# MEI: Multimodal Emotional Intelligence

Angelica Lim

# Abstract

In this thesis, we design and implement a multimodal emotion system for robots. The overreaching goal is to advance the fundamentals of robot primary emotions by using clues from infant development. Our approach has three prime characteristics, which set it apart from current robot emotion systems.

Firstly, it is multimodal. Humans express and recognize emotions through a variety of dynamic channels, such as voice, movement and music. Our paradigm uses speed, intensity, irregularity and extent (SIRE) to colour a robot's voice, gesture and gait with emotion, using a simple 4-dimensional representation. Secondly, it models emotion statistically. Many emotion models are hand-defined based on a posteriori rules, yet humans are known to be statistical learning machines. Our MEI (multimodal emotional intelligence) module once trained, can recognize emotion in a context it has never encountered, and generate statistically probable emotion expressions. Finally, it is developed through a social process found in caregiver-infant interactions. Emotions are thought to be innate, but according to evidence in developmental psychology, much development happens between the ages of zero and one. In this thesis, we model this first year of life where emotional intelligence grows rapidly, possibly due to a universal phenomenon called motherese.

This thesis consists of seven chapters. In Chapter 1, we show the motivation for developing an emotion system for robots, and highlight the technical problems towards this goal.

In Chapter 2, we review the literature related to developing emotional intelligence at the primary emotion level. We give a basic overview of current definitions of emotion in psychology, neuroscience and computing, and describe related theoretical emotion models. We also give the three basic requirements for a robot emotion system: expres-

sion, recognition and representation, and describe the current state of the art for robots with these capabilities.

In Chapter 3, we describe a novel emotion representation called SIRE, and show its efficacy by developing the SIRE Emotion Transfer system. We model the dynamics of an emotion as four perceptual features – speed, intensity, irregularity and extent (each between 0 and 1). Experimental results show that SIRE transfer system can accurately *analyze* a vocal expression of happiness, sadness, anger and fear, convert it to SIRE space, and transfer the same emotion to an *expressive* gesture. Since SIRE can succinctly represent an emotion with only four numbers across multiple modalities, we use it as a basis for the rest of this thesis.

In Chapter 4, we extend our SIRE model to allow emotion recognition, even in a novel modality. We develop the Multimodal Emotional Intelligence (MEI) system, composed of a SIRE Gaussian Mixture Model for each emotion class. It provides a unique combination of advantages: 1) it integrates both recognition and expression into the same system, 2) it recognizes with a confidence measure for four classes, 3) it allows introspection to view an emotion class' most typical dynamics, 4) it produces non-repetitive emotional expressions, and 5) it accounts for individual differences in training data. Experimental results show that MEI can perform cross-modal emotion recognition with results almost comparable to inter-modal recognition.

In Chapter 5, we implement a robot system that develops multimodal emotional intelligence through real-time social interaction with humans. This interaction is inspired by an infant development phenomenon called "motherese", where a caregiver speaks in an emotional way to a baby. During the motherese interaction, the robot uses its MEI to simultaneously learn and entrain its gestural dynamics to the human's voice. It associates its internal state (physical "feelings" of flourishing or distress) to the emotional SIRE dynamics presented by the human. We show that a robot trained in an empathetic motherese loop will associate sad dynamics with its own physical distress, and happy dynamics with flourishing.

In Chapter 6, observations, general discussion and future work stemming from this thesis are described, and Chapter 7 concludes the thesis.

# Acknowledgments

This work was accomplished as part of the Okuno Laboratory, Graduate School of Informatics, Kyoto University and at Honda Research Institute – Japan in Wako, Saitama. It was sponsored in part by a generous grant from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

First of all, I would like to thank my supervisor, professor and mentor, IEEE Fellow Dr. Hiroshi G. Okuno. I still remember our conversations that inspired SIRE, elaborated on the back of a napkin in a fish restaurant next to Kyoto University. Thank you so much for your unwavering belief in me, kindness, enthusiasm, and contagious persistence that proves, "where there is a will, there is a way". I could not have asked for more in a supervisor for both my Master's and PhD. Thank you, Gitchang.

I would also like to thank Dr. Kazuhiro Nakadai for hosting me at HRI-JP. You were instrumental in allowing me to complete this thesis. I am very grateful for the time I have spent with you and your team, learning about multimodal fusion, auditory analysis, machine learning, and also what it takes to make successful and enjoyable research. I feel very lucky to have been given the opportunity to work at HRI-JP with you.

I also would like to thank my jury members, especially Dr. Toyoaki Nishida, who is an expert in the fields of AI and social robotics, and provided excellent guidance throughout my years at Kyoto University.

Next, I would like to express my gratitude to my research mentor, Dr. Takeshi Mizumoto. For nearly 5 years, he has provided fruitful discussions and many "aha!" moments. From my research beginnings with music robots to HARK, he has been my inspiration, collaborator, and source of many laughs. Thank you.

I am also grateful to the colleagues who have provided a wonderful research environment at HRI-JP and Kyoto University. Dr. Keisuke Nakamura, Dr. Randy Gomez

and the interns from Canada, India and France who made research fun! I am also grateful for the support of Takuma Otsuka, Dr. Katsutoshi Itoyama, Dr. Eui-Hyun Kim, and my other colleagues in Okuno-ken.

I am also indebted to my colleagues at Aldebaran Robotics, in particular my former supervisor Jérome Monceaux. Jérome's enthusiasm for robotics and animated phone conversations inspired me from the other side of the world. His support, even through the toughest of times, was fundamental in allowing me to finish this PhD, and I will never forget this. Thank you to my amazing colleague Sébastien Cagnon, together with whom we opened the doors at Aldebaran Robotics Japan. I am also grateful to A-Lab director Jean-Christophe Baillie for the inspiring conversations that led to new ideas for my thesis. I thank all of the wonderful colleagues in Paris and Shanghai, especially Benoît Libeau, Qingyi Lelay, Viviane Dejeammes, the Studio team, Interaction team, Platform team, and Internal communication. Last but not least, thank you to Bruno Maisonnier, who is one of the most inspiring leaders I have ever met, and who encouraged me to pursue my dream of completing my PhD.

I thank our lab secretary, Ms. Hiromi Okazaki, for the valuable support that helped make my research here in Japan possible. I am so grateful to my coach, Soness Stevens, who not only turned my life around through her positive energy and life strategies, but also taught me the meaning of true gratitude and emotion. I also thank Ben Humphreys, whose partnership, support, humour has made the thesis-writing process and living in Japan a joy.

Finally, I thank my parents Delia and Philip Lim, my brothers Christian and Jacob, my family including Abbie and the Cu family, who have supported me from abroad for years. It has been a difficult journey, but knowing that I have your love and support throughout it all has made it possible.

# Contents

# List of Figures

# List of Tables

# 1

# Introduction

*"When dealing with people, remember you are not dealing with creatures of logic, but creatures of emotion."*

– Dale Carnegie

## 1.1 Motivation

The most emotional moments of our lives are the most memorable. Our best friends help us lead lives that are happy and bright, and are endearingly empathetic when we're down. Emotions colour our world, our interactions, our words, from humming in the morning over breakfast, to smiles before sleeping at night. Positive emotions help us to be more creative, be more optimistic, and even work harder. Negative emotions help us focus, narrow our field of view to attack a problem, or change course when one direction isn't working out.

Robots show promise in helping us in these emotion-governed lives. Just as the Internet and mobile technology have made us more connected, new robotic technologies are opening a door towards supporting an aging society. In Japan, almost 25% of the population is over 65 years old[1], and they seek a life of retirement with independence in the community, physical activities, and an active social life [12]. To meet the rising demand for healthcare workers and more, the government has estimated that the service robot market will reach over 4900 billion yen by 2035[2], exceeding the demand for robot

[1] http://www.stat.go.jp/data/jinsui/pdf/201311.pdf
[2] http://www.meti.go.jp/english/press/2013/pdf/0718_01.pdf

manufacturing by almost twofold. It is hoped, for example, that robots can help the bedridden become mobile, and the dependent become independent.

