Converting emotional voice to motion for robot telepresence

Angelica Lim, Tetsuya Ogata, and Hiroshi G. Okuno Graduate School of Informatics Sakyo, Kyoto, Japan

angelica@kuis.kyoto-u.ac.jp, ogata@i.kyoto-u.ac.jp, okuno@kuis.kyoto-u.ac.jp

Abstract—In this paper we present a new method for producing affective motion for humanoid robots. The NAO robot, like other humanoids, does not possess facial features to convey emotion. Instead, our proposed system generates pose-independent robot movement using a description of emotion through *speed*, *intensity, regularity and extent (DESIRE)*. We show how the DESIRE framework can link the emotional content of voice and gesture, without the need for an emotion recognition system. Our results show that DESIRE movement can be used to effectively convey at least four emotions with user agreement 60-75%, and that voices converted to motion through SIRE maintained the same emotion significantly higher than chance, even across cultures (German to Japanese). Additionally, portrayals recognized as happiness were rated significantly easier to understand with motion over voice alone.

I. INTRODUCTION

Robot telepresence has recently become a popular way to "be in two places at once". A typical application is office presence: an employee can control a remote robot, allowing mobility and video interaction with co-workers far away. These systems are touted because embodiment of remotely operated robots provide a "presence" that exceeds communication by videoconference [1]. Despite its utility in an increasingly internationalized world, telepresence has been relatively unexplored for home use, to connect distant family members (such as the elderly [2]).

Social communication should be the primary focus in a telepresence application connecting loved ones. One way to improve social communication is to make a telepresence system that can convey emotion as clearly as possible; displays of emotion are known to be important for communication and social bonding [3] [4]. So far, very few telepresence robots possess a humanoid form (cf. Telenoid¹) and thus cannot convey emotion through the rich, engaging bond of body language.

Implementing emotionally-charged telepresence is difficult because typical approaches "classify" affect. Fellous argues: "Implementing emotions as 'states' fails to capture the way emotions emerge, wax and wane, and subside." [5]. Yet most systems focus on categorizing speech into one of several emotions. For instance, Kismet [6] classified voice into one of 5 states (approval, attention, prohibition, soothing and neutral), and the result was fed into the robot's behavioral system to generate an appropriate response. Neurobaby [7] was a





Fig. 1: Our emotion transfer system the context of a multi-modal telepresence application.

simulated infant that responded to changes in voice, using a neural network to detect one of four emotional states. In contrast with these systems, a telepresence robot user may convey any one of a number of emotions (or even a mixture of emotions) at any time.

Generating recognizable robot emotions through body language is also not without limitations. Conventional approaches play pre-defined robot poses such as raised arms for surprise [8], or an aggressive stance for anger [9]. These allow for scaling so that the gesture can be more or less activated [10] [11], but its usage is restricted: 1) the poses and emotions are limited to a hand-designed set, and 2) the robot cannot do any other gestures (e.g., emblematic, interactive, punctuative [12]) at the same time. For example, "angrily pointing" would not be possible. Another promising method is the use of the Laban Movement analysis [13], which prefers to convey affect through features relating to weight, space, and time.

In this paper, we propose an emotional telepresence framework to transfer emotional voice to robot gesture (Fig. 1) that 1) does not require emotion classification and 2) can be applied to virtually any gesture. 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.

The rest of this paper is organized as follows. First, we will give an overview of our general approach, and describe in detail our implementation. Experiments and results will then be presented.

II. AN EMOTION TRANSFER FRAMEWORK

We first overview our general philosophy and requirements for an emotion transfer system.

A. Pose independence

To be pose-independent, we do not focus on gestures themselves, but on their dynamics. Recent studies in neuroscience show that movement alone may induce emotion. In [14], it was shown that the observation of angry hand actions recruited the same areas of the brain as when viewing an angry face. A similar result was found in the comparison of whole body expressions of fear, in both dynamic as well as static conditions [15]. In psychology, point-light displays of dancers portraying fear, anger, grief, joy, surprise and disgust were recognized significantly above chance (63%) in [16]. These impoverished displays of movement were recognized significantly above chance even when inverted. Another study recorded point-light displays of actors performing "drinking and knocking movements" in 10 different affects [17]. Their results showed that the Circumplex affect model's [18] pleasantness and activation dimensions could be recovered in the movements. We do not claim that pose is irrelevant (their importance is well-known [15] [19]), but these studies give evidence to a strategy already well-known in the animation industry: we can use motion to imply "emotions" for humans as well as non-humans, and even for objects like vacuum cleaners [20].

