The MEI Robot: Towards Using Motherese to Develop Multimodal Emotional Intelligence

Angelica Lim, *Student Member, IEEE*, and Hiroshi G. Okuno, *Fellow, IEEE*

***Abstract*—We introduce the first steps in a developmental robot called MEI (multimodal emotional intelligence), a robot that can understand and express emotions in voice, gesture and gait using a controller trained only on voice. Whereas it is known that humans can perceive affect in voice, movement, music and even as little as point light displays, it is not clear how humans develop this skill. Is it innate? If not, how does this emotional intelligence develop in infants? The MEI robot develops these skills through vocal input and perceptual mapping of vocal features to other modalities. We base MEI's development on the idea that motherese is used as a way to associate dynamic vocal contours to facial emotion from an early age. MEI uses these dynamic contours to both understand and express multimodal emotions using a unified model called SIRE (Speed, Intensity, irRegularity, and Extent). Offline experiments with MEI support its cross-modal generalization ability: a model trained with voice data can recognize happiness, sadness, and fear in a completely different modality—human gait. User evaluations of the MEI robot speaking, gesturing and walking show that it can reliably express multimodal happiness and sadness using only the voice-trained model as a basis.**

***Index Terms*—Cross-modal recognition, emotion recognition, gait, gaussian mixture, gesture, motherese, SIRE, voice.**

I. Introduction

EMOTIONS can be conveyed in many ways outside of facial expression. Consider the sympathy we feel for a quivering puppy—he looks scared, we might say. Or the shouts of neighbors fighting in a foreign language; they can still sound angry even without knowing what they are saying. Even a singer on stage can belt out a tune with such emotional intensity that listeners are moved to tears. It is a curious phenomenon: how can mere movements or sounds affect us in this way? This kind of "emotional intelligence"—to sense emotions through various means—appears to be built into any normal-functioning human and even some animals. We propose that robots, too, can develop the ability to understand emotions, no matter the communication channel. But first we must investigate how we as humans develop this ability, whether the channel is movement, voice, or any other type of sound.

Manuscript received May 22, 2013; revised December 02, 2013; accepted March 31, 2014. Date of publication April 15, 2014; date of current version June 10, 2014. This work was supported by KAKENHI (S) 24220006 and the Honda Research Institute Japan.

The authors are with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: angelica@kuis.kyoto-u.ac.jp; okuno@kuis.kyoto-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAMD.2014.2317513

First, consider that any movement can be colored with emotion. In the 1980's, the neurologist Manfred Clynes performed extensive cross-cultural studies using his sentograph, a device to measure touch [16]. He asked subjects to tap the device at regular intervals while imagining emotions such as love, hate, and grief. The resulting dynamic forms of the movements appear similar across cultures, e.g., abrupt, jabbing movements for hate, and soft, lethargic taps for sadness. More recently, psychologists show the importance of movement by attaching balls of light to actors' joints, turning off the lights, and recording these so-called 'point-light' displays. Actors in [57] made "drinking and knocking" movements in 10 different emotions, and despite the impoverished format, raters could still recognize emotional information. Walking style, or gait, can also reveal the walker's emotional state [49], [59]. For instance, heavyfootedness can signify anger, and slow walking speed can signify grief. For a given emotion, the dynamics of gesturing and walking already appear to have underlying similarities.

Another common way we express emotions is through the voice. In a typical study on emotional voice, researchers ask actors to utter gibberish words in various emotions. Van Bezooijen *et al.* [76] asked native Dutch speakers to say *twee maanden zwanger* ("two months pregnant") in neutral and nine other emotions, and then played them to Dutch and Japanese subjects. Changes in properties like pitch, tempo and loudness of speech due to physiological changes appear to create universally perceptible emotional differences [64]. Juslin and Laukka [29] reviewed dozens of studies of this kind, and found that hearers can judge anger, fear, happiness, sadness and tenderness in voice almost as well as facial expressions, around 70%. Emotion in sounds may even stretch to the animal kingdom; among some animals, alarm calls mimic human fear vocalizations, with high-pitches and abrupt onset times [69]. In primates, dominant males often emit threatening vocalizations with characteristics similar to those of human anger [69].

It has long been speculated that whether it be a step, tone of voice, or even a musical phrase, the expression of emotions have the same underlying dynamic "code" [16], [29], [73]. For example, both loud, intense voices and large, forceful movements convey anger. Sadness can be conveyed through small and slow movements and quiet, slow speech. Indeed, a stomping gait can indicate fury, and a lethargic walk can portray depression.

In this paper, we address this open question in developmental robotics: how might a robot develop this multimodal emotional intelligence (MEI), based on evidence in human development of emotional intelligence in voice, movement, and even music?

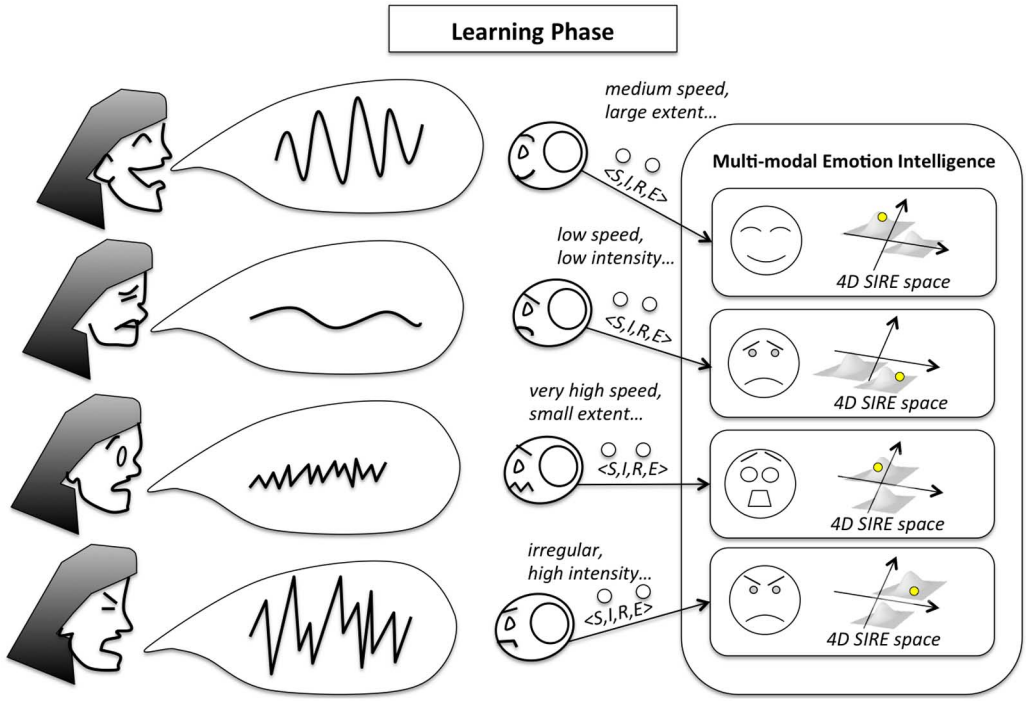

Fig. 1. Overview of the learning phase of MEI through infant-directed speech. The robot (center) observes, for example, a happy face concurrent with a happy voice. The speed, intensity, irregularity and extent (SIRE) are then extracted from auditory input. Finally, this SIRE tuple is added to the relevant class model, strengthening the association between the facial class and vocal dynamics. In our experiments, the face emotion recognition is represented as a class tag, to be replaced in the future by the output of an FACS engine such as [42] (or other affective groundtruth).

A. The Case for Emotion and Motherese

We propose that emotional voice is learned through associative learning with emotional face, and that other MEI abilities scaffold onto these vocal dynamics (see Fig. 1.) Consider the universal phenomenon of infant-directed (ID) speech, or "motherese." ID speech is a highly varying style of speaking with contours and properties (e.g., pitch, intensity) also found in exaggerated adult-directed (AD) emotional speech: "Acoustic analyses showed few differences between the ID and AD samples, but robust differences across the emotions [74]."

In other words, motherese is emotional speech, and it co-occurs with exaggerated emotional facial expressions [70]. For deaf children, facial expression accompanying emotive signing is called "visual motherese":

> "Hearing babies know when their parents are happy, worried, angry, or excited from their voices, even when the baby cannot see the parent's face. Your deaf baby needs to see your facial expression and your body movements to get the same information. Are you smiling, and letting your signs flow? Are you frowning and signing sharp, emphatic signs as you run to cover the electric outlet? Are you pretending to cry as you see a sad character in a story?[1]

Motherese is known to be necessary for social and verbal development and exists across cultures (e.g., [21] and [34]). Yet, compared to most studies of ID speech which concentrate on language acquisition (e.g., [51]), ID speech and its role on the comprehension of prosody has received little attention [19], [62]. Soken and Pick [70] suggest an important role played by motherese for developing the correspondence between the face and voice:

> "It has been shown that infants are attracted by and attend to motherese, which is characterized by more exaggerated intonation and higher pitch than adult-to-adult speech. Concurrent with the exaggerated speech of motherese, there are probably exaggerated facial displays, allowing infants to explore the particular aspects of the face (e.g., exaggerated mouth and brow movement). [. . .] Child-centered displays may serve as opportunities for learning about affective events."

Lewis [37] proposed that young infants selectively respond to the strong affective character in speech since prosody is initially more salient than phonetic information in the development of language. Fernald [19] also notes that ID speech's "melodies are characterized not only by fundamental frequency, but also by intensity or amplitude envelope, and by temporal structure. For example, expressions of approval such as 'Good!' or 'Clever girl!' are typically spoken using exaggerated rise-fall F0 contours [and] expressions of prohibition or warning such as 'No!' or 'Don't touch that!' are spoken with low pitch and high intensity." How are these affective expressions processed by the infant?