Yet robots have to overcome the challenge of navigating our world, because it is not always black and white. Imagine a healthcare robot overseeing an elderly patient named Linda at the hospital – the robot is set to close the room by 9pm. Soaked by the rain, the patient's daughter, Mary, knocks on the hospital room door. She has driven 50 kilometers from the airport, but a thunderstorm has delayed her arrival. Mary yearns to hold her mother's hand – it has been 3 years since their last meeting. Linda is delighted to see her daughter through the hospital room window, but it is now 9:01pm. Crestfallen, the mom and daughter eyes meet, as the healthcare robot locks the door with a loud thud. The rules are rules.

"The heart is a strange beast and not ruled by logic." (Maria V. Snyder)

Robots do not share our capacity for emotion. In science fiction, Star Trek's android lieutenant Data was described as human-like in many ways, except that he lacked emotions: "human behavior flows from three main sources: desire, emotion, and knowledge," Plato once said, and Data had goals and knowledge, but no emotion. In many futuristic movies, this emotional shortfall drives robots to take over the world. Like history's worst dictators, these robots' calculating brilliance, logic, and lack of empathy bind together in cruel combination.

It is easy to see why, in a 2012 survey, 60% of EU citizens stated that robots should be banned in the care of children, the elderly, or the disabled[3]. Large majorities would also agree to ban robots from 'human' areas such as education (34%), healthcare (27%) and leisure (20%) [their quotes]. Of course, in certain environments like factories, bomb-detection or remote operating tables, the precision and predictability of robots is a necessity. Yet a new breed of "service robots" are advancing to our doorstep quickly, with the potential to change the lives of children and the elderly, able-bodied and disabled, students and more. For robots to be accepted in our daily lives as helpers, we must release robots from their pure, programmed logic and make them more emotional, more empathetic, to interact with humans on their own terms.

[3] http://ec.europa.eu/public_opinion/archives/ebs/ebs_382_en.pdf

How do we start to build such a robot? One guiding principle could be to look to human development for inspiration. Just as each human has linguistic abilities (whether through voice or sign-language), each human is equipped with the capacities of emotion expression and understanding. And whether they were raised in Japan, the USA, China or France, each person is unique based on their upbringing and environment. They may express happiness loudly or quietly, they may fear snakes or love snakes. They may be more or less sympathetic. They may openly declare displeasure or only show it through one eyebrow. Their abilities may fall on a spectrum of what we consider underdeveloped emotional intelligence, or autism. Clearly, there is no one-size-fits-all definition, and likewise, a robot's emotions should be adaptive, too. Sometimes this zealous focus on pliable, human-like models may appear to be a detriment to the short-term accuracy of the systems we engineer. But with the goal of autonomous, ever-learning robots, our hope is that in the long-term, we will be building the foundation of a powerful artificial emotional intelligence.

## 1.2 Technical Problems and Solutions

This section summarizes the main technical problems covered in this thesis.

### 1.2.1 Problem 1 – Expressing emotion without a face

How can a robot express emotion? Consider that many robots do not have movable facial features, including robots like NAO[4], Robovie[5], or animal-type robots. One typical approach is to hand-define poses or create emotional animations, but once the animation or pose is finished, the robot appears to "lose" its emotion.

We thus solve this problem through multimodal expression of emotion through vocal, gestural, or gait dynamics. We construct a framework based on speed, intensity, irregularity, and extent (SIRE), and map high-level perceptual SIRE features to low-level features such as speech rate or velocity of movement. We show, for example, that slow, non-intense, regular and small dynamics can express sadness in robot voice, movement and music, and that fast, intense, irregular, small dynamics can express fear.

[4] www.aldebaran-robotics.com
[5] http://www.vstone.co.jp/

### 1.2.2 Problem 2 – Cross-modal emotion recognition

Consider that we can watch a cartoon lamp jump for joy, and share its happiness[6]. Humans have the ability to recognize emotion in new contexts, yet this remains a major challenge for robots. This is because current paradigms would typically train a separate recognition module for each new context (e.g., one for singing voice, one for animal emotion recognition, for Japanese, English, indoor environment, outdoor environment, etc.), and test in the same modality/context. This specialization of intelligence is not suitable for robots exploring new environments.

Our solution to cross-modal emotion recognition has three characteristics. First, we assume that each individual expresses themselves differently. We create a normalization scheme for their dynamic emotion features (from Problem 1) to SIRE space based on their own statistical distribution, allowing, for example, a person that speaks quickly or a person with a hoarse voice, to contribute to or use data from the same trained model. Secondly, we map features from different modalities (e.g., gait, voice) to the same space using common perceptual features. Thirdly, we train Gaussian Mixture Models (GMM) with these SIRE features, and optimize their granularity with automatic selection of components through a minimization of their Bayesian Information Criterion score. This allows us to construct a single Multimodal Emotional Intelligence (MEI) module that can recognize emotions in a modality it has never seen before: human gait.

### 1.2.3 Problem 3 – Integration of recognition and expression systems

A major challenge in robotics is integration, and emotion is no exception. Modules for emotion recognition and expression are usually separate. For example, Kismet, one of the few integrated emotional robot systems, has a voice emotion recognition module that is separate from the emotional voice expression module [13]. This is undesirable because even hundreds of hours of this emotional voice input, though recognized, will never improve the way the robot's own emotions are expressed.

To address the integration challenge, we design a machine learning model that performs both emotion 1) recognition, 2) representation, and 3) expression. Our SIRE

[6] In Pixar's two-minute short film Luxo Jr., a small, "young" lamp plays exuberantly with a ball.

Gaussian Mixture Model design allows a) recognition of an emotional expression with a confidence score b) model transparency through inspection of GMM means, and c) non-repetitive expression via distribution sampling.

### 1.2.4 Problem 4 – Grounding emotional expression in low-level feelings

It is not clear how a robot could truly "have" emotion. In practice, hand-defined labels or symbols are used to represent emotions in current systems [3]. Simulated models of emotion propose grounding schemes based on concepts such as intrinsic drives [14] [15], but do not explain how, for example, feeling could be felt when listening to sad music.

We attack this problem by taking a developmental approach. A robot trains its MEI through real-time interaction with a caregiver, associating its internal state (physical "feelings" of *flourishing* or *distress* based on battery levels) with emotional SIRE dynamics. It expresses emotions through a combination of a) empathetic mirroring [16] and b) its own internal physical "feeling" state. We show that a robot trained in an empathetic caregiver loop will associate sad voices with its own distress at 84%, and happy voices with flourishing at 90%.

## 1.3 Organization

In this thesis, we report the design, implementation, and evaluation of a multimodal emotional intelligence system towards meeting these challenges. An overview of chapters 3-5 is shown in Figure 1.1.

In chapter 2, we start by a review of robot emotion literature and theoretical emotion models in psychology, neuroscience and computing. The distinction between primary and secondary emotions is discussed, and we define our contribution within the context of a robot's primary emotion system.

In chapter 3, we explore the multimodal nature of our emotional expression: how do simple sounds, music, dance, and even cartoon animations express emotions? We develop the SIRE Emotion Transfer system to test various values of speed, intensity,

irregularity and extent, to see how well four numbers can capture an emotion through dynamics.

In chapter 4, we describe a system for emotional intelligence called MEI, composed of SIRE GMMs. It performs both recognition and expression, and has an interpretable representation. We show how a trained MEI can recognize an emotion in a situation it has not encountered.

Finally, in chapter 5, we describe how a MEI can associate low-level feelings and emotional expressions through human, caregiver-like interaction.

Chapter 6 we describe general observations, provide avenues for future research.

Chapter 7 concludes this thesis.