B. Classifier-free

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) [6]. It is common to use these models for affect generation, but it is not always clear how to map the dimensions to affective input. For example, what makes for a pleasant or unpleasant voice? Features ranging from speed to voice quality to pitch changes have been found to be correlated with pleasantness [21], which is why emotion classification of feature vectors is such a popular approach (e.g., [22] [23]).

C. Extensibility to other domains and applications

We design our emotion transfer framework to be extendible to other modalities and applications. Although we focus here on speech input and gesture output, a framework should allow the inverse (emotional movement as input, and speech as output), or human speech to robot speech (Fig. 1). We also propose that such a framework be adaptable to modalities other than speech and gesture; for example, transferring emotion from a conductor's gesture to music. Finally, although we focus on Aldebaran Robotics' humanoid robot NAO², ideally the framework should be easily applied to other robot hardware.



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

III. DESIRE: DESCRIPTION OF EMOTION THROUGH SPEED, INTENSITY, REGULARITY AND EXTENT

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) and graphics communities to convey some aspects of emotion [9] [20] [24]. For example, fast or large motions give an impression of energy and may convey anger or happiness [13]. In this study, we examine speed and extent, along with two other parameters called regularity and intensity which are well-known in the fields of affective speech and music. 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 (see Fig. 2):

- Hardware-independent layer: A mapping from DESIRE to perceptual features (based on discipline-specific studies).
- *Hardware-specific layer*: A mapping the perceptual features to a hardware-specific implementation (by the system designer).

A. Hardware-independent layer

The DESIRE framework was inspired by commonalities found between emotion in movement, voice and music ([28], [25]). 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 [21], or *animation* in gesture [27]. We have summarized our literature review in Table I. This table is not meant to be comprehensive, but rather to give some practical guidelines for how to map SIRE to various modalities. In some cases, multiple interpretations are proposed. For example, it may not be clear how to implement joint "phase shift" on robots with few degrees of freedom. In this case, we hypothesize that another interpretation of regularity, such as "directness", can

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

		Modality mappings to relevant emotional features			
Parameter	Description	Voice	Gesture		
Speed Intensity Regularity Extent	slow vs. fast gradual vs. abrupt smooth vs. rough small vs. large	speech rate [21], pauses [25] voice onset rapidity [25], articulation [21] jitter [25], voice quality [21] [25] pitch range [21], loudness [25]	velocity [26], animation [27], quantity of motion [28] acceleration [26], power [29] directness [26], phase shift [24] [17], fluidity [30] spatial expansiveness [29] [27], contraction index [26]		

be used in practice for mapping, though verifying this is set for future work. How to evaluate a particular mapping choice will be examined in Section IV-A.

B. 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.

1) Extracting SIRE from Voice: 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 [31]. 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. We note informally that, over a short utterance, a miscalculation of a few syllables can have a significant influence on the calculated speed between 0 and 1. Therefore, future work should explore the extent of this practical limitation, for example by reliably extracting syllables with an Automatic Speech Recognition system such as HTK [32], using long utterance frames, and so on.

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 k of size n (in our experiments, n = 1024) by:

$$p(k) = \sum_{i=0}^{n-1} x(k \cdot n + i)^2 \tag{1}$$

and onset rapidity is:

$$\max_{k=1,...,N/n} p(k) - p(k-1).$$
 (2)

Regularity is mapped here to the inverse of jitter in the voice sample, as jitter has been related to vocal "roughness" in [33]. Jitter is defined for each utterance as:

$$\frac{1}{N-1}\sum_{t=1}^{N}|x(t)-x(t-1)|$$
(3)

Extent is defined as the range of pitch in the speaker's voice. We used the Snack sound toolkit³ implementation of the average magnitude difference function (AMDF) [34], an autocorrelation function, 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. 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. As for speed, we used a minimum speech rate of 2 syllables per second and a maximum speech rate of 7 syllables per second, based on our input set. Similarly, these values should ideally be set according to context or the speaker's learned history.