[1]"My Baby's Hearing" guide, Boys Town National Research Hospital for childhood deafness, visual impairment and related communication disorders: http://www.babyhearing.org/languagelearning/buildconversations/Motherese.asp.

Psychological studies show that development of vocal emotion recognition already occurs within the 1st year of life (see [25] for a recent review). At 5-months-old, infants look longer at happy, sad, or angry voices when they co-occur with facial photos but not with black and white checkerboards. According to Walker–Andrews [77], this result suggests that, "the presence of the face acts as a setting for attending to the affective quality of the voice." A neurological study using event-related potentials (ERP) showed that 7-month-olds are able to recognize happiness or anger when they co-occur with the matching voice, even when all voices are presented asynchronously [26]. At 12-months, the development of recognition of angry voices appears complete, coinciding with the onset of crawling which increases access to expressions of anger [11], [25].

When are infants able to generalize emotion to other modalities? **Across languages**: When listening to an unfamiliar language in ID-speech, 5-month-olds already smile more often at approving voices and show negative affect when listening to prohibitions [20]. **Across voice and movement**: At 7-months-old, babies look longer at affectively concordant point-light displays [70] of facial movements and voices. **In music**: whereas 3-month-old infants cannot discriminate between sad and happy music, 9-month-olds can make the distinction [23]. In between, at 5 and 7 months, infants showed order bias, for instance being able to distinguish when sad music was presented before happy, but not the inverse.

In summary, it is clear that multimodal emotional intelligence is available within the first year of life [77]—even before the onset of speech [34]—yet no detailed proposals exist for the development of these multimodal skills, either in human psychology or robotics. In developmental psychology, [26] and [77] discuss only the possibility of associative learning between the affective voice and face. In developmental robotics, the "intuitive parenting" paradigm has been proposed for grounding emotional face [7], [78], but other modalities have not yet been examined. We therefore describe here a developmental robot system with the goal of: 1) advancing computational models of emotion in both developmental psychology and autonomous robotics; and 2) using infant development as a clue to develop a robot with a powerful emotion system.

II. MEI Based on the SIRE Model

The aim of the MEI system is to avoid the compartmentalization of emotional intelligence. As mentioned in the introduction, humans have the ability to generalize emotion to new contexts, yet this remains a major challenge for robots. This is because current paradigms would typically train a separate model for each of the cases we imagined: an emotion module to interpret the movements of the quivering puppy (such a system does not exist, though many do for human gestures, e.g., [13]), an emotion module for a novel language (e.g., cross-language emotion recognition [5] is a recent topic), an emotion module for the operatic singer (many emotion recognition systems exist for music [33], but none exist for singing voice). In fact, twice the number of these modules is typically implemented: one for recognition of the above-mentioned cases, and one for their expression. For example, Kismet, one of the few integrated emotional robot systems, has a voice emotion recognition module that is independent from the emotional voice expression module [8]. This means that emotional voice input, though recognized, will never improve the way the robot's own emotions are expressed.

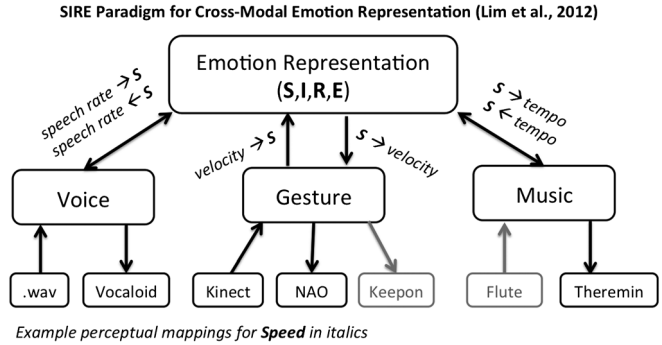

Fig. 2. SIRE paradigm from [39] for experiments across voice, gesture and music.

Unfortunately, this multiplication of specialized systems is not scalable for an autonomous robot. Therefore, we seek an integrated emotion system with the following requirements: 1) a low-dimensional emotion representation; 2) for multiple modalities; and 3) for analysis and synthesis. A model that fulfills these requirements remains an open problem according to a recent review of affect models [27]. To address this challenge, the MEI system uses a: 1) 4-dimensional; 2) cross-modal SIRE [38] emotion paradigm, coupled with a statistical Gaussian Mixture Model (GMM); 3) capable of both recognition and expression.

The SIRE paradigm has shown promise in finding emotion "universals" across voice, movement, and music [39] (see Fig. 2). This was tested by mapping low-level features to high-level perceptual features, such as those in Table I. SIRE stands for Speed, Intensity, irRegularity, and Extent [39], [40] where the tuple contains four values (S, I, R, E) on [0,1]. By extracting the dynamics from a voice and mapping it to a gesture, [38] found that, for instance, an expression of sadness is slow with low intensity, whether expressed in voice, gesture or music. Fear is fast, intense and irregular. A recent study in psychology supports these findings: in Sievers *et al.*'s bouncing ball experiment, emotion dynamics in an animated ball mirrored those in music, even across cultures [68].

While SIRE has been tested for particular values of speed, intensity, irregularity and extent, it remains to be seen if the same results emerge with a large number of training samples, for example to account for the many different expressions of sadness. To address this statistical learning problem, we turn to modeling using probabilistic Gaussian Mixtures in the 4-D SIRE space.

The MEI module is composed of four GMM's in SIRE space, one representing each basic emotion (see Fig. 3). For each emotion class C of *happiness, sadness, anger* and *fear*, we define an m-mixture Gaussian in 4D SIRE space

$$SIRE\_Emotion_c(X_c) = \sum_{k=1}^{m} \pi_k \mathcal{N}(X_c|\mu_k, \sigma_k) \qquad (1)$$

where X_c is a vector of SIRE tuples corresponding to the class C, and m is the optimal number of components to minimize

TABLE I
SIRE PARAMETERS AND ASSOCIATED EMOTIONAL FEATURES FOR VOICE, GESTURE AND MUSIC BASED ON A LITERATURE SURVEY IN [39]

		Modality mappings to relevant emotional features		
Parameter	Description	Voice	Gesture	Music
Speed	slow vs. fast	speech rate [17], pauses [29]	velocity [45], animation [24], quantity of motion [12]	tempo [47] [43]
Intensity	gradual vs. abrupt	voice onset rapidity [29], articulation [17]	acceleration [45]	note attack[47], articulation [43]
IrRegularity	smooth vs. rough	jitter [29], high-freq. energy, voice quality [17] [29]	directness [45], phase shift [1] [57], fluidity [55]	microstructural irregularity [43], timbral roughness [47]
Extent	small vs. large	pitch range [17], loudness [29]	spatial expansiveness [24], contraction index [45]	volume [47] [43]

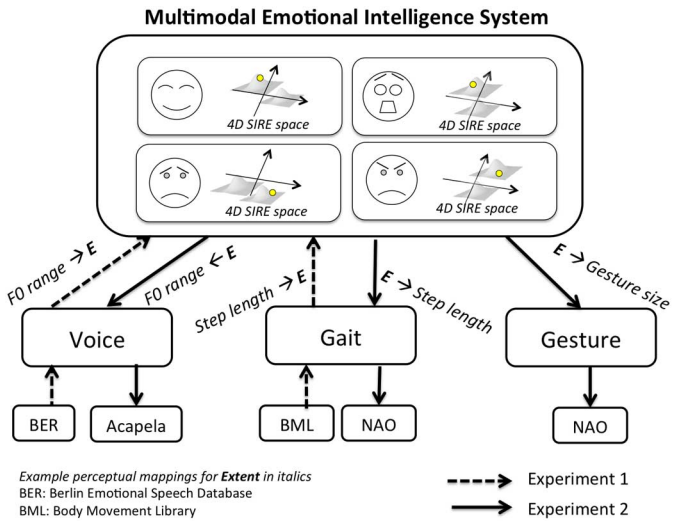

Fig. 3. Present system performs cross-modal recognition and expression based on a GMM representation. In Experiment 1, we test how well the model trained with emotional voice can recognize emotional gait. In Experiment 2, we use the model to generate emotional voice, gait and gesture.

the Bayesian information criterion (BIC) over X_c [66]. The above four emotion classes are the focus of the present study for two reasons. First, we study emotions verified in infants less than one year old [25]. At this point in development, infants do not yet have the notion of self and therefore the capacity for complex emotions such as embarrassment, pride, or guilt [36]. Second, we choose these emotions because they are the most commonly studied across our target modalities; it is relatively rare to find, for instance, music conveying disgust or surprise [29].

The Gaussian mixture model is selected for affect modeling because it proposes several advantages. First and most importantly, the GMM, as opposed to support vector machines [9], K-means [30], or linear regression models [6], can be used for both recognition and expression. For instance, a GMM trained on sad SIRE tuples can give the likelihood that a new, observed movement looks sad (Experiment 1 of this paper). And, a GMM trained on happy SIRE tuples can be sampled when the robot wishes to express joy (Experiment 2 of this paper), while avoiding repetitious values. Secondly, a GMM provides interpretability. Like prototype methods [71], we can inspect the means of the GMM to find the most prototypical set of parameters. For example, we can check whether the trained "fear" GMM components correspond to the anxious or terror fear found in psychology (as we shall see in Section VII). Or, we can see exactly how one emotion might different from another by comparing their means (e.g., elation differing from terror along the extent–but not speed–dimension.) Finally, having a GMM score for each emotion allows us to know relative emotional content. For instance, if an energetic vocal emotion sounds both happy and angry, the model should output high scores for these two emotions, and lower scores for sadness and fear. This could eventually be useful if the system is combined with another detector [e.g., a facial action coding system (FACS) detector [42] or contextual information] which could further differentiate between the top confusions.

III. TRAINING MEI

An overview of the training of the MEI module is given in Fig. 1. From emotional speech input, SIRE parameters are extracted and taken in conjunction with an emotional tag. These samples are used as training data for MEI.