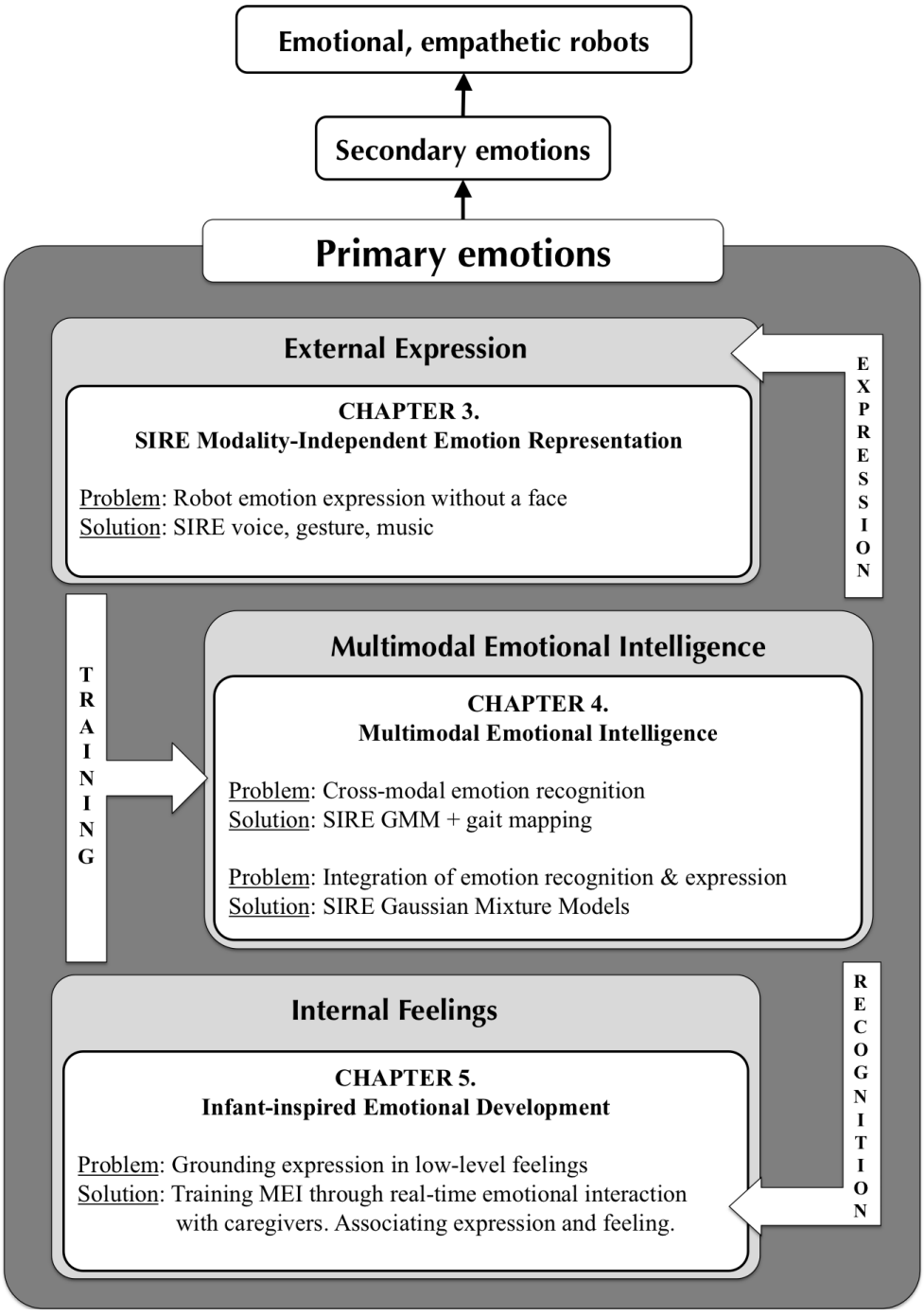



Figure 1.1: Thesis organization.

# 2

# Literature Review

*"Robots can serve as a test-bed for evaluating models of human and animal social learning."*

– Cynthia Breazeal

## 2.1 What are emotions?

Emotion has been defined in many ways. Researchers in 'appraisal theory' (for example, Scherer [17]) study emotion as "relatively brief and intense reactions to goal-relevant changes in the environment" [18]. For example, if our goal is to drive down a road, and a car cuts in front of us, we may experience anger after a quick appraisal of the event. Many systems for artificial emotion reason about an emotional state using this appraisal approach. On the other hand, consider that emotions can be evoked without any specific goal-related event. For example, a rousing musical piece can stir up feelings of joy, and slow, vibrato-heavy music can move us to tears.

### 2.1.1 Primary and secondary emotions

In the field of neuroscience, Damasio [2] distinguishes between the two types of emotions above: primary and secondary emotions. Primary emotions, he says, are initial, supposedly innate emotional responses processed in the lower-level limbic system of the brain. For instance, fear can arise before the signals even reach the cortex to perform reasoning [2]. In affective computing, Picard calls this the *physical* or *bodily*

*component* [4]. Secondary emotions comprise emotions such as grief, in response to the understanding of a loss of a loved one, for example. This involves slowing-acting, higher-level reasoning, and are what Picard refers to as the *cognitive component* of emotion [4].

A primary emotion may be very different from its ensuing secondary emotion. Imagine, for example, seeing your sister excited that she found a job in a different country. Your primary emotion may be happiness as you share her joy, but a few seconds later, your secondary emotion could be sadness when you realize you will no longer see her. Parrot [19] suggests a tree-like hierarchy of primary, secondary and tertiary emotions: primary emotions comprise love, joy, surprise, anger, sadness and fear. Secondary emotions stem from these primary emotions after cognitive appraisal. For example, secondary emotions for anger include rage, envy, exasperation; joy can be further broken down into contentment, enthrallment, relief, and so on.

Primary emotions are sometimes also called basic emotions. For example, Ekman [20] in his work on universal facial expressions lists 6 emotions: fear, anger, sadness, happiness, disgust and surprise. Tomkins [21] includes interest and shame; Johnson-Laird and Oatley list 5 emotions: fear, anger, sadness, happiness and disgust [22]. It should be emphasized, however, that there is no clear consensus, and there is even a growing movement of researchers who refute the idea that simple, basic emotions even exist [23]. Nevertheless, from these lists and research across disciplines, the most common four emotions studied (combining near synonyms like joy and happiness) are joy, sadness, anger and fear [4].

We use the definition of *emotion* offered by Juslin in [24]:

> **Emotion**: a quite brief but intense affective reaction that usually involves a number of sub-components – subjective feeling, expression, action tendency, physiological arousal, and regulation – that are more or less 'synchronized'.

### 2.1.2 Other definitions

We offer a few other definitions related to emotion. *Affect* is the general word that covers phenomena involving all valenced (positive/negative) states, including emotion, mood,

and preference [24]. *Moods* are lower in intensity, and last longer than emotions (i.e. several hours to days) and are not necessarily directed towards any clear 'object' [24]. The word *feeling* can also be distinguished here; Damasio [2] defines feeling as the private, mental experience of emotion. Unlike emotions, they are not observable.

*Empathy* is defined as "the ability to understand and share the feelings of another" [25]. Related to empathy is *emotional contagion*, defined as "the tendency to automatically mimic and synchronize facial expressions, vocalizations, postures, and movements with those of another person's and, consequently, to converge emotionally," [26] or more generally "a process in which a person or group influences the emotions of another person or group through the conscious or unconscious induction of emotion states." The latter definition is more appropriate, for example, in the case where music induces emotions in listeners [24]. Finally, we make the distinction between *emotion* (singular), which is the phenomenon described earlier, and *emotions* (plural), which we will use here interchangeably for emotion classes (e.g., happiness, sadness).

## 2.2 Human emotion: what, when, how?

To develop a robot emotion system, we should first understand the basic processes in human emotion. The literature on emotion is vast, so we will review some theories from three perspectives: *what* comprises an emotion, *when* emotion happens, and *how* it develops.

- **What**: This refers to the components of emotions, such as subjective feeling, physiological arousal, and expression.

- **When**: This refers to timing and phases of an emotion. We will examine the timeline for the beginning of an emotional reaction.