2) Gestural mappings for NAO Humanoid: In this section we describe how we implement the perception of speed, intensity, regularity and extent on the NAO humanoid robot.

A gesture is considered here as a simple motion from a "base posture" p_0 to an "extended posture" p_1 and back to the "base posture" to be achieved at three target times t_0, t_1, t_2 (Fig. 3). Intuitively, speed S is mapped by performing a simple linear down-scaling of times t_1 and t_2 (e.g., see Algorithm 1, lines 4-5). Intensity I is modulated by bringing the start position and middle position times temporally closer together, effectively increasing the relative acceleration to reach the middle position (e.g., Algorithm 1, line 4). Regularity Ris 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 (Algorithm 2, line 3), and for the head, an irregular movement is created by adding side-to-side movement (Algorithm 3, lines 1-3). The amount of side-to-side movement δ_{s1} , δ_{s2} 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 (e.g., Algorithm 1, line 2).

³http://www.speech.kth.se/snack/



Fig. 3: Timeline of an arm gesture.

Formally, we define gestures for three of NAO's end effectors: the head, left arm, and right arm.

Let us define the arm gesture $((p_0, p_1), (t_0, t_1, t_2))$, where p_0 is the base position of the hand in 3D, p_1 the extended position of the gesture, and t_0 , t_1 and t_2 are the times in seconds at which the base, extended, and base positions are to be reached, respectively. We say that \underline{m} is the minimum time needed for the robot to change position from p_0 to p_1 safely. We find a temporal offset δ_t using \underline{r} , a maximum time length used to offset joint movements.

We define the head movement $((\kappa_0, \kappa_1), (t_0, t_1, t_2))$ where κ_0 is the base configuration (pitch and yaw values) of the head, and κ_1 the extended posture. For the head, we find δ_{s1} and δ_{s2} , spatial offsets for the base and extended yaw values, taken from a normal distribution with variance proportional to $\sigma = (1 - R)$.

The mappings for the left and right arms, and the head, are outlined in Algorithm 1, 2 and 3.

 $\begin{array}{l} \textbf{Algorithm 1 MAPSIRE} \left\langle \mathcal{G}_{\textbf{NAO}, \textbf{left arm}} \right\rangle \\ \hline \textbf{Require:} \quad (S, I, R, E) \in [0, 1]^4 \\ \textbf{Require:} \quad p_0, p_1, \in \mathbb{R}^3 \\ \textbf{Require:} \quad t_0 \leq t_1 \leq t_2 \in \mathbb{R} \\ \textbf{Ensure:} \quad g_{out}.t_1, g_{out}.t_2 \geq \underline{m} \\ 1: \quad g_{out}.p_0 = p_0 \\ 2: \quad g_{out}.p_1 = p_0 + E \cdot (p_1 - p_0) \\ 3: \quad g_{out}.t_1 = \max((1 - S) \cdot I \cdot t_1, \underline{m}) \\ 5: \quad g_{out}.t_2 = \max((1 - S) \cdot t_2, \underline{m}) \\ 6: \quad \textbf{return} \quad g_{out} \end{array}$

IV. EVALUATION

We performed three experiments to answer the following research questions:

- **Q1**: How well do the implementations (described in Sec. III) map to the SIRE parameters?
- **Q2**: Can the robot produce happiness, sadness, anger and fear by modulating SIRE, and how well?
- Q3: What DESIRE values produce the highest recognition of happiness, anger, fear and sadness respectively?
- Q4: How well can our system transfer emotion from speech to gesture?