In detail, we train MEI's happiness, sadness, anger and fear SIRE emotion models using three steps.

1) **Low-level feature extraction**. We select and extract low-level, modality-specific features representing Speed, Intensity, irRegularity, and Extent (SIRE). For example, *speech rate* in syllables per second is an indicator of speed in speech.
2) **Mapping samples to SIRE space**. We normalize each sample's four low-level features to [0,1] based on an individual's mean and standard deviation. This takes into account that individuals may have varying speaking styles, for example.
3) **Training** the models in SIRE space using expectation-maximization.

A. Low-Level Feature Extraction

In the SIRE paradigm, we select features that may perceptually be mapped to speed, intensity, irregularity and extent. These are dynamic features that are found as principal characteristics in emotion studies across voice [17], [22], music [47], [43], and motion [1], [55], [12]. For the purposes of this experiment, we selected the features in Table II to map voice and gait to SIRE parameters. In general, the maximum is selected because it has been shown to be highly relevant (more so than mean) in a cross-lingual recognition task [58]. We also examine samples

TABLE II
LOW-LEVEL FEATURE TO SIRE MAPPINGS

Voice feature	Parameter	Gait feature
Speech rate *(syllables/sec)*	Speed	Walking speed *(steps/min)*
Volume range *(dB)*	Intensity	Maximum foot acceleration *(cm/sec^2)*
High-frequency energy ratio *(dB)*	irRegularity	Step timing variance *(sec)*
Pitch range *(Hz)*	Extent	Maximum step length *(m)*

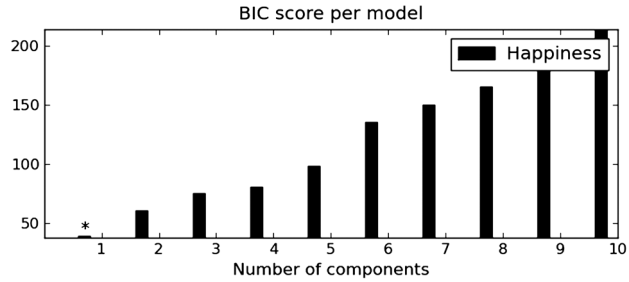

Fig. 4. Example of the system selecting a 1-component GMM to model the happiness dataset of the Berlin database used in Experiment 1.

with a maximum length of 15 seconds, to roughly parallel the length of short-term memory [56].

We use the Snack Toolkit[2] to extract the following features from a given recorded utterance. In future work, we plan to use the HARK robot audition system [50] for online extraction of features.

Speed: The number of syllables per second is calculated as the number of syllables divided by the number of seconds from the beginning to the end of an utterance's voiced segment.

Intensity: The intensity is the change in power (volume) in the voiced segment of the entire utterance, defined as maximum power subtracted by the minimum power (in dB).

Irregularity: This is defined as the utterance's average high frequency energy content (5–8 kHz) during the voiced segments, normalized frame-wise by power.

Extent: This is the utterance's pitch range, defined as the utterance's maximum F0 subtracted by the utterance's minimum F0.

We do not claim that these mappings are the optimal set, but rather show examples of sensory-specific mappings for these high-level perceptual features.

B. Mapping Samples to SIRE Space

How do we map real-world values to [0,1]? Our general idea is to take into account individual differences, so that any person (e.g. older or younger people, or generally fast or slow speakers) can still contribute to the emotion model. In this work, we used a very simple mapping method based on an individual's mean and variance, as described below. Nonlinear methods such as a logistic sigmoid function are likely more appropriate, however, and should be used in future work.

In this paper, we transform a datapoint by calculating its Z-score (standard score) relative to the mean and variance over an individual's dataset X_s. Since Z-scores fall between [-1,1] (i.e., a positive Z-score means the sample is greater than the dataset average, and a negative Z-score indicates the sample is less than the dataset average) the Z-scores are then shifted and scaled to [0,1]. Values less than -2σ or greater than 2σ are assigned to 0 and 1, respectively. Specifically

$$x_{s'} = \begin{cases} 0 & \text{if } x_s \leq -2\sigma \\ 1 & \text{if } x_s \geq 2\sigma \\ 0.5 + \frac{x_s - \mu(X_s)}{4 * \sigma(X_s)} & \text{otherwise,} \end{cases} \quad (2)$$

[2]http://www.speech.kth.se/snack/

where $\mu(X_s)$ and $\sigma(X_s)$ are the mean and variance of the speech rates for that individual. This transformation is defined in the same way for Intensity, Irregularity and Extent (see examples of usage in Fig. 3). In this way, we can ensure that "fast speech" ($x_{s'} \approx 1.0$) in a happiness sample, for example, is "fast" relative to that person's average speech rate, not an absolute definition of "fast."

C. Training

The above mapping procedure results in a multispeaker dataset X_C which contain SIRE values for an emotion class C (labeled a priori). We use this to train the corresponding $SIRE\_Emotion_c(X_C)$ GMM using expectation maximization [18]. In our experiments, the SciKit Learn Toolkit [54] is used to model and train each GMM, where the number of components is automatically selected by using the model with the lowest BIC score over a maximum of 10 components (see Fig. 4).

IV. RECOGNIZING EMOTIONS WITH MEI

We now describe how we can use MEI's voice-trained model to recognize emotion in a modality different from voice: human gait. We limit the scope to gait in this first attempt, but future work could be examined in two ways: 1) online perception; and 2) dynamics in other modalities such as gestures, music, facial features, and others suggested in [39].

A. Gait Feature Extraction

Gait studies such as [32], [59] analyze data from multiple participants walking in various emotional styles. They may take into account walker's posture, arm swing, speed, and may use measurement instruments such as force pads, motion capture, or a combination of both: Montepare [49] and Janssen [28] considered the force of the steps, and Unuma *et al.* [75] took into account step-length and hip position. Montepare [48] also found correlations between emotions and perceptual cues such as smooth-jerky, stiff-loose, expanded-contracted, and so on. Many cues have been found to be linked to emotion, and we attempt to extract the simplest, most important features.

To extract SIRE parameters, we consider the positions of feet through time. Our current study uses the Body Movement Library [44], which contains emotional walking by nonprofessionals, in neutral, happy, sad, angry, and a few samples of

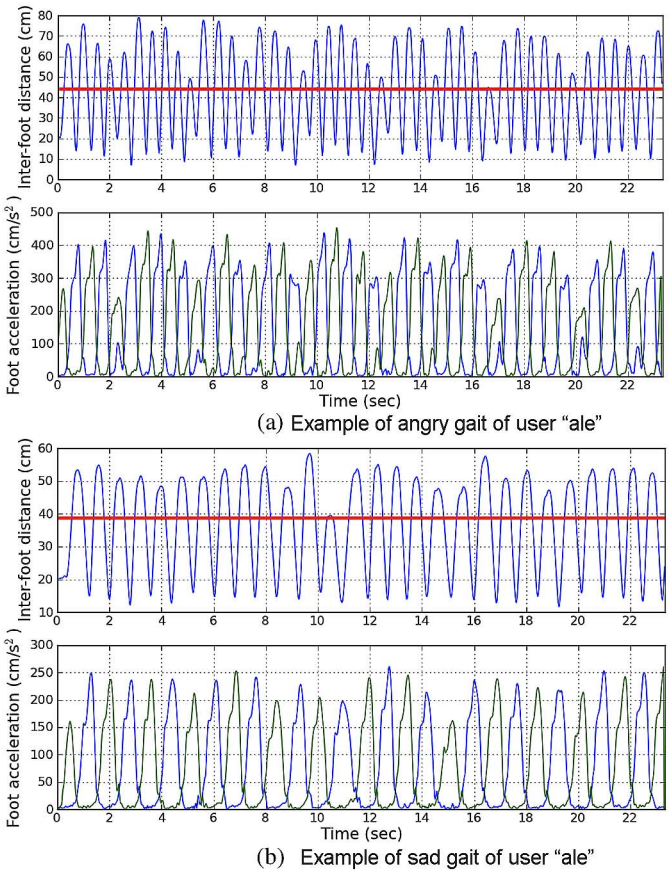

Fig. 5. Examples of gait analysis. The horizontal line indicates the threshold for peak-picking (mean value). For sad gaits, the step lengths (interfoot distances) are shorter, and foot acceleration is lower.

afraid. We use the data points of the ankle joints in x, y, z space, where z is the vertical axis.

Speed. We calculate speed in steps per minute. We subtract the position of one foot from the other in the horizontal (x, y) plane. We then perform peak picking (using average foot distance as the threshold, as in Fig. 5), assuming that feet are at their maximum horizontal distance when stepping. The centroids of these peaks determine the time of each step.

Intensity. We calculate the maximum acceleration achieved in the sample in x, y, z space. In a real-time situation, this may need to be used in conjunction with a sliding window. Intuitively, intensity corresponds to the "heavy-footedness" of the steps. In [4]'s emotion recognition approach for knocking movements, average acceleration was used. It's not clear whether one formulation over the other offers any advantage.

Irregularity. Step timing variance is calculated as the standard deviation in the step lengths, in seconds. For instance, walking with a "regular" pace may give a different impression compared to an "irregular" pacing which stops and starts.

Extent. This is the maximum step length in x, y space.

B. Mapping to SIRE Space and Personalization

After the features are extracted, the next step of mapping the features to SIRE space is performed identically to the procedure in III-B, using the new mean and standard deviations in the gait dataset.