- **How**: This refers to the mechanisms behind emotion. Compared to other bodily processes, this is relatively poorly understood.

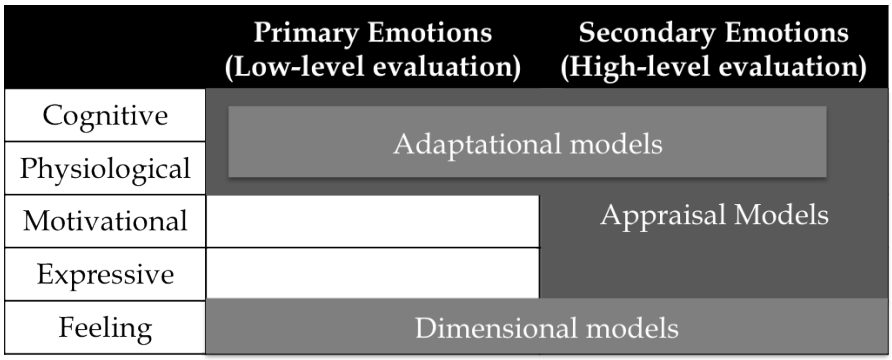


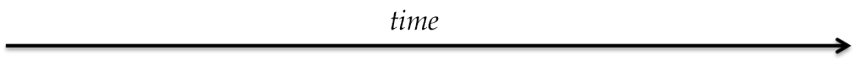

Figure 2.1: Current emotional models and their scope, shown in grey, reproduced in part from [1]. The horizontal axis outlines the temporal phases in the emotion process, and the vertical axis shows the emotion components.

### 2.2.1 What: Emotion components

What happens during an emotional reaction? Our earlier definition of emotion supposed that there is a synchronized reaction of subcomponents [1] [24], to name a few:

- *Subjective feeling*, the 'felt' part of an emotion
- *Expression*, including postures, gestures, facial and vocal expressions
- *Action tendency*, a probable action in response to the emotional event, e.g., flee or fight in response to fear
- *Physiological arousal*, including cardiovascular, electrodermal, gastrointestinal, hormonal activity
- *Regulation*, (cognitive) processing and integration of emotion components, e.g., to correct inappropriate emotional responses

Psychologists, neuroscientists, and cognitive scientists have proposed models for most of these components. A recent table from [1], gives an overview emotion theories touching on some of these aspects, listed vertically in Figure 2.1. *Appraisal models* touch on high-level evaluation of a situation, including low-level cognitive and physiological aspects. The field of appraisal theory was pioneered by Magda Arnold in the 1960's, who proposed that emotion results from an appraisal of the situation [27].




| Primary Emotions | | Secondary Emotions | |
|---|---|---|---|
| Reactive | Routine | Reflective | *Ortony* |
| Physical | | Cognitive | *Picard* |
| Low-level | | High-level | *Scherer* |

*Damasio*

Figure 2.2: Different labels for the stages of emotion, according to neuroscientist Antonio Damasio [2], cognitive scientist Andrew Ortony [3], computer scientist Rosalind Picard [4], and cognitive psychologist Klaus Scherer [1].

Lazarus further developed the field [27], with the more recent models in psychology by Scherer [10]. Appraisal modeling has spurred designs for virtual agents, such as the Ortony-Clore-Collins (OCC) model [28], for example. *Adaptation models*, proposed for example by neuroscientist LeDoux [3], suggest that organisms are biologically primed by evolution to detect and respond to stimuli important for their survival. These theories focus particularly on emotional responses related to fear-inducing stimuli, such as snakes or electric shocks [1]. The *dimensional theories* focus on the representation of emotion, for instance, as a point, for example on a 3-dimensional pleasure-arousal-dominance space [29] or 2-dimensional valence-arousal "circumplex" space [30]. Although being adapted to capturing the dimensions of feeling, it is not always clear how to map them to emotional expressions [31].

Interestingly, we notice a large gap in primary emotion models. Specifically, low-level, multimodal expression (facial configuration, gestures, vocal tones) is unaddressed by current emotion models, likely because it is assumed that emotional expressions are innately prepared. Since secondary emotions (modeled most broadly by appraisal models) are thought to depend in part on their preceding primary emotions [10], we suggest that it is worth building a solid theoretical foundation of primary emotion to support these models. We next examine this temporal process, in which primary emotions occur, followed by secondary emotions.

### 2.2.2 When: Emotion phases

What are the phases in an emotional reaction? As shown in Figure 2.2, researchers suggest that primary or low-level appraisal probably occurs first, followed by secondary/high-level appraisal. Primary emotion appraisal is supposed to be fast, though it is not clear to what extent primary emotions wired in at birth. Humans, for example, are not innately wired to be afraid of bears or eagles, notes Damasio [2]. But he posits that there may be certain stimuli that can automatically evoke an emotion, such as size (e.g., large animals), type of motion (e.g., reptiles), sounds (e.g., growling), configurations of body state (e.g., pain of a heart attack) [2].

Ortony gives more insight into primary emotions. He proposes that so-called hard-wired emotional reactions can be further categorized into *reactive* and *routine* [3], as shown in Figure 2.2. At the *reactive level*, he calls the simplest form of affect *proto-affect*: the reactive level assigns values along two output dimensions of positive and negative. For example, warmth, rhythmic beats, and sweet tastes automatically produce positive valence, and extreme heat or cold, loud or dissonant sounds, and bitter tastes induce negative valence. He considers this reactive level as automatic and biologically prepared.

At the next stage, the *routine level*, he defines *primitive emotions*. Whereas reactive-level processes are fixed by biology, much of the content at the routine level is learned thoughout life. These are the automatic, implicit result of the "accumulation of experiences that forms a general model of likely, or 'normal,' outcomes or events". The output of emotion processing at the routine level motion are then fed into the *reflective* level for further cognitive processing. He discerns four categories of primitive emotions:

- **Happiness**, a (positive) feeling about a good thing (present)
- **Distress**, a (negative) feeling about a bad thing (present)
- **Excitement**, a (positive) feeling about a potential good thing (possible future)
- **Fear**, a (negative) feeling about a potential bad thing (possible future)

We can notice here a special distinction: anger is actually different from happiness,

sadness, and fear because, unlike Ortony's four primitive emotions, it requires cognitive appraisal at the *reflective* level.

We therefore consider at least 3 temporal stages in an emotional reaction (Figure 2.2, Ortony):

$$Reactive \rightarrow Routine \rightarrow Reflective.$$

### 2.2.3 How: Emotion development

How are emotions (simultaneous expressions, physiological responses, etc.) developed? The "how" question is a particularly important issue for roboticists, because its answer determines the implementation for an artificial emotion system. Yet, the answer to this question is not at all clear. Based on a review by Scherer ( [1], p. 11-15), most theorists suggest an evolutionary view: "basic emotion" theorists argue that fundamental emotions are phylogentically stable neuromotor programs, created in a Darwinian evolutionary sense. Circuit theorists assume that fundamental emotions are determined by genetically coded neural circuits. Yet, how does one explain the differences in emotional expression between people of different personalities or cultures? Ekman and Friesen suggest that humans learn 'display rules' [32]: cultural norms to express different emotions in social situations. How exactly these rules are learned is not elaborated in detail.

To give an alternative perspective on this answer, we give a brief literature review of research in emotion development in infant psychology. In particular, we consider the primitive, emotional intelligence of infants before the age of 1 year old, during which dramatic development occurs:

> "Although the development of emotion perception extends beyond infancy–perhaps throughout the lifespan–[...] *dramatic changes in emotion perception competencies [...] are observed over this period of development [during the first year of life].* Furthermore, it may be that infants reared in situations with impoverished affective expression information, such as those, for example, from caregivers with clinical depression, or in contexts where actions and expressions are discrepant, may be particularly influenced in their comprehension of expressions." [33]

This early boom in emotional expressions typically evolves in a set sequence: joy and sadness expressions appear at 3 months, anger between 2 to 6 months, and fear at 6-8 months [34].