Algorithm 2 MAPSIRE $\langle \mathcal{G}_{NAO,right arm} \rangle$ Require: $(S, I, R, E) \in [0, 1]^4$ Require: $p_0, p_1, \in \mathbb{R}^3$ Require: $t_0 \leq t_1 \leq t_2 \in \mathbb{R}$ Ensure: $g_{out}.t_1, g_{out}.t_2 \geq \underline{m}$ 1: $g_{out}.p_0 = p_0$ 2: $g_{out}.p_1 = p_0 + E \cdot (p_1 - p_0)$ 3: $\delta_t = (1 - R) \cdot \underline{r}$ 4: $g_{out}.t_0 = \delta_t + t_0$ 5: $g_{out}.t_1 = \delta_t + \max((1 - S) \cdot I \cdot t_1, \underline{m})$ 6: $g_{out}.t_2 = \delta_t + \max((1 - S) \cdot t_2, \underline{m})$ 7: return g_{out}

Algorithm 3 MAPSIRE $\langle \mathcal{G}_{NAO,head} \rangle$

 $\begin{array}{l} \textbf{Require:} \quad (S, I, R, E) \in [0, 1]^4 \\ \textbf{Require:} \quad \kappa_0, \kappa_1 \in [0, \pi/2] \times [0, \pi] \\ \textbf{Require:} \quad t_0 \leq t_1 \leq t_2 \in \mathbb{R} \\ \textbf{Ensure:} \quad g_{out}.t_1, g_{out}.t_2 \geq \underline{m} \\ 1: \quad \delta_{s1}, \delta_{s2} \sim \mathcal{N}_{0,\underline{\sigma}} \\ 2: \quad g_{out}.\kappa_0 = (\kappa_0.pitch, \kappa_0.yaw + \delta_{s1}) \\ 3: \quad g_{out}.\kappa_1 = (\kappa_0.pitch + (1 - S) \cdot E \cdot (\kappa_1.pitch - \kappa_0.pitch), \kappa_0.yaw + \delta_{s2}) \\ 4: \quad g_{out}.t_0 = t_0 \\ 5: \quad g_{out}.t_1 = \max((1 - S) \cdot I \cdot t_1, \underline{m}) \\ 6: \quad g_{out}.t_2 = \max((1 - S) \cdot t_2, \underline{m}) \\ 7: \quad \textbf{return} \quad g_{out} \end{array}$

• **Q5**: Does the addition of motion improve recognition of emotion?

A. Experiment 1: SIRE Parameters

In this experiment, our goal was to verify that our SIRE mappings agree with evaluators' perceptions of speed, intensity, regularity and extent (Q1). We recruited 29 self-reported native English speakers through the Internet, 79% male and 21% female, and asked them to rate videos of robot gestures generated by modulating each SIRE parameter independently. For each of S, I, R, E respectively, we compared the values 0.1 and 0.9 while keeping the other parameters constant at 0.5, with the exception of regularity, which was kept constant at 0.9 to avoid random perturbations. Two gestures were tested: a head nodding movement, and an arm extension movement (Fig. 4).

Each participant was asked to compare a total of eight pairs of videos – four for head-nodding, four for arm gestures. Depending on the parameter being compared, the participant was asked to choose one of the two videos (e.g., comparing a head nod at extent 0.1 and 0.9):

- Which has higher speed?
- Which is more intense?
- Which is more regular?
- Which has larger extent?



Fig. 4: Gesture positions for Experiment 1

TABLE II: Recognition of high-low mappings of SIRE parameters, for arm gesture (AG) and head nod (HN) and average difficulty from 1 (very easy) to 5 (very difficult)

Parameter	% AG	Difficulty	% HN	Difficulty
Speed	100	1.3	100	1.4
Intensity	86	2	93	2
Regularity	93	1.7	86	1.9
Extent	97	1.6	100	1.4

Following each question, the participant was asked to rate the difficulty of the question on a 5-point Likert scale: very easy, somewhat easy, neutral, somewhat difficult, very difficult. Evaluators could freely give comments on each choice, and had no time limit.

Results and discussion. As Table II shows, the mappings as described agree with raters' perceptions at more than 86% in all cases (Q1). The prototypical features of Speed and Extent are the most easily recognized. This is shown by the nearly perfect recognition rates and subjective evaluation of "very easy" to distinguish. Intensity and regularity are slightly less recognized, but still more than 86% of raters were still able to tell the difference between "low intensity" and "high intensity", as well as "irregular" and "regular", with an average rating of "somewhat easy". One explanation for the lower ratings of regularity and intensity may have been the choice of wording. According to rater comments, the feeling of "intensity" could also be given in a slow, purposeful stare or movement. This suggests that future tests should use unambiguous words to describe the dimension, such as gradual vs. abupt. The word "regular" was also understood as "normal" by at least one evaluator. This is a problem when evaluating motions using one axis only; they look "robotic" rather than "normal". In this case, the addition of irregularity could make the movement look more human-like, and therefore more "regular". Like intensity, we suggest to use another description to evaluate regularity, like smooth vs. rough.