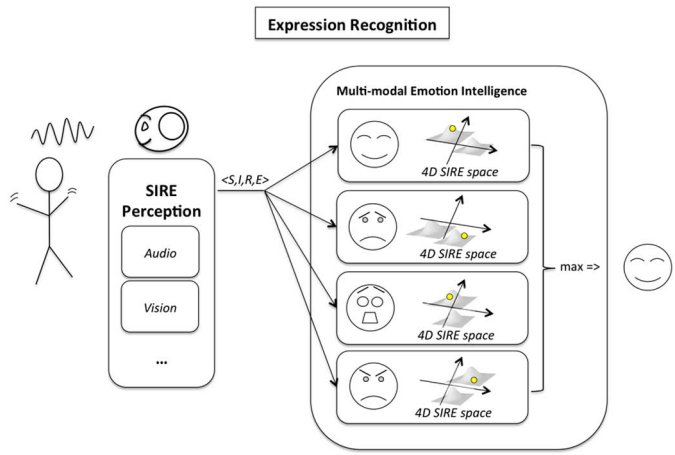

Fig. 6. Overview of how MEI can perform recognition. The SIRE perception module extracts S,I,R,E parameters through audio or video, and evaluates the SIRE tuple to find the most likely emotion being portrayed. In the present experiment, we use offline data from motion capture, but in previous work a Kinect has been used to perceive emotional motion [39].

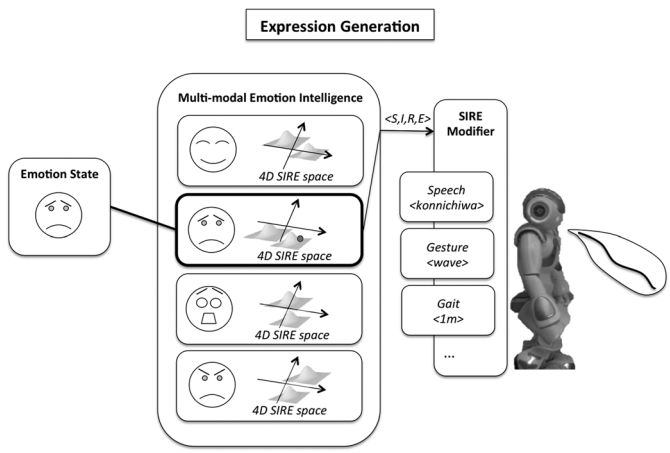

Fig. 7. Overview of how MEI is used to generate emotionally colored speech and movements on the robot. The desired emotional state is used to select the relevant class model, which is then sampled to generate a SIRE tuple. The tuple is used to modify the speed, intensity, irregularity, and extent of existing utterances and movements.

C. Recognizing Emotion in Gait

The emotion class of a given input SIRE vector X can be found simply by evaluating the sample in the Gaussian Mixture $SIRE\_Emotion_c(X)$ for each of the classes C, and selecting the class producing the maximum probability (see Fig. 6).

V. Generating Emotional Expression using MEI

It is straightforward to generate an emotional expression using MEI's SIRE Model (see Fig. 7). We first generate a SIRE tuple for a given emotion, then perform the mapping from the SIRE to the desired modalities. Generating emotional gesture from SIRE parameters has been explored in [39]. Here, we modulate the speech, gesture and gait on the NAO model robot from Aldebaran Robotics[3] (see Fig. 7).

[3]http://www.aldebaran-robotics.com

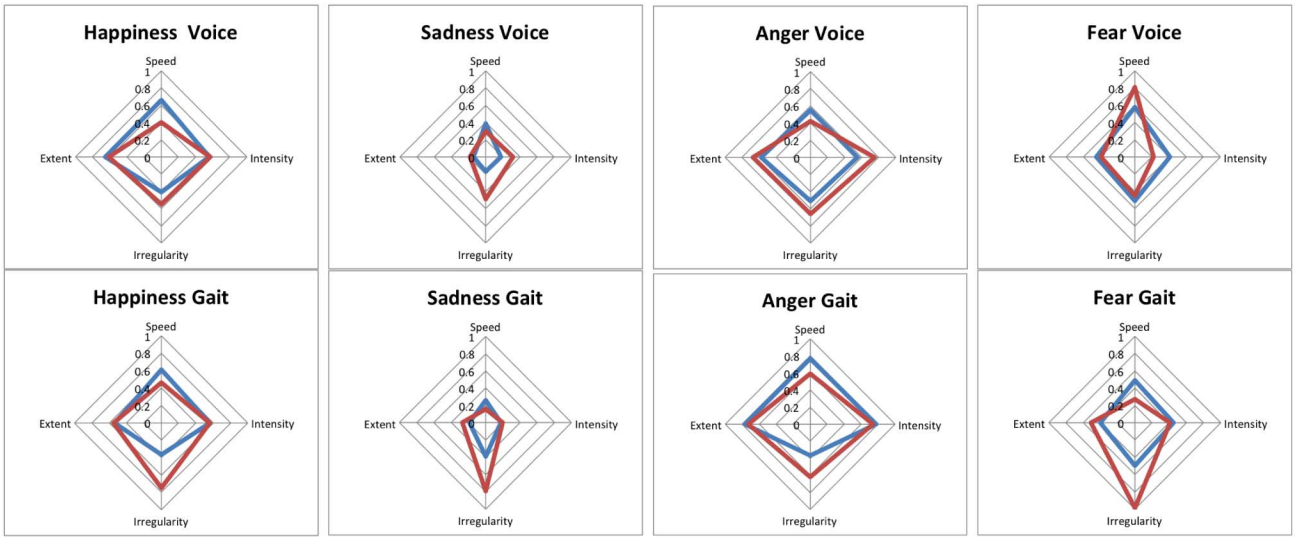

Fig. 8. Comparison of voice and gait means of GMMs trained with the full voice dataset (> 62 samples per emotion) and full gait dataset (> 42 samples per emotion). Red and blue lines correspond to the two 4-dimensional components per GMM, which were fixed at 2-components for visualization purposes. We can notice the similarity across voice and gait, with the exception of fear. This illustrates that the voice database likely contains "terror" fear samples, and the gait database primarily "anxious" fear samples [2].

A. MEI: Generating A SIRE Tuple

Given a desired emotion class C, we generate a SIRE tuple by sampling the appropriate Gaussian mixture $SIRE\_Emotion_c(X)$. Note that here we manually set the robot's emotion class (happiness, sadness, anger, or fear). How to automatically decide a "current emotional state" is complex and outside the scope of this paper. For more information, see for example [52] on deriving an emotional state based on cognitive appraisal of the robot's goals and surroundings. Additionally, this method does not allow for sampling an emotion subcategory (such as anxious fear) directly. This should be explored in future work.

B. SIRE Modifier: Mapping SIRE to Speech

The MEI robot speaks a string of Japanese syllables with no perceptible meaning, similar to infant-babbling. This choice of a human-incomprehensible language allows us to explore purely prosodic communication without any semantically charged meaning, similar to the developmental robotics work by Oudeyer [53]. We use the NAO's built-in Japanese TTS to generate an utterance W composed of words $[w_0, \ldots w_n]$, using Acapela[4] markup or the Aldebaran API to change the utterance's speed, intensity, irregularity, and extent.

Speed. We map W's relative speed linearly between 50% and 130% of the default rate.

Intensity. W's volume is modified by mapping I linearly between 0% and 100% of the maximum volume provided by the API.

irRegularity. We add pauses of length m after every word in W, where m is sampled randomly from a normal distribution with $\mu = 0$ and $\sigma = R/3$ seconds.

[4]http://www.acapela-group.com/

Extent. Given a pitch range between 90% and 140% of NAO's base pitch, we augment the pitch linearly by E for the first syllable of every word w, and set the other syllables to the minimum pitch 90%.

C. Mapping SIRE to Gesture

We use the same approach as in [39]. We also adjust the head of the robot such that Extent is mapped to the head. $E = 0$ is mapped to a downward-gazing head angle, and $E = 1$ mapped to a upward-gazing head angle, with linear interpolation in between. This follows the general SIRE design principle that higher values of extent for bodies should correspond to larger spatial expansion [39].

D. Mapping SIRE to Gait

We adjust parameters in the NAO Motion API, to modify using SIRE as follows:

Speed. S is mapped linearly to step frequency between 5% and 100% of the maximum speed provided by the API.

Intensity. I is mapped linearly to the height of the steps between 0.5 cm and 4 cm.

irRegularity. We calculate pauses of length m, where m is sampled randomly from a normal distribution with $\mu = 0$ and $\sigma = R$ seconds. The robot checks every 2s if $m > 1$, and if so stops for m seconds.

Extent. E is mapped linearly to the length of the steps between 3 and 8 cm.

It should be noted that automatic arm animations to match the rate of the walk are automatically added in Aldebaran NAO's default gait; these were not modified, with one exception. Hands were mapped in a similar manner to the head, with smaller values of E corresponding to a closed hand, and larger values of E for an open hand.

TABLE III
CROSS-MODAL RECOGNITION (BASELINE): RECOGNITION OF EMOTIONAL GAIT INPUT. A 4-CLASS MEI CLASSIFIER WAS TRAINED WITH RAW VOICE FEATURES AND TESTED RAW GAIT FEATURES (ACCURACY: 25%)

Detected Input	Happiness (%)	Sadness (%)	Anger (%)	Neutral (%)	*p-value*
Happiness	**100**	0	0	0	<0.0001
Sadness	99	0	**1**	0	<0.0001
Anger	**100**	0	0	0	<0.0001
Neutral	**100**	0	0	0	<0.0001

VI. EXPERIMENT 1: CROSS-MODAL EMOTION RECOGNITION

A. Purpose

Cross-language emotion recognition has been explored with limited success [58] (65%–72% accuracy), but to our knowledge, cross-modal emotion recognition has never been performed. In this experiment, we test whether MEI can be trained with voice and then recognize emotional gait. This simulates the situation where a robot encounters a modality it has never seen before.