For instance, vocal emotion development occurs within the $1^{st}$ year of life (see [35] for a recent review). In terms of expression, emotional vocalizations are distinguishable in the 5th or 7th month [36]. Similarly, for recognition, 5-month old infants pay attention longer to happy, sad, or angry voices when they co-occur with facial photos, but not with black and white checkerboards. 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 [37]. 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 [35] [38].

Infants also develop emotional capabilities in other modalities within the first year. When listening to an unfamiliar language in infant-directed speech, 5-month-olds smile more often at approving voices and show negative affect when listening to prohibitions [39]. At 7-months old, babies look longer at affectively concordant point-light displays [40] 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 [41]. 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 basic multimodal emotional intelligence is developed within the first year of life [33] — even before the onset of speech [42] — providing a promising perspective for our "how" question. In developmental psychology, [33] and [37] suggested a kind of associative learning between the affective voice and face. In developmental robotics, the "intuitive parenting" paradigm has been proposed for grounding emotional face [43] [44]. Recent research has suggested that a universal mechanism called "motherese" (a.k.a., infant-directed speech) is at the crossroads for developing emotions, cognition and language [45]. In Chapter 5, we will examine in depth these promising ideas of associative learning, intuitive parenting and motherese.

## 2.3 Designing emotional robots

Now that we have a better understanding of the human phenomenon of emotion, we turn to the state of the art in robotics. Robotics research has mostly taken a functional perspective toward emotional robots, and typically addresses one or more of the following aspects when implementing emotion: expression of emotions, recognition of emotions, and representation of emotions. It should be mentioned that research in virtual, sociable agents is a large source of inspiration for robot implementations, for instance work by Picard [4], Nishida [46] and Pelachaud [47]. Due to the wealth of existing studies in each of these domains, only research implemented on robot platforms will be covered in the following review.

### 2.3.1 Expression of emotions

The most natural place to start for robot emotions is the *expression* of emotions. Many robots perform this function using a moveable face, whether they are hyper-realistic [48] or cartoon-like [13]. We describe in depth three more modalities which will most useful for a robot without a moveable face (e.g., NAO): bodily expression, vocal expression, and musical expression.

#### Bodily expression

Humanoids are natural platforms for expressing emotion through bodily configurations. Using pre-defined robot poses such as raised arms for surprise [49], or an aggressive stance for anger are common approaches. These allow for scaling so that the gesture can be more or less activated [50] [51]. We can note that the use of poses is restricted in two ways: 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 [52]) at the same time. Using this pose-based design, for example, “angrily pointing” would not be possible.

Movement of the body is another promising method. For example, Kismet drew its head backwards to indicate fear [13]. Nakata et al. used a technique from acting called Laban Movement analysis [53], which conveys affect through features relating to

weight, space, and time. One study on the Roomba robot showed that even without a humanoid form, acceleration and curvature can contribute to impressions of valence [54].

### Vocal expression

Emotional voice synthesis typically involves modulating neutral speech to make it more expressive. Specifically, one would vary a text-to-speech (TTS) system's vocal characteristics such as pitch, amplitude and speech rate. Kismet modulates fifteen "vocal affect parameters," which can be grouped into pitch, timing, voice quality and articulation aspects [13]. iCat uses a similar scheme with the EmoFilt [55] synthesis system. Part of the work is mapping these perceptual features onto the parameters of the TTS system (such as DECTalk for Kismet.)

After mapping, these perceptual features are modulated according to results of studies on vocal emotions. For instance, fear typically manifests with fast speech rate, high pitch, wide pitch range, and irregular voicing. Sorrow is slower, lower in pitch, with a narrow pitch range and resonant voicing. Comprehensive tables can be found in [56] [57], for example.

### Musical expression

In the field of music robotics, a few studies modulate parameters known to be important in emotional expression through music. Ogata et al. [58] developed a violin playing machine which could change timbre according to adjectives such as "moist" or "dry". Solis et al. [59] developed an expressive flute-playing robot, which learned parameters such as note length and vibrato based on a human flutist's performance. Nakano and Goto extracted the vocal [60] and facial expression from a famous singer, and later reproduced the performance on the hyper-realistic android HRP-4C. Lim et al. [61] used a programming by playing approach to transfer expression from a human flute player's performance to a robot thereminist. Robots such as Shimon, the marimba playing robot [62] and the trumpet playing robot from Toyota move their bodies in order to add expression [63]. On the other hand, no experiments have yet been conducted to assess the emotional communication between music robots and humans.

### 2.3.2 Analysis of emotions for robots

Secondly, robots should not only express, but also *analyze* and *recognize* emotions in the world around them. While emotional expression has been often attempted in robotics, analysis of emotions – that is, in a real-time, embodied setting – has been less studied.

A few robot systems perform emotional analysis. Kismet [13] used an indirect approach to analyze the emotional content of interlocutor voices. First, it classified voices into one of 5 categories (approval, prohibition, comfort, attention and neutral) using a Gaussian Mixture Model (GMM) trained on 12 features. These categories were then mapped by hand onto affective dimensions of arousal, valence, and stance (called an [A, V, S] trio). For instance, approval was mapped to medium high arousal, high positive valence, and an approaching stance. These [A, V, S] values in turn were fed into Kismet's emotional system to produce an appropriate social response. Neurobaby [64] was a simulated infant that responded to changes in voice, using a neural network to detect one of four emotional states, and a robotic hand interface that detected intensity. Although robots reacting to emotional movements is still an open topic, Mancini et al. [65], created a real-time system to detect emotion from videos of dancers by tracking just several features such as extent. Common difficulties include the real-time factor, adaptation to speaker, context, and frame size.

The current approaches for emotional analysis still have room for improvement. State of the art emotion recognition systems in voice and music (e.g., [57] [66] [67]) typically use the a feature-classification approach: first, high-dimensional vectors of low-level features such as f0, MFCCs, spectral flux, decibel power are extracted from the audio signal; then, a classifier is used for training and classification. For example, Tahon et al. recently studied a dataset of children and elderly interacting with a robot [68]. Their results show that a GMM trained on the features of one corpus could not be used to reliably detect emotions in the other. This generalization problem has been touched on in music information retrieval as well. The review in [69] suggests that this high dimensional approach has reached a glass ceiling, and new approaches involving higher level processing are needed.

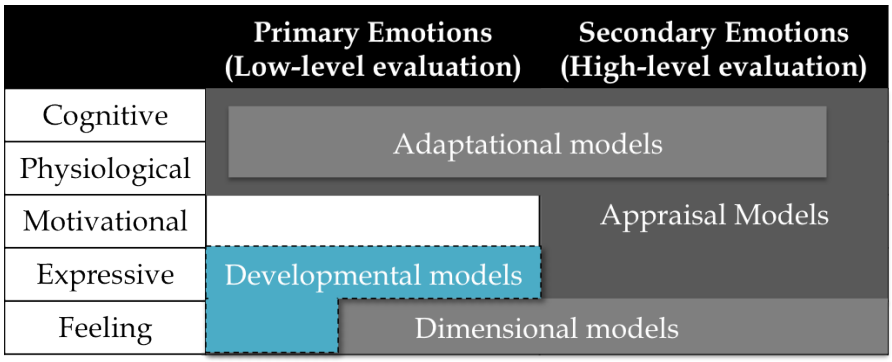



Figure 2.3: Current emotional models and their scope, shown in grey, reproduced in part from [1]. We define a new area called developmental models (in blue) which addresses low-level expression. We touch briefly on the subject of feeling in Chapter 5.

### 2.3.3 Emotion representation for robots

Finally, a robot should keep track of its internal affective state using an *emotion representation*.

Categorical representations (e.g., [20]) of emotion presume that the emotional state can be classified into any one of several discrete classes. These models use labels such as happiness and sadness, and is the most straightforward way to represent emotions. On the other hand, this representation fails to take into account the fact that emotions may be subtle and continuous, or as Fellous puts it, "wax and wane over time" [70]. Velasquez's Simón the Toddler [15] is an example of a simulated artificial emotion system with levels of activation for happiness, sadness, fear, anger, disgust and surprise.