B. Experiment 2: Motion only

In this experiment, our goal was to find out whether the robot could produce recognizable emotions through motion (Q2) and if so, find out what SIRE values produced each of four emotions (Q3). 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 (Q4).

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 Speech⁴, which is a database of emotional speech recorded by professional German actors [35]. Each sample was a wave file at 16kHz, all with the same semantic content: "Heute abend könnte ich es ihm sagen." (Translated in English as "Tonight I could tell him."). The volume of all files had been previously normalized. Four samples each of happiness, sadness, fear, and anger were used as input, for a total of 16 samples ranging from 1.5 - 3.9 seconds. We selected only utterances with recognition rates of 80% or higher by German evaluators.

Given the SIRE values extracted from these audio samples, we generated 16 movement sequences and showed them to the participants 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. After each sequence, the participants were given 5 seconds to choose one of happiness, sadness, anger, or fear in a forced-choice questionnaire.

Results and discussion. In Table III, we outline the movements which had the highest agreement between evaluators for each of the four emotions. It shows that the robot, by changing the dynamics of the same gesture, could produce recognizable emotions at more than 60% inter-rater agreement (**Q2**). Table III also gives the SIRE parameters which achieved them (**Q3**). In summary, we can see that:

- happiness can be produced with med-high speed, medlow intensity, med-low regularity, and med-large extent
- *sadness* can be produced with low speed, medium intensity, med-high regularity, and medium extent
- *anger* can be produced with med-high speed, high intensity, med-low regularity, and large extent
- *fear* can be produced high speed, med-high intensity, medium-low regularity, medium extent

This is not an exhaustive list of possibilities (to do so, we would need to examine much more points in the 4-D SIRE space), but it gives a useful hint for designing motions with these emotions. For example, one motion sample was recognized in the majority as anger, though the source speech file was happiness, with (S, I, R, E) = (0.71, 0.75, 0.75, 0.91).

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



Fig. 5: Experiment 2: Recognition results of gestures generated by voice samples



Fig. 6: Recognition results of motion (Exp. 2) and voice+motion (Exp. 3) compared to voice only (Exp. 3).

TABLE III: Sequences with best agreement between evaluators and their corresponding SIRE values.

Emotion	Agreement (%)	S	Ι	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

In a detailed analysis (Fig. 5), we note that some "happiness" samples were recognized as anger with greater agreement (48%). The reason for this may be explained through the results of Experiment 3. In fact, the voice-only condition of h3 and h4 (Fig. 5) were the most poorly recognized of all samples, at 33% and 38% respectively, meaning that the German vocal expressions of happiness were difficult to recognize for Japanese. Since both voice and motion were similarly poorly recognized, this gives evidence to our hypothesis that the emotion for both voice and motion have the same fundamental basis in our dynamic parameter set.

Figure 6 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 (Q4). We compare them here with the voice recognition results from Experiment 3 (Section IV-C), which act as an upper bound.

C. Experiment 3: Adding motion to voice

In this experiment, we assessed the usefulness of the DESIRE system for emotional expression via telepresence (Q5). We compare the emotion recognition of 1) a humanoid playing a voice only with 2) a humanoid playing a voice *and* performing the associated DESIRE motion. Additionally, as opposed to Experiment 2 in which only one type of arm gesture was used, here we test two different arm gestures and head motion. Our hypothesis is that adding motion will increase recognition of emotions, or make the impression of the emotion stronger over voice only.

We recruited 21 evaluators with normal (or corrected to normal) vision and hearing from Kyoto University Graduate School of Informatics. The participants were male, ranging in age from 21-27 (mean=24.5, stdev=4.1). This experiment was performed with a NAO robot placed on a table as shown in Fig. 7. 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 Section IV-B-3.