B. Materials and Procedure

As training data, we used German utterances from 10 subjects (5 female, 5 male) from the Berlin emotional speech (Emo-DB) database [10]. This database is suitable because: 1) acted emotional utterances, like ID speech, are known to be more "full-blown" than everyday speech [10]; and 2) due to Emo-DB's widespread availability and use in many emotional voice studies (e.g., [67]), its use will facilitate follow-up experiments. Up to ten different sentences in four styles were used: happy (71 samples), sad (62 samples), angry (127 samples), and fear (69 samples). For all samples, the recognition rate by German-speaking adults was at least 80%. We used this data to train MEI's four SIRE emotion models as described in Section III.

As test data, we used foot motion capture data from 28 subjects from the Body Movement Library [44]. Each individual provided two 30 second samples of expressive walking per emotion class, except for fear which had fewer samples. For this experiment, each sample was split into 8 second segments, for a total of 168 happiness, 168 sadness, 168 anger, and 42 fear samples. Note that only the ankle joint data was used; leg, body, and posture data were not used at all.

VII. RESULTS AND DISCUSSION

How well can MEI recognize emotion in a new context: gait? In Tables III-VI, we show the results of recognizing emotional gait samples of happiness, sadness, anger and fear. *P-values* were calculated using the chi-square test with a null hypothesis of a uniform distribution over the four categories. As a baseline, Table III illustrates that cross-modal recognition is not possible with the standard low-level feature approach: training in one modality (voice) and testing in another (gait) results in chance level recognition.

Using our SIRE paradigm, we can see that the overall cross-modal recognition rate is 63%, without using any data from the target modality (see Table IV). Happiness, sadness, and fear were recognized at significant levels, though anger was sometimes confused with happiness (discussed later in this section.) In fact, training with emotional voice gives almost comparable results to intramodal training, that is, training and testing with emotional gait data. As an upper-bound, we compare our cross-modal result to the recognition rate when gait information is available: 72% in [31] and 75% here (Tables VI and V). This suggests that cross-modal recognition can be achieved by first abstracting data features to a higher-level perceptual space, such as SIRE.

TABLE IV
CROSS-MODAL RECOGNITION (OUR METHOD): RECOGNITION OF EMOTIONAL GAIT INPUT. A 4-CLASS MEI CLASSIFIER WAS TRAINED WITH VOICE SAMPLES IN SIRE SPACE AND TESTED RAW GAIT SAMPLES IN SIRE SPACE (ACCURACY: 63%)

Detected Input	Happiness (%)	Sadness (%)	Anger (%)	Fear (%)	*p-value*
Happiness	**62**	0	19	19	<0.0001
Sadness	2	**90**	0	6	<0.0001
Anger	**55**	0	43	2	<0.0001
Fear	21	12	12	**55**	<0.0001

TABLE V
INTRA-MODAL RECOGNITION (OUR METHOD): RECOGNITION OF EMOTIONAL GAIT INPUT. TRAINING AND TESTING IS PERFORMED USING GAIT SAMPLES IN SIRE SPACE, IN OPEN TESTS (ACCURACY: 75%)

Detected Input	Happiness (%)	Sadness (%)	Anger (%)	Fear (%)	*p-value*
Happiness	**70**	0	5	25	<0.0001
Sadness	0	**80**	0	20	<0.0001
Anger	20	0	**80**	0	<0.0001
Fear	30	0	0	**70**	<0.0001

TABLE VI
INTRA-MODAL RECOGNITION (EIGENWALKERS METHOD [31]): RECOGNITION OF EMOTIONAL GAIT INPUT TRAINED IN 20 DIMENSIONS (ACCURACY: 72%)

Detected Input	Happiness (%)	Sadness (%)	Anger (%)	Neutral (%)	*p-value*
Happiness	**76**	3	14	7	<0.0001
Sadness	10	**76**	7	7	<0.0001
Anger	21	7	**69**	3	<0.0001
Neutral	17	3	10	**69**	<0.0001

This result is also comparable to human performance. Consider that human emotion recognition in a new context is also low: in [65], participants from nine countries and three continents rated emotional German samples over five emotions. The recognition accuracy ranged from a maximum of 74% by native Germans participants, to 52% by Indonesian participants. There was variability between emotions, too; the Dutch participants rated, for example, German Joy portrayals with an accuracy of 30%.

Next, we analyze the structure of the voice and gait GMMs, and give possible explanations for confusions. In Fig. 8, two-component GMMs are plotted for both voice and gait for ease of comparison.

In Table IV, we see that fear in gait was not as well recognized as happiness or sadness. One explanation could be that the voice

training dataset may have contained almost uniquely "terror" fear, and the gait dataset mostly "anxious" fear. The dynamics of these two subtypes of fear have been shown to differ greatly [2]. Indeed, upon comparing the voice means and gait means (see Fig. 8), it appears that the voice dataset contained fast (terrified) voices, while the gait dataset contained slow, irregular (anxious) walks. According to [10], the voice actors were asked not to whisper when producing fear utterances, whereas whispering may be necessary to produce "anxious fear" in voice. This suggests that recognition rates may improve by adding samples of slower, "anxious" voice to our training database.

Next, anger was most often recognized as happiness. Upon inspection of angry gait misclassifications, MEI consistently output high probabilities of both anger and happiness. Why confusion with happiness? According to an experiment with human evaluators of voice data in [2], "elation was relatively often confused with despair, hot anger, and panic fear, which differ strongly in quality but are similar in intensity." Inspection of Fig. 8 supports this; we can notice that the dynamics of anger and happiness are relatively similar. How to overcome this confusion must be examined in future work, for example by including another modality such as face to overcome the difference in valence.

VIII. Experiment 2: Cross-modal Emotion Expression

A. Purpose

Our goal is to test whether a robot trained with emotional voice can express emotions through speaking, gesturing and walking, as shown in Fig. 7). Expression itself is a particularly difficult challenge, because the robot: a) does not use an expressive face (as in [8], [79]); b) does not use any custom emotion animations (such as weeping for sadness) [3]; and c) does not use hand-defined parameters to control its movement [41]. Importantly, we are also testing whether emotion parameters learned from voice data could be a basis for expression in multiple modalities.

B. Materials and Procedure

We first outline the many design considerations for a human evaluation of emotion, especially in humanoid robots. Firstly, we must remember that many cues may interfere with emotional expression because the humanoid form is already socially charged. For instance, a robot speaking happily with a stationary body can be confusing for observers: a robot with an immobile head was suggested to look angry (as if staring) in [38]. Similarly, looking away can also express embarrassment or social disinterest, according to gaze studies [14]. Closed hands may look like angry fists or have other cultural meanings. The appearance of the robot itself, with an infant-like size or bold color could implicitly play a role in the perception of personality or stereotypical emotions. In implementation, motor noise can also have unintended effects (such as "sad sounds" in [38]) and even a one second latency could imply negative hesitation. There is a plethora of cues to consider, so we do our best effort to control for these parameters; we use a neutral grey-colored instead of orange NAO, omit heavy processing for a real-time response, and try to use semantically-ambiguous gestures. Secondly, whereas

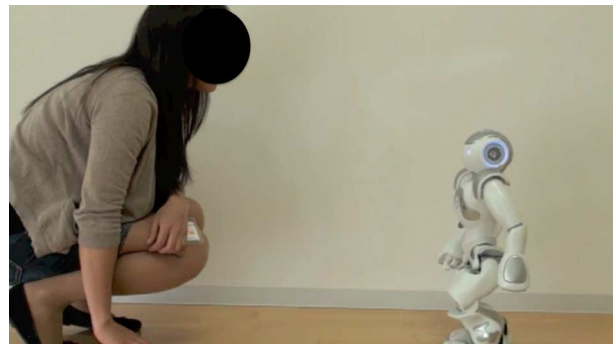

Fig. 9. Stimulus used in Experiment 2 of robot interacting with human with various emotions. The robot spoke, gestured, then walked toward the human in all stimuli.

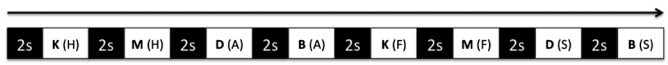

Fig. 10. Order of presented stimuli for all subjects. The letter in bold corresponds to the interaction utterances: K–Konnichiwa, M–Mite, D–Dame, B–Baibai. The letter in parentheses is the robot's emotional SIRE modification: H–Happiness, S–Sadness, A–Anger, and F–Fear.

many studies test emotional expression in a independent context, humans use many cues, including social context, to decide the emotion of a person. For instance, [15] showed that full-body point-light displays of humans expressing love and joy were understood when presented in a two person context, but not when shown alone.

For these reasons, and also due to the fact that the platform is small and child-like, we design our experiment to evaluate the MEI and SIRE paradigm using a short but realistic progression in an adult-child interaction: 1) a greeting, 2) showing a toy, 3) revoking the toy, and 4) saying goodbye. We filmed a woman speaking in Japanese to a white and grey NAO robot controlled with MEI (Fig. 9). Four interactions were created:

1) the human said "Konnichiwa" (Hello) while waving at the robot;
2) the human said "Mite" (Look) and held out a toy;
3) the human said "Dame" (No) and clasped the toy in the direction away from the robot;
4) the human said "Baibai" (Bye bye) while waving.

The robot responded in SIRE-modified nonsense words with accompanying gesture as described in Table VII, then walked toward the human. The gesture contained 2 movements used in [38], starting with the robot's hands close together: 1) the hands moved apart to either side; and 2) one hand moved upwards and the other moved downwards. As shown in Table VII, for each interaction, the robot's speech, gesture and gait were subject to one of two emotional modifications, depending on the stimuli (see Fig. 10). The emotional responses were chosen to emulate typical social responses of a child to the situations. For example, a child meeting a person may be happy to see them or afraid.