A more common approach is called the dimensional representation. The most famous representation is Russell's circumplex model of affect [30], which represents emotion along two axes: arousal and valence. The arousal dimension is continuous from relaxed to aroused, and valence axis from unpleasant to pleasant. Similar models exist, such as the 3-dimensional emotional space (PAD) [29] which adds a dominance dimension, a 4-dimensional space with expectation [71]. Breazeal's Kismet used a variant of PAD to continuously model the robot's internal emotional state.

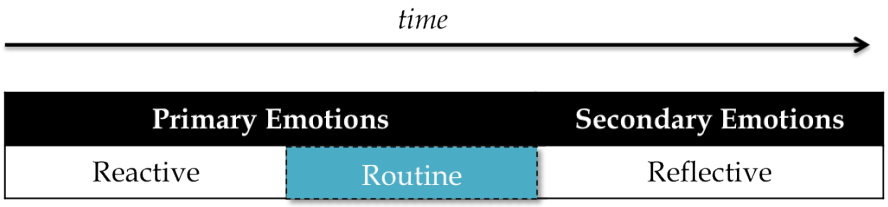



Figure 2.4: The stages of emotion. The focus of this thesis, primary emotions at the routine level, is highlighted in blue.

## 2.4 Scope

Amidst this sea of theoretical and engineering emotion research, we limit the scope of this thesis in three ways:

- **What**: Non-facial, multimodal expression and feeling aspects of emotion (Figure 2.3)
- **When**: Low-level, primary emotions at the routine stage (Figure 2.4)
- **How**: Emotional development inspired by infant development (Figure 2.3)

We illustrate the scope in Figure 2.4 and 2.3. We will focus on developing an emotion system, specifically at the *routine* level – according to Ortony, this is developed thoughout life through an accumulation of experiences. We focus on their development through *infant-caregiver interaction*, and define a new area called *developmental models*, which touch on the expressive component in emotion modeling. Indeed, this research area is only beginning (e.g., work by Asada et al. [43]), and our aim is to build a stronger foundation for the complex emotional, social, and cognitive processes that stem from a primary emotion system.

In terms of robot design, we will touch on all of expression, recognition, and representation of emotion. For expression, we will specifically address emotional expressions that are dynamic through time, such those expressed through movement, voice and music. In terms of recognition and representation, the majority of our work will, due to availability of training databases across modalities, address the four emotions most studied (happiness, sadness, anger and fear). In Chapter 5, we will also consider other classes such as distress and flourishing.

# 3

# The SIRE Paradigm: Modality-Independent Emotion Representation

*Music can describe emotions far more accurately than words ever can.*

– John Lydon

## 3.1 Introduction

In this chapter, we explore the many ways that emotions can be conveyed outside of facial expression. For example, imagine 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 convey such emotion? This kind of 'emotional intelligence' – to sense and express emotions through various means – appears to be built into any normal-functioning human and even some animals. We propose that robots, too, can develop this expressive ability, no matter the communication channel. But first we must investigate how humans express emotion, whether the channel is movement, voice, or even music.

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 sento-

graph, a device to measure touch [72]. 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 [73] 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 [74] [75]. 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 9 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 [77]. Juslin and Laukka [18] 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 [78]. In primates, dominant males often emit threatening vocalizations with characteristics similar to those of human anger [78].

Musicians also have the power to evoke emotions in their listeners. Movie scores are written with the purpose of matching a film script; a slow, vibrato-heavy violin ballad matches a sorrowful scene, whereas the same violin playing quick, anxious repeated notes with unexpected attacks would accompany a horror film. This musical capacity appears somewhat universal; for instance, Japanese listeners are able to recognize joy, anger and sadness in both Hindustani and Western music [79] just as well as they do in Japanese music [80]. Babies by the age of nine months can discriminate between happy and sad music [41], and by the age of six they can identify sadness, fear and anger in

music [81].

Researchers suggest that whether it be a step, tone of voice, or even a musical phrase, the expression of emotions have the same underlying dynamic 'code' [72] [18] [82]. For example, both loud, intense musical phrases and large, forceful movements conveyed 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. A recent study by Sievers et al. [83] found strong evidence that music and movement share the same dynamic structure. Sievers asked participants to control an animated bouncing ball, using five parameters: rate, jitter, smoothness, step size, and direction. Using these five parameters, they created *movements* corresponding to "happy", "angry", "sad", "scared", and "peaceful". The participants were then asked to use the same five parameters to create synthesized *music* for each of the emotions. Upon comparing the parameter settings for a given emotion, they found that the dynamics for both music and movement were similar. Furthermore, upon comparing results in the USA and a rural village in Cambodia, they found that dynamic features of emotion expression appear cross-culturally universal. (They do not claim that their 5-parameter emotional space describes the parameter space optimally.)

In a similar way, we propose emotional expression through a four-parameter representation: speed, intensity, irregularity and extent. We target voice, movement, and music, to endow a robot to express emotion across a broad set of modalities. For short, we call our parameter set **SIRE**. *Note: For the remainder of this chapter only, the parameter Regularity – not IrRegularity – will be used, due to our definition at the time of experiments. Each can be derived from the other through a simple inversion.*

## 3.2 Modality-independent emotional parameters

The rationale of the SIRE representation stems from common factors across modalities. We refer the reader to key reviews for both emotion *recognition* and *generation* in music, speech, and gesture. For instance, Livingstone et al. provide an excellent up-to-date review of the last half century of research in musical emotion [84]. In speech, Cowie et al. present a review of factors for recognizing or synthesizing expressive and emotional

speech [57]. Emotional gesture has been less studied, though Pelachaud's work in animated characters [47] may be the most state-of-the-art. Amidst a sea of features and modality-specific variables, several factors repeatedly resurface, albeit under different names. We found that the most salient factors for emotion recognition and generation could be clustered perceptually into *speed, intensity, regularity*, and *extent*. The results of our review are summarized in Table 3.1.

*Speed* is the most robust feature and has been called speech rate, velocity of gesture, or tempo. The dichotomy between fast and slow is the simplest way to distinguish between happy and sad voices, music and gestures.

*Intensity* is another useful feature, which we define as the perceptual distinction between gradual and abrupt. For instance, anger is often characterized with abrupt movements or attacked words and musical notes [85] [18]; sad voices, music and gestures are correlated with low intensity, gradual changes.

*Regularity* is the perception of smooth versus rough. For example, fear can be indicated in a voice with a breathy voice quality [57], quivering (as opposed to smooth) gestures, and music with irregular, sporadically played notes.

*Extent* may be the second most important feature after speed: for gesture, large, expansive movements could be characteristic of happy or (hot) anger. Smaller movements can indicate depression or making oneself small due to fear. In music and voice, the same effects are observed for loud versus soft.

## 3.3 The SIRE model

Conceptually, SIRE is a way to represent emotions in the four dimensional space of speed, intensity, regularity, and extent. For example, "sadness" may be represented as a point with low speed, low intensity, high regularity and low extent. Furthermore, this representation can be applied for both synthesis and analysis of emotional expressions (as will be shown in the Experiments section) for multiple modalities. Figure 3.1 illustrates how we can both extract a SIRE representation, and express it through different modalities.