We presented the participants with two robot conditions.

- Condition 1: Voice only. The robot stayed still in a neutral position (Fig. 7) 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.

Given the 16 utterances, 32 robot sequences were generated given the two conditions. 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".

Results and discussion. In this experiment, we explore the result of adding motion. Since we saw in Experiment 2 that



Fig. 7: Experimental setup for experiment 3, position of robot during voiceonly condition.



Fig. 8: Experiment 3: Comparison of ease of understanding, from difficult (1) to easy (4), for correctly recognized samples.

motion was generally less recognized than voice, the expected result is that adding motion would not improve recognition compared to voice. This was the case for happiness, sadness and anger (Fig. 6). On the contrary, for the emotion that was the most difficult to recognize through voice only–fear–the addition of motion increased recognition from 49% to 55% (Q5). This may be explained by the fact that, according to evaluator comments, vocal expressions of fear and sadness were both gave a negative impression, and were easily confused between each other. However, according to Table III, we see that fear movements differ greatly from sadness along the speed dimension, which may explain this increase in perceptual separation.

Next, we compare the evaluator's ratings for "ease of understanding", i.e., how clearly was the emotion expressed? Intuitively, this is the factor we wish to increase by adding robot motion. For a given rater, when the samples was recognized correctly for both voice and voice+motion, we compared the rater's ease of understanding for two sequences.

In Fig. 8, we notice that anger was better understood when the robot was still than when the robot was moving. This could be due to the choice of "neutral" stance during the voiceonly condition; the robot was 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 [36] [37]. 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). This suggests that to provide reliable anger movements 1) regularity should not be implemented with left-to-right spatial movement, but perhaps using temporal regularity instead or 2) the head should remain still (i.e., only arm gestures should be used). Experiment 2 gives further evidence to this hypothesis, because recognition of anger was relatively high, and the samples only used arm movements. This also suggests that a humanoid that maintains a forward-facing stare may be more easily viewed as angry, which could have general implications in HRI as to how robots are perceived.

We also see that in Fig. 8 that happiness in particular was 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 the use of a gesticulating humanoid may be most useful for portraying joy through a telepresence robot.

V. CONCLUSIONS AND FUTURE WORK

In this study, we studied a hypothesis that emotion from voice could be effectively transferred to motion through only four features (speed, intensity, regularity and extent). Our cross-cultural experiments suggest that this is indeed the case. These exploratory results are promising, but should be replicated with other robots, movements and/or cultures to add support to the data presented here. Our analysis provided two other surprising results. The first is that German fear voices were poorly recognized by Japanese, and that adding SIRE-generated motion improved the recognition of fear by 6 points. Secondly, we found that when SIRE-motion matched the voice, portrayals of happiness were rated significantly easier to understand than for voice alone.

Our results suggested that future work should study how to best integrate emotional motion with other cues, like pose. For example, a multi-modal system integrating prototypical poses could be useful for situations when cues contain conflictual meanings, such as sarcasm. Interactions with other cues could also be tested, such as adding flashing red eyes for anger. An ideal voice-to-motion system would also integrate the semantic content of the voice, for example using keyword-spotting to choose the appropriate gesture template on which to add a SIRE-motion.

In the future, we plan to test the system's real-time capabilities and the extent of this system's emotional vocabulary. In our study, voice actors were used to convey emotion that was easy to recognize; it would be useful to examine how non-extreme emotions, or emotions like surprise, disgust, and love, are portrayed. We also plan to extend the framework to other modalities which may benefit from emotion generation, such as robot music or dance (Fig. 2).

REFERENCES

- D. Sakamoto, T. Kanda, T. Ono, H. Ishiguro, and N. Hagita, "Android as a Telecommunication Medium with a Human-like Presence," *HRI*, pp. 193–200, 2007.
- [2] F. Michaud, P. Boissy, and D. Labont, "Telepresence Robot for Home Care Assistance," AAAI Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics, 2004.
- [3] E. T. Rolls, "Précis of The brain and emotion," *Behavioral and Brain Sciences*, pp. 177–233, 2000.