We created a video containing a total of 8 different scenes, comprised of two sets of the four interactions as shown in Fig. 10, separated by 2 second black frames, for a total of 2 min 10 s. In the first set of four, we chose a progression of happy and angry robot emotional reactions to portray an "outgoing" robot. In the last set of four, we chose the remaining emotions

TABLE VII
INTERACTIONS BETWEEN HUMAN AND ROBOT, AND SIRE MODIFICATIONS USED IN EXPERIMENT 2 (JP: JAPANESE LANGUAGE)

Human utterance in JP	Robot response in nonsense syllables	SIRE Modification
Konnichiwa *(Hello)*	Bama mufe ikefu	Happiness, fear
Mite *(Look)*	Bifu buse bamasu	Happiness, fear
Dame *(No)*	Bamasu muhe bushibe	Anger, sadness
Baibai *(Bye bye)*	Bama muse nojebu	Anger, sadness

TABLE VIII
OUR EXPECTED PAD VALUES FOR HAPPINESS, SADNESS, ANGER AND FEAR PORTRAYALS IN EXPERIMENT 2, BASED ON EMOTION TERMS PROVIDED IN [46]

P	A	D	Emotion terms from [46]	This study
+	+	+	Bold, excited, triumphant	Happiness
+	+	-	Fascinated, amazed, respectful	
+	-	+	At ease, relaxed, unperturbed	
+	-	-	Docile, protected, sleepy	
-	+	+	Angry, defiant, hostile	Anger
-	+	-	Aghast, distressed, insecure	Fear
-	-	+	Disdainful, uncaring, unconcerned	
-	-	-	Despairing, lonely, sad	Sadness

of sadness and fear, portraying a "reserved" robot. We chose to use these logical progressions because pilot trials with a random order showed that users were perturbed by the robot showing wildly varying and inconsistent "personalities".

For the experiment, the robot's MEI module generated the following SIRE parameters, which we held constant throughout the experiment:

- happiness: [0.713, 0.552, 0.422, 0.630];
- sadness: [0.112, 0.307, 0.816, 0.195];
- fear: [0.912, 0.465, 0.205, 0.351];
- anger: [0.157, 0.946, 0.198, 0.459].

We recruited 20 Japanese-speaking participants (6 female) to view the stimulus video and rate the robot's emotional expression. The users were given a modified version of the self-assessment manikin (SAM) Measurement Scale for Japanese called REM [35] to rate the pleasure, arousal and dominance (PAD) of the robot in each scene [46].

In Table VIII, we show the expected positive/negative PAD values for the 4 emotion classes used in our study. We used the table from [46], which provides PAD permutations and associated emotional tags. For example, we expect that a robot with happiness SIRE modifications using our MEI should result in positive P, A, and D values (assuming that "excited, triumphant" are near adjectives to happiness), and so on. PAD is expected to be more useful than simple emotion categorization because PAD can provide both an emotional category and explanation for that choice. For instance, PAD could be useful to see that an expression was not recognized because of a missing pleasure component.

The procedure was as follows:

1) the participant read an introduction of the robot which described it as speaking a nonsense language;
2) the participant watched the video once on a laptop with external speakers, in the order of Fig. 10;
3) the participant watched the video again and chose how they believed the robot felt during the scene, on each of the PAD scales. He/she was given as much time as needed after each scene before proceeding to the next.

C. Results and Discussion

We compare the average ratings for each scene with the expected PAD result. For happiness, we expect $+\mathrm{P}$, $+\mathrm{A}$, $+\mathrm{D}$ ratings, and for sadness, we expect $-\mathrm{P}$, $-\mathrm{A}$, $-\mathrm{D}$. Anger portrayals are expected to give $-\mathrm{P}$, $+\mathrm{A}$, $+\mathrm{D}$, and fear is expected as $-\mathrm{P}$, $+\mathrm{A}$, $-\mathrm{D}$.

According to the ratings shown in Fig. 11, happiness and sadness were well expressed. We find that the portrayals of happiness had $+\mathrm{P}$, $+\mathrm{A}$, $+\mathrm{D}$, (.53,.44,.28) and (.48,.46,.24). Both portrayals of sadness also were shown to have $-\mathrm{P}$, $-\mathrm{A}$, $-\mathrm{D}$, $(-.66, -.32, -.28)$ and $(-.58, -.46, -.3)$. Importantly, these portrayals are not confused with other emotions. For instance, happiness is not confused with anger nor fear, other emotions with relatively high dynamics.

Fear, which Mehrabian defines as $-\mathrm{P}, +\mathrm{A}, -\mathrm{D}$, was not well captured in our scenarios. The assessments as $+\mathrm{P}, +\mathrm{A}, +/-\mathrm{D}$ show that they were somewhat confused with happiness, with a positive pleasure component (though not as high as the happiness portrayals). The explanation for this may stem from the fact that, over all conditions, the robot was shown to be approaching the human, whereas fear is an avoidance behavior [60]. Indeed, based on our data analysis in Experiment 1, the original voice samples appear to contain terror fear, resulting in MEI-controlled gestures of the robot were fast and jerky, yet the robot moved at a fast rate (with small steps) *toward* the human. Subjective reports are consistent with this: when participants were told that the target emotion was fear, some stated that the robot moving towards the object was incongruent. This suggests that in future work, a "direction" parameter should be added in an embodied robot situation.

Raters also found difficulty in assessing the angry expressions. Mehrabian defines anger as $-\mathrm{P}$, $+\mathrm{A}$, $+\mathrm{D}$, but participants rated the expressions as $-\mathrm{P}$, $+\mathrm{A}$, $-\mathrm{D}$, a difference in the dominance dimension which suggests the raters tended to confuse the anger portrayals as slightly fearful. Whereas anger is characterized by a high dominance component, the robot was rated to be slightly submissive ($\mathrm{D} = -0.15$). In examining the SIRE values produced by MEI, the values appear to characterize irritation (cold anger), i.e., with a low speed and high intensity. In the future, similar to fear, we may also need to explore either producing rage (hot anger) towards the person/object, or cold anger away. Another direction for future work is to notice that the robot was only slightly submissive-looking, at $\mathrm{D} = -0.15$, compared to sadness, which was more submissive at $\mathrm{D} = -0.3$. It may be interesting to check the effect of a robot's relative size; with a robot of equal or larger size to a human, the dominance dimension may possibly be pushed to $+\mathrm{D}$, making the robot look angry using our technique.

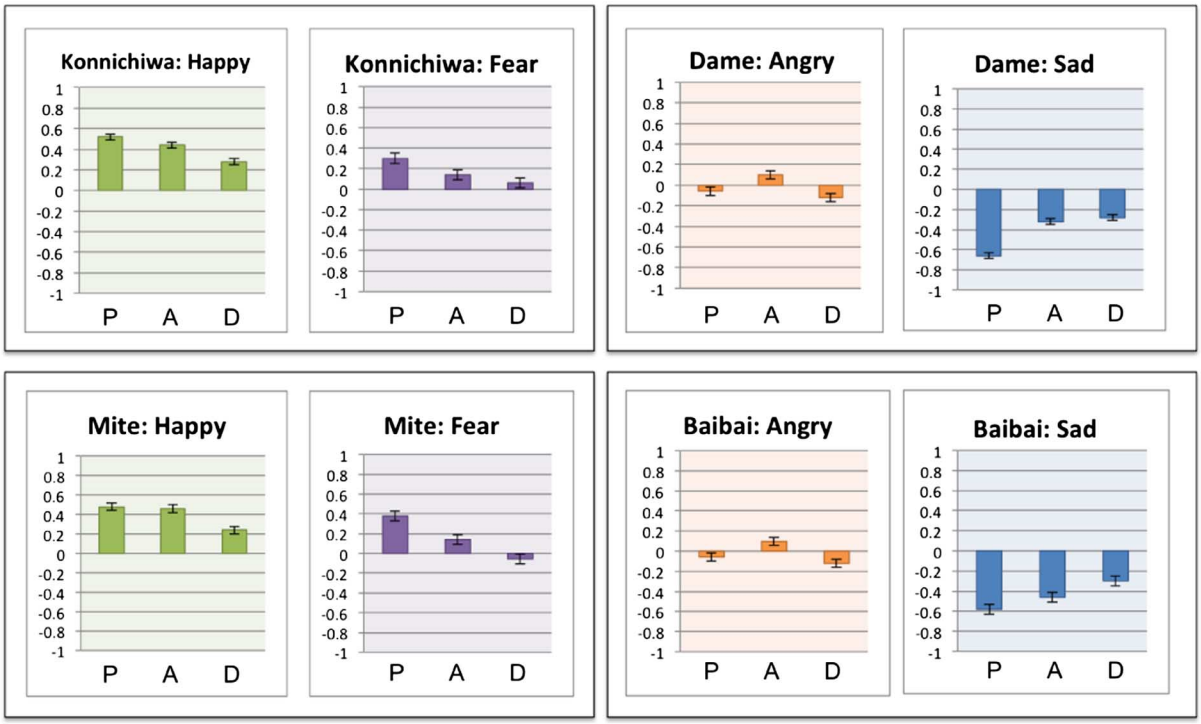

Fig. 11. Results of user evaluations, where P = pleasure, A = arousal, D = dominance. Happy and sad emotional expressions conform to expected values PAD values from [46]. We can also note that fear was perceived to have less dominance than happy, but the pleasure component was not dropped as expected. The angry and sad dyads were easily distinguished from each other, though dominance in anger was not greater than 0 as expected.

There are limitations to these results. Firstly, Experiment 2 cannot ascertain how much voice, gesture or gait contributed each to the overall impression of the robot. Ideally, a similar experiment could be run without speech, gesture or gait respectively, keeping in mind that a lack of speech (silence) may also convey negative emotions. Theories for how humans and robots develop the ability to perceive speed, intensity, irregularity and extent across modalities should also be investigated [72].

IX. Conclusion and Future Work

In this paper, we proposed a robot named MEI, which could develop emotional expressions through a universal parent-infant interaction called motherese. By associating an emotional face ground truth with vocal dynamics, it could develop its MEI: the ability to recognize and produce emotion in multiple modalities, such as voice, gesture, gait, or music. Although very simple, the MEI had three significant characteristics: a recognition ability, an interpretable model, and an expression ability. We implemented MEI in an Aldebaran NAO model robot, and performed two experiments to test our hypothesis.

The first experiment was a cross-modal emotion recognition task. Our goal was to check whether the voice-trained MEI was powerful enough to recognize emotion in a completely new modality. To verify this, we trained the MEI with emotional voice, and attempted to use the same MEI to recognize emotional gait. We found that it achieved 63% cross-modal recognition accuracy, which is significant compared to the baseline of 25% using raw features, and an upper-bound of 75% intramodal accuracy. This result was achieved by mapping both modalities to a common perceptual space called SIRE (speed, intensity, irregularity and extent). It also suggests a promising approach for the highly sought-after generalization ability in artificial intelligence: by abstracting low-level features to a common high-level perceptual space, it is possible for a classifier to generalize to a new context.

The second experiment was an emotion production task. Our hypothesis was that the voice-trained MEI could also provide a basis for expressing emotional speech, gesture and gait. To test this, we asked participants to deduce the emotion of a MEI-controlled NAO robot speaking, gesturing and walking towards a human. The results show that the robot could reliably portray expressions of happiness and sadness. For robot fear and anger, we suggested that an additional parameter–direction *away* to show avoidance or approaching–could be added to increase understanding. This provides a practical result: a data-driven approach for generating long-term, continuous emotional expression for robots, without relying on a moveable face, custom animations, or hand-tweaked parameters. Although only verified here for happiness and sadness, these are perhaps the most important emotions for a robot to express in a human-robot interaction (happiness to cheer up humans, and sadness to express empathy or remorse for a mistake). In addition, it shows the effectiveness of integrating controllers for multiple modalities. The multiplication of specialized systems is not scalable for autonomous robots, and this paper contributes a way to simplify multiple expression systems into one.

Together, this paper provides a theoretical, scientific result for emotion learning. It is not clear how humans "develop" emotion expression and recognition, though it is known that there is

rapid progress in these abilities in the first year of human life. This paper contributes an explicit theory for robot development of emotional intelligence through expressive facial and vocal interaction with a caregiver. This is a new idea that should be investigated for human infant development as well, with possible implications in autism treatment or other delays in emotion understanding and expression.

In the future, a promising direction for investigation is integration with the face or with contextual information. It has been shown in psychology that, whereas voice provides activation and is weak for valence [61], face readily provides valence information. As previously suggested, face information could be used along with MEI output scores, to distinguish, for example between confusions of happiness and anger. In terms of expression, colored eye LEDs or semantically charged words could contribute to the expressions of anger or fear. It would also be greatly interesting to test whether visual face information could be used to further improve or expand the emotional repertoire, for instance to express complex emotions such as pride or embarrassment, or even sarcasm. Furthermore, we could take into account the context of the motherese interaction, such as the current goal or internal state. As an example, imagine an infant being scolded by a parent with an angry voice and angry face. It has likely had a goal that has been stopped (e.g., reaching for an electrical outlet). Later in life, a child or adult whose goal has been thwarted may express that which he has associated with such a situation: an angry face and an angry voice.

Finally, future work may consider the relation between emotion and language development, as emotion processes appear before language is acquired. Emotional vocalizations are distinguishable in the 5th or 7th month [63]. Yet, only around the age of three do babies acquire the ability to speak full sentences [34]. A recent literature review by Saint-Georges *et al.* [62] describes the nature of motherese and its links to emotion, cognition and language development: roboticists could use this is a guide for considering emotional processes as a basis for developing language and meaning.

REFERENCES

[1] K. Amaya and A. Bruderlin *et al.*, "Emotion from motion," *Graphics Interface*, vol. 96, pp. 222–229, 1996.

[2] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Personality Soc. Psychol.*, vol. 70, no. 3, p. 614, 1996.

[3] A. Beck and A. Hiolle *et al.*, "Interpretation of emotional body language displayed by robots," in *Proc. 3rd Int. Workshop Affective Interaction Natural Environments*, 2010, pp. 37–42, ACM.

[4] D. Bernhardt and P. Robinson, "Detecting emotions from connected action sequences," in *IVIC '09*. Berlin, Germany: Springer-Verlag, 2009, pp. 1–11.

[5] M. Bhaykar and J. Yadav *et al.*, "Speaker dependent, speaker independent, and cross language emotion recognition from speech using gmm, and hmm," in *NCC*. Piscataway, NJ, USA: IEEE, 2013, pp. 1–5.

[6] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer-Verlag, 2006, vol. 1.

[7] S. Boucenna and P. Gaussier *et al.*, "Imitation as a communication tool for online facial expression learning, and recognition," in *Proc. IROS*, 2010, pp. 5323–5328, IEEE.

[8] C. L. Breazeal, *Designing Sociable Robots*. Cambridge, MA, USA: MIT press, 2004.

[9] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[10] F. Burkhardt and A. Paeschke *et al.*, "A database of german emotional speech," in *Interspeech*, 2005, pp. 1517–1520.

[11] J. J. Campos and D. I. Anderson *et al.*, "Travel broadens the mind," *Infancy*, vol. 1, no. 2, pp. 149–219, 2000.

[12] A. Camurri and G. Volpe *et al.*, "Communicating expressiveness and affect in multimodal interactive systems," *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 43–53, Jan. 2005.

[13] G. Castellano and S. D. Villalba *et al.*, "Recognising human emotions from body movement, and gesture dynamics," in *Proc. Affective Comput. Intell. Interaction*, 2007, pp. 71–82, Springer.

[14] J. J. Choi and Y. Kim *et al.*, "Have you ever lied?: The impacts of gaze avoidance on people's perception of a robot," in *HRI*. Piscataway, NJ, USA: IEEE Press, 2013, pp. 105–106.

[15] T. J. Clarke and M. F. Bradshaw *et al.*, "The perception of emotion from body movement in point-light displays of interpersonal dialogue," *Perception*, vol. 34, no. 10, pp. 1171–1180, 2005.

[16] M. Clynes, *Sentics: The touch of emotions*. New York, NY, USA: Anchor Press Garden, 1977.

[17] R. Cowie and E. Douglas-Cowie *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.

[18] A. P. Dempster and N. M. Laird *et al.*, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy. Statist. Soc Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[19] A. Fernald, "Intonation and communicative intent in mothers' speech to infants: Is the melody the message?," *Child Develop.*, vol. 60, no. 6, pp. 1497–1510, 1989.

[20] A. Fernald, "Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages," *Child Develop.*, vol. 64, no. 3, pp. 657–674, 1993.

[21] A. Fernald and T. Taeschner *et al.*, "A cross-language study of prosodic modifications in mothers and fathers speech to preverbal infants," *J. Child Lang.*, vol. 16, no. 3, pp. 477–501, 1989.

[22] R. Fernandez and R. W. Picard, "Classical and novel discriminant features for affect recognition from speech," in *Proc. Interspeech*, 2005, pp. 473–476.

[23] R. Flom and D. A. Gentile *et al.*, "Infants discrimination of happy and sad music," *Infant Behav. Develop.*, vol. 31, no. 4, pp. 716–728, 2008.

[24] P. E. Gallaher, "Individual differences in nonverbal behavior: Dimensions of style," *J. Personality Soc. Psychol.*, vol. 63, no. 1, p. 133, 1992.

[25] T. Grossmann, "The development of emotion perception in face and voice during infancy," *Restorative Neurol. Neurosci.*, vol. 28, no. 2, pp. 219–236, 2010.

[26] T. Grossmann and T. Striano *et al.*, "Crossmodal integration of emotional information from face and voice in the infant brain," *Develop. Sci.*, vol. 9, no. 3, pp. 309–315, 2006.

[27] H. Gunes and B. Schuller *et al.*, "Emotion representation, analysis, and synthesis in continuous space: A survey," in *FG*. Piscataway, NJ, USA: IEEE, 2011, pp. 827–834.

[28] D. Janssen and W. I. Schöllhorn *et al.*, "Recognition of emotions in gait patterns by means of artificial neural nets," *J. Nonverbal Behav.*, vol. 32, no. 2, pp. 79–92, 2008.

[29] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," *Psychol. Bulletin*, vol. 129, no. 5, p. 770, 2003.

[30] T. Kanungo and D. M. Mount *et al.*, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.

[31] M. Karg and R. Jenke *et al.*, "A two-fold pca-approach for inter-individual recognition of emotions in natural walking," in *Proc. MLDM Posters*, 2009, pp. 51–61.

[32] M. Karg and K. Kuhnlenz *et al.*, "Recognition of affect based on gait patterns," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 1050–1061, Apr. 2010.

[33] Y. E. Kim and E. M. Schmidt *et al.*, "Music emotion recognition: A state of the art review," in *Proc. ISMIR*, 2010, pp. 255–266.

[34] P. K. Kuhl, "Early language acquisition: Cracking the speech code," *Nature Rev. Neurosci.*, vol. 5, no. 11, pp. 831–843, 2004.

[35] K. Kuji, "Applicability of "rem", revised emotions measurement, to marketing analysis," *Adv. Consumer Studies*, vol. 15, pp. 57–76, 2009.

[36] M. Lewis, "Self-conscious emotions," *Handbook of Emotions*, vol. 2, pp. 623–636, 2000.

[37] M. M. Lewis, *Infant speech; a study of the beginnings of language*. London, U.K.: Routledge & Kegan Paul, 1936.

[38] A. Lim and T. Ogata *et al.*, "Converting emotional voice to motion for robot telepresence," in *Humanoids*. Piscataway, NJ, USA: IEEE, 2011, pp. 472–479.

[39] A. Lim and T. Ogata *et al.*, "Towards expressive musical robots: A cross-modal framework for emotional gesture, voice and music," *EURASIP J. Audio, Speech, Music Process.*, vol. 2012, no. 1, pp. 1–12, 2012.

[40] A. Lim and H. G. Okuno, "Using speech data to recognize emotion in human gait," in *HBU*. Berlin, Germany: Springer-Verlag, 2012, pp. 52–64.
[41] H.-o. Lim and A. Ishii *et al.*, "Emotion-based biped walking," *Robotica*, vol. 22, no. 5, pp. 577–586, 2004.
[42] G. Littlewort and J. Whitehill *et al.*, "The computer expression recognition toolbox (cert)," in *FG*. Piscataway, NJ, USA: IEEE, 2011, pp. 298–305.
[43] S. R. Livingstone and R. Muhlberger *et al.*, "Changing musical emotion: A computational rule system for modifying score and performance," *Computer Music J.*, vol. 34, no. 1, pp. 41–64, 2010.
[44] Y. Ma and H. M. Paterson *et al.*, "A motion capture library for the study of identity, gender, and emotion perception from biological motion," *Behav. Res. Methods*, vol. 38, no. 1, pp. 134–141, 2006.
[45] M. Mancini and G. Castellano, "Real-time analysis, and synthesis of emotional gesture expressivity," in *Proc. Doctoral Consortium Int. Conf. Affective Comput. Intell. Interaction*, 2007.
[46] A. Mehrabian, "Framework for a comprehensive description and measurement of emotional states," *Genetic, Soc., General Psychol. Monographs*, vol. 121, no. 3, pp. 339–361, 1995.
[47] L. Mion and G. De Poli, "Score-independent audio features for description of music expression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 458–466, 2008.
[48] J. Montepare and E. Koff *et al.*, "The use of body movements and gestures as cues to emotions in younger and older adults," *J. Nonverbal Behav.*, vol. 23, no. 2, pp. 133–152, 1999.
[49] J. M. Montepare and S. B. Goldstein *et al.*, "The identification of emotions from gait information," *J. Nonverbal Behav.*, vol. 11, no. 1, pp. 33–42, 1987.
[50] K. Nakadai and T. Takahashi *et al.*, "Design and implementation of robot audition system'hark' open source software for listening to three simultaneous speakers," *Adv. Robot.*, vol. 24, no. 5–6, pp. 739–761, 2010.
[51] D. G. K. Nelson and K. Hirsh-Pasek *et al.*, "How the prosodic cues in motherese might assist language learning," *J. Child Lang.*, vol. 16, no. 01, pp. 55–68, 1989.
[52] A. Ortony, *The cognitive structure of emotions*. Cambridge, MA, USA: Cambridge Univ. Press, 1990.
[53] P.-y. Oudeyer, "The synthesis of cartoon emotional speech," presented at the Speech Prosody 2002 Int. Conf., 2002.
[54] F. Pedregosa and G. Varoquaux *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[55] C. Pelachaud, "Studies on gesture expressivity for a virtual agent," *Speech Commun.*, vol. 51, no. 7, pp. 630–639, 2009.
[56] L. Peterson and M. J. Peterson, "Short-term retention of individual verbal items," *J. Exp. Psychol.*, vol. 58, no. 3, p. 193, 1959.
[57] F. E. Pollick and H. M. Paterson *et al.*, "Perceiving affect from arm movement," *Cognition*, vol. 82, no. 2, pp. B51–B61, 2001.
[58] T. Polzehl and A. Schmitt *et al.*, "Approaching multi-lingual emotion recognition from speech-on language dependency of acoustic/prosodic features for anger detection," in *Proc. Speech Prosody*, 2010.
[59] C. L. Roether and L. Omlor *et al.*, "Critical features for the perception of emotion from gait," *J. Vision*, vol. 9, no. 6, 2009.
[60] E. T. Rolls, "Precis of the brain and emotion," *Behav. Brain Sci.*, vol. 23, no. 2, pp. 177–191, 2000.
[61] J. A. Russell and J.-A. Bachorowski *et al.*, "Facial and vocal expressions of emotion," *Annu. Rev. Psychol.*, vol. 54, no. 1, pp. 329–349, 2003.
[62] C. Saint-Georges and M. Chetouani *et al.*, "Motherese in interaction: At the cross-road of emotion and cognition?(a systematic review)," *PloS One*, vol. 8, no. 10, p. e78103, 2013.
[63] E. Scheiner and K. Hammerschmidt *et al.*, "Acoustic analyses of developmental changes and emotional expression in the preverbal vocalizations of infants," *J. Voice*, vol. 16, no. 4, pp. 509–529, 2002.
[64] K. R. Scherer, "Vocal affect expression: A review and a model for future research," *Psychol. Bulletin*, vol. 99, no. 2, p. 143, 1986.
[65] K. R. Scherer, "A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology," in *Proc. Interspeech*, 2000, pp. 379–382.
[66] G. Schwarz, "Estimating the dimension of a model," *Annal. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[67] P. Shen and Z. Changjun *et al.*, "Automatic speech emotion recognition using support vector machine," in *EMEIT*. Piscataway, NJ, USA: IEEE Press, 2011, vol. 2, pp. 621–625.
[68] B. Sievers and L. Polansky *et al.*, "Music and movement share a dynamic structure that supports universal expressions of emotion," *Proc. Nat. Acad. Sci.*, vol. 110, no. 1, pp. 70–75, 2013.
[69] C. T. Snowdon, "Expression of emotion in non-human animals," in *Handbook of Affective Sciences*. London, U.K.: Oxford Univ. Press, 2003, pp. 457–480.
[70] N. H. Soken and A. D. Pick, "Intermodal perception of happy and angry expressive behaviors by seven-month-old infants," *Child Develop.*, vol. 63, no. 4, pp. 787–795, 1992.
[71] P. Somervuo and T. Kohonen, "Self-organizing maps and learning vector quantization for feature sequences," *Neural Process. Lett.*, vol. 10, no. 2, pp. 151–159, 1999.
[72] F. Spector and D. Maurer, "Synesthesia: A new approach to understanding the development of perception," *Develop. Psychol.*, vol. 45, no. 1, p. 175, 2009.
[73] H. Spencer, "The origin and function of music," *Frasers Mag.*, vol. 56, pp. 396–408, 1857.
[74] L. J. Trainor and C. M. Austin *et al.*, "Is infant-directed speech prosody a result of the vocal expression of emotion?," *Psychol. Sci.*, vol. 11, no. 3, pp. 188–195, 2000.
[75] M. Unuma and K. Anjyo *et al.*, "Fourier principles for emotion-based human figure animation," in *SIGGRAPH*, 1995, pp. 91–96, ACM.
[76] R. Van Bezooijen and S. A. Otto *et al.*, "Recognition of vocal expressions of emotion a three-nation study to identify universal characteristics," *J. Cross-Cultural Psychol.*, vol. 14, no. 4, pp. 387–406, 1983.
[77] A. S. Walker-Andrews, "Infants' perception of expressive behaviors: Differentiation of multimodal information," *Psychol. Bulletin*, vol. 121, no. 3, p. 437, 1997.
[78] A. Watanabe and M. Ogino *et al.*, "Mapping facial expression to internal states based on intuitive parenting," *J. Robot. Mechatron.*, vol. 19, no. 3, p. 315, 2007.
[79] M. Zecca and Y. Mizoguchi *et al.*, "Whole body emotion expressions for kobian humanoid robot–preliminary experiments with different emotional patterns," in *RO-MAN*. Piscataway, NJ, USA: IEEE Press, 2009, pp. 381–386.

Angelica Lim (S'10) received the B.Sc. degree in computing science from Simon Fraser University, Burnaby, BC, Canada, and the M.Sc. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2014.

She worked for Aldebaran Robotics Japan, and has interned at Google, Honda Research Institute Japan, and I3S-CNRS, France. Since 2006, she has been engaged in robotics and artificial intelligence, and is currently interested in signal processing, machine learning, and developmental robotics for intelligent systems, particularly in the field of emotions.

Dr. Lim was a Guest Editor for the International Journal of Synthetic Emotions, and has received various awards including CITEC Award for Excellence in Doctoral HRI Research (2014), NTF Award for Entertainment Robots and Systems at IROS 2010, and the Google Canada Anita Borg Scholarship (2008). She is also a contributor to IEEE Spectrum Robotics Blog.

Hiroshi G. Okuno (M'03–SM'06–F'12) received the B.A. and Ph.D degrees from the University of Tokyo, Tokyo, Japan, in 1972 and 1996, respectively.

He worked for NTT, JST, the Tokyo University of Science, and Kyoto University. He is currently a Professor in the Graduate Program of Embodiment Informatics, Waseda University, and a Professor Emeritus of Kyoto University. He was a visiting scholar at Stanford University from 1986 to 1988. He is currently engaged in computational auditory scene analysis, music information processing, and robot audition. He coedited "Computational Auditory Scene Analysis" (Lawrence Erlbaum Associates, 1998), "Advanced Lisp Technology" (London, U.K., Taylor and Francis, 2002), and "New Trends in Applied Artificial Intelligence (IEA/AIE)" (Springer, 2007).

Dr. Okuno has received various awards including the 1990 Best Paper Award of JSAI, the Best Paper Award of IEA/AIE-2001, 2005, 2010, and 2013, the IEEE/RSJ IROS-2001 and 2006 Best Paper Nomination Finalist Award, and the NTF Award for Entertainment Robots and Systems in 2010. He is a Fellow of the Japanese Society for Artificial Intelligence, and a Member of AAAI, ACM, ASA, RSJ, IPSJ, JSSST, and JCSST.