In short, the SIRE framework is:

Table 3.1: SIRE parameters and associated emotional features for modalities of voice, gesture and music. Features in *italics* are used in our experiments ( [5], p. 4).

| | | **Modality mappings to relevant emotional features** | | |
|---|---|---|---|---|
| **Parameter** | **Description** | **Voice** | **Gesture** | **Music** |
| Speed | slow vs. fast | *speech rate* [57], pauses [18] | *velocity* [65], animation [86], quantity of motion [85] | *tempo* [84, 87] |
| Intensity | gradual vs. abrupt | *voice onset rapidity* [18], articulation [57] | *acceleration* [65], power [88] | *note attack* [87], articulation [84] |
| Regularity | smooth vs. rough | *jitter* [18], voice quality [18, 57] | directness [65], *phase shift* [73, 89], fluidity [47] | *microstructural irregularity* [84], timbral roughness [87] |
| Extent | small vs. large | *pitch range* [57], *loudness* [18] | *spatial expansiveness* [86, 88], contraction index [65] | *volume* [84, 87] |

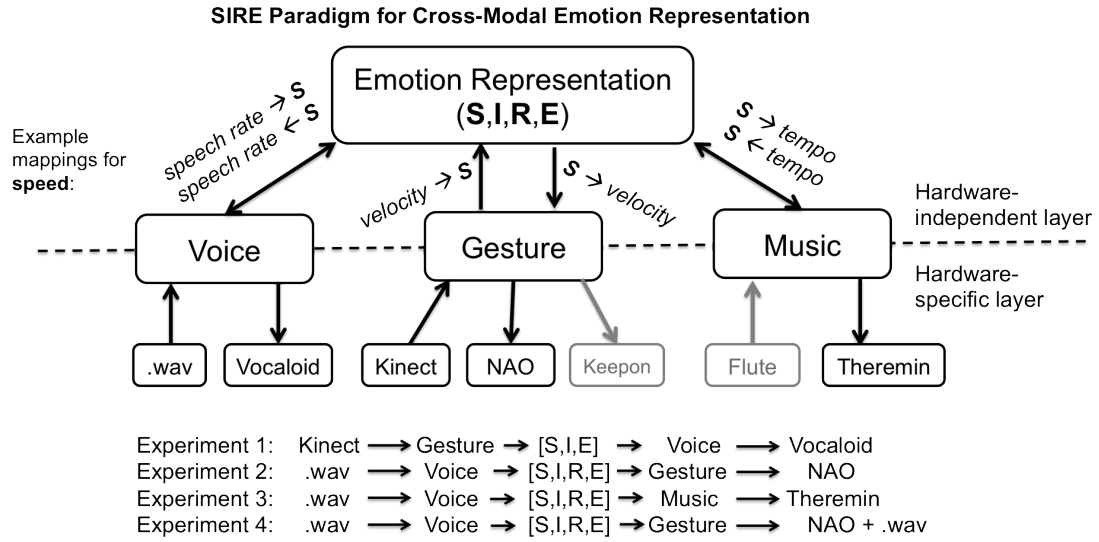



Figure 3.1: Overview of the SIRE emotion framework, in which emotions are represented only though speed, intensity, regularity, and extent. Greyed boxes show other input/output types as examples for future work.

1. *SIRE representation*, 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 hardware-specific implementation.

The *parameter mappings* can be divided into two layers (see Figure 3.1):

- *Hardware-independent layer*: A mapping from SIRE to perceptual features. These mappings are those outlined in Table 3.1.

- *Hardware-specific layer*: A mapping of perceptual features to a hardware-specific implementation.

We have implemented the SIRE framework on three systems representing three modalities:

1. **Voice**: HRP-2 singing with Vocaloid real-time opera synthesizer (Figure 3.2, used in [90])

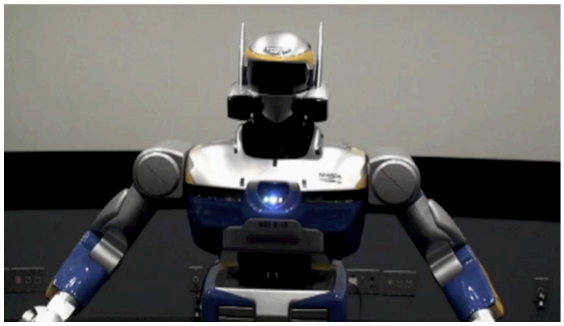


Figure 3.2: HRP-2 singing robot: platform for gesture to voice experiment ( [5], p. 8)

Figure 3.3: NAO gesturing robot: platform for voice to gesture experiment ( [5], p. 8)

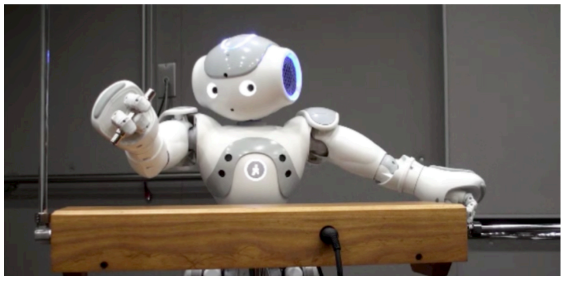


Figure 3.4: NAO thereminist: platform for voice to music experiment ( [5], p. 8)

2. **Gesture**: NAO[1] gesturing robot (Figure 3.3, reported in [6]).

3. **Music**: NAO theremin-playing robot (Figure 3.4, based on [91]).

### 3.3.1 Emotional voice using SIRE

We describe here how we implement analysis and generation of emotional voice using the SIRE model.

We have tested several hardware-independent mappings:

- **Speed**: speech rate
- **Intensity**: voice onset rapidity
- **Regularity**: voice quality
- **Extent**: pitch range, loudness

We implement the hardware-dependent mappings as follows.

#### Analysis of emotional speech waveforms

In this section, we assume an input speech sample $x(t)$ with sample rate $f_s$ and length $N$. In the experiments in this chapter, 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 [92]. For this reason, we manually provide the number of syllables $b$ for the experiments in this chapter. 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. In Chapter 5, we explore extracting syllables with an Automatic Speech Recognition system.

*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:

[1] www.aldebaran-robotics.com

$$p(k) = \sum_{i=0}^{n-1} x(k \cdot n + i)^2 \tag{3.1}$$

and onset rapidity is:

$$\max_{k=1,\ldots,N/n} p(k) - p(k-1). \tag{3.2}$$

*Regularity* is mapped here to the inverse of jitter in the voice sample, as jitter has been related to vocal "roughness" in [93]. Jitter is defined for each utterance as:

$$\frac{1}{N-1} \sum_{t=1}^{N} |x(t) - x(t-1)| \tag{3.3}$$

*Extent* is defined as the range of pitch in the speaker's voice. We used the Snack sound toolkit[2] implementation of the average magnitude difference function (AMDF) [94], an autocorrelation function, to extract the utterance's F0 trajectory, taking extent as the difference between the lowest F0 and the highest F0.

In this chapter, scaling of SIRE parameters was performed in a simple linear fashion. Given the minimum and maximum values for each parameter (experimentally chosen), we linearly scale to achieve a parameter between 0 and 1. For instance, pitch range was linearly scaled between a minimum F0 of 40 Hz and a maximum F0 of 255 Hz. As for speed, we used a minimum speech rate of 2 syllables per second and a maximum speech rate of 7 syllables per second, based on our input set. In Chapter 4, we study how scaling can be adapted to the speaker by defining, for instance, extent as the user's deviation from their pitch average.

**Generation of voice using Vocaloid**

We use Yamaha's Tonio Vocaloid singing synthesis software [95] to output a voice for our robot. It takes a list of phonemes and note values as input, and can adjust parameters such as tempo and speed of note onset in real-time. Although meant for synthetic opera singing, its flexibility is advantageous for reproducing speech utterances. For instance, it has an advantage over typical Text-To-Speech software since parameters such as pitch, brightness, note attack and delay times are readily accessible for each note. Here, we

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

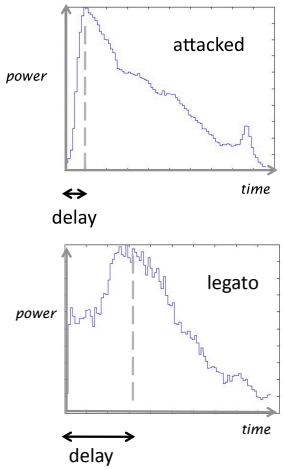



Figure 3.5: Illustration of power envelopes for attacked and legato articulations, taken from flute samples. Attacked articulations have higher intensity than legato.

define a note as a syllable-pitch-length triple. For example, the utterance *"I am go-ing to the store"* would contain 7 notes. Defining the pitch and length values of each note is a simple way to add prosody.

*Speed* is defined here as the inter-onset-interval (IOI) between notes, which is inversely proportional to tempo. For instance, 60 BPM is equivalent to IOI = 1000ms. Given an SIRE $(S,I,R,E)$, we can calculate the corresponding speech tempo as:

$$\text{IOI}(S) = \text{IOI}_{MIN} + (1-S) * (\text{IOI}_{MAX} - \text{IOI}_{MIN}) \tag{3.4}$$

where $\text{IOI}_{MAX}$ and $\text{IOI}_{MIN}$ correspond respectively to the IOI of the slowest and fastest speech possible. We also experimentally found these parameters, taking into account the length of each note, since this can also affect perceived speech rate.

*Intensity* of speech is characterized here as the articulation of each syllable. A smooth, low intensity utterance would sound *legato*, whereas a high intensity utterance, with loud emphasis on every syllable, would sound attacked. Thus, we define intensity using the concepts of note onset delay. Figure 3.5 shows an illustration of two types of notes: attacked and legato. Notice that the delay (time to reach a maximum volume) for an attacked note is shorter than that of the legato. Like tempo, we have an inverse relationship: a shorter delay makes a higher intensity. We define delay, given an intensity $I$, as:

$$\mathrm{DEL}(I) = \mathrm{DEL}_{MIN} + (1-I) * (\mathrm{DEL}_{MAX} - \mathrm{DEL}_{MIN}) \tag{3.5}$$

We experimentally set $\mathrm{DEL}_{MIN}$ to 100 ms and $\mathrm{DEL}_{MAX}$ to 400 ms. Additionally, we found that changing the decay parameter (the "falloff" time after the note's maximum volume has been reached) to match $\mathrm{DEL}(I)$ resulted in an attack that was easier to perceive.

*Regularity* is modulated using the Vocaloid breathiness controller: the more breathy the voice, the less regular. Breathiness, given regularity R, is defined similar to intensity:

$$\mathrm{BRE}(R) = \mathrm{BRE}_{MIN} + (1-R) * (\mathrm{BRE}_{MAX} - \mathrm{BRE}_{MIN}) \tag{3.6}$$

The *extent* of a human voice could be understood semantically in two ways: the volume range in decibels, or the F0 range. Some languages have wider or narrower F0 ranges to accommodate tones; to be as general as possible, we assign range to volume, though F0 range could be interesting to explore in the future. Due to our use of MIDI for sending Vocaloid parameters, we send this as the volume in the MIDI message range $[0, 127]$ along with NOTE_ON messages.

$$\mathrm{VOL}(E) = E * 127 \tag{3.7}$$

### 3.3.2 Emotional gesture using SIRE

Similarly, we have implemented systems for both analysis and generation of gesture.

#### Analysis of emotional gesture using 3D hand tracking

Our first step is to extract a SIRE representation from human gesture. In practice, we use the Microsoft Kinect, a color and depth sensor, to extract the pose of a human's body in 3-dimensional space. We simply use the OpenNI library to extract the $(x, y, z)$ positions of the human's left and right hands over a time sequence. Our SIRE values $(S, I, R, E)$ are calculated as follows.

*Speed* is calculated as the average rate of change in hand position from frame to frame. Let $(x_L(f), y_L(f), z_L(f)) = H_L(f)$ be the left hand's position in $\mathbb{R}^3$ at frame $f$,

and $(x_R(f), y_R(f), z_R(f)) = H_R(f)$ be the right hand's position at frame $f$. Then we can calculate velocity as the Euclidean distance divided by the time $t$ elapsed between two subsequent frames, $f-1$ and $f$.

$$V_R(f) = \frac{||H_R(f) - H_R(f-1)||_2}{t_f - t_{f-1}} \tag{3.8}$$

$$V_L(f) = \frac{||H_L(f) - H_L(f-1)||_2}{t_f - t_{f-1}} \tag{3.9}$$

Next, we find the maximum velocity of either hand achieved over the last $k$ frames,

$$V_{MAX(k)}(f) = \max_{i \in \{f,\ldots,f-k\}} \max_{h \in \{L,R\}} V_h(i) \tag{3.10}$$

Finally, we set the speed to be the maximum velocity over the last $k$ frames, normalized by a maximum velocity $V_{MAX}$ of a human hand movement, found experimentally. For experiments, we set this to $V_{MAX} = 5$m/s.

$$S_k(f) = V_{MAX(k)}(f)/V_{MAX} \tag{3.11}$$

*Intensity* is defined as the acceleration of the hands. Acceleration is found in the same manner as velocity; maximum acceleration over the last $k$ frames $A_{MAX(k)}(f)$ is calculated by replacing $H_R(f)$ and $H_L(f)$ in Equation (3.8) and Equation (3.9) with $V_R(f)$ and $V_L(f)$, respectively.

We then define intensity as the normalized acceleration. Experimentally, we set $A_{MAX} = 50m/s^2$

$$I_k(f) = A_{MAX(k)}(f)/A_{MAX} \tag{3.12}$$

*Regularity* may be defined as the movement correlation between the hands (i.e., higher regularity when the hands are moving synchronously) or the smoothness of the trajectory. This parameter has not been implemented in our experiments, but we refer the reader to the implementation in [65], in which a similar *directness* parameter is used.

*Extent* is defined as the Euclidean distance between the two hands for a given frame $f$, which we call extent $E(f)$, normalized by the maximum distance when hands are

spread apart, $E_{MAX}$, also found experimentally. This also could have been defined differently, such as maximum 3D volume taken by the human's pose. For our experiments, we define it simply as the extent between the two hands:

$$E(f) = ||H_L(f) - H_R(f)||_2 \tag{3.13}$$

$$R_k(f) = E_{MAX(k)}(f)/E_{MAX} \tag{3.14}$$

**Generation of gesture on NAO humanoid**

In this section we describe how we implement the perception of speed, intensity, regularity and extent on the NAO humanoid robot.

In our implementation, 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 $t_0, t_1, t_2$ (Figure 3.6). 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 $R$ is implemented either as joint phase shift and directness, which can be thought of as temporal and spatial regularity respectively; for arms, a more irregular movement is created by temporally "shifting" one of the arm movements (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 $\delta_{s1}$, $\delta_{s2}$ 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).

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

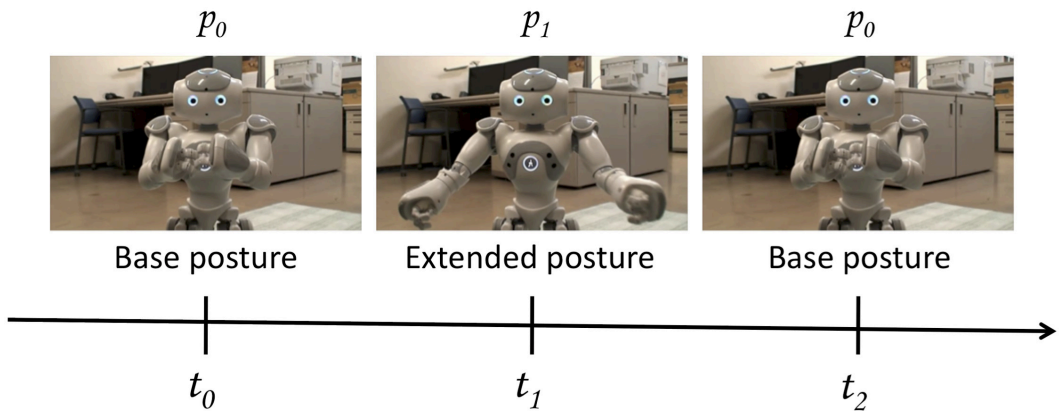



Figure 3.6: Timeline of an arm gesture ( [6], p. 4)

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 $\delta_t$ 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 $\kappa_0$ is the base configuration (pitch and yaw values