- [4] P. Ekman and R. J. Davidson, *The Nature of Emotion: Fundamental Questions*, 1st ed. Oxford University Press, USA, Dec. 1994.
- [5] J. Fellous, "From Human Emotions to Robot Emotions," Architectures for Modeling Emotion: Cross-Disciplinary Foundations, American Association for Artificial Intelligence, pp. 39–46, 2004.
- [6] C. Breazeal, *Designing sociable robots*, 1st ed. The MIT Press, May 2004.
- [7] T. Yamada, H. Hashimoto, and N. Tosa, "Pattern recognition of emotion with neural network," *IECON*, pp. 183–187, 1995.
- [8] 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.
- [9] H. Robot, "Effective Emotional Expressions with Emotion Expression Humanoid Robot WE-4RII," *IROS*, pp. 2203 – 2208, 2003.
- [10] A. Beck, A. Hiolle, A. Mazel, and R. Losserand, "Interpretation of Emotional Body Language Displayed by Robots," *AFFINE*, pp. 37–42, 2010.
- [11] A. Beck, L. Cañamero, and K. A. Bard, "Towards an Affect Space for Robots to Display Emotional Body," *RO-MAN*, pp. 464–469, 2010.
- [12] M. L. Knapp and J. A. Hall, Nonverbal Communication in Human Interaction. Cengage Learning, Mar. 2009.
- [13] 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.
- [14] 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.
- [15] J. Grèzes, S. Pichon, and B. D. Gelder, "Perceiving fear in dynamic body expressions," *NeuroImage*, vol. 35, pp. 959–967, 2007.
- [16] 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.
- [17] 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.
- [18] J. Russell, "A circumplex model of affect." Journal of personality and social psychology, vol. 39, no. 6, p. 1161, 1980.
- [19] B. D. Gelder, "Towards the neurobiology of emotional body language," *Neuroscience*, vol. 7, no. March, pp. 242–249, 2006.
- [20] M. Saerbeck and C. Bartneck, "Perception of Affect Elicited by Robot Motion," in *HRI*, pp. 53–60, 2010.
- [21] 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.
- [22] R. Fernandez, R. W. Picard, I. B. M. T. J. Watson, and Y. Heights, "Classical and Novel Discriminant Features for Affect Recognition from Speech," *INTERSPEECH*, pp. 4–8, 2005.
- [23] Z. Zeng, I. C. Society, M. Pantic, and S. Member, "A Survey of Affect Recognition Methods : Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.
- [24] K. Amaya, A. Bruderlin, and T. Calvert, "Emotion from Motion," *Graphics Interface*, pp. 222–229, 1996.
 [25] P. Juslin and P. Laukka, "Communication of emotions in vocal
- [25] 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.
- [26] M. Mancini and G. Castellano, "Real-time analysis and synthesis of emotional gesture expressivity," in *Proc. of the Doctoral Consortium* of ACII, 2007.
- [27] P. E. Gallaher, "Individual differences in nonverbal behavior: Dimensions of style." *Journal of Personality and Social Psychology*, vol. 63, no. 1, pp. 133–145, 1992.
- [28] A. Camurri and G. Volpe, "Communicating Expressiveness and Affect in Multimodal Interactive Systems," *IEEE Multimedia*, vol. 12, no. 1, pp. 43–53, 2005.
- [29] H. G. Wallbott, "Bodily expression of emotion," *European Journal of Social Psychology*, vol. 28, no. 6, pp. 879–896, 1998.
- [30] C. Pelachaud, "Studies on gesture expressivity for a virtual agent," Speech Communication, vol. 51, no. 7, pp. 630–639, 2009.
- [31] 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.

- [32] S. J. Young and S. Young, "The HTK Hidden Markov Model Toolkit: Design and Philosophy," *Entropic Cambridge Research Laboratory*, *Ltd* (technical report), vol. 2, pp. 2–44, 1994.
- [33] 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.
- [34] 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.
- [35] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "Database of German Emotional Speech," *Interspeech*, pp. 1517–1520, 2005.
- [36] 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.
- [37] R. A. Hinde, T. E. Rowell, "Communication by Postures and Facial Expressions in the Rhesus Monkey" *Proceedings of the Zoological Society of London*, vol. 138, pp.1–21, 1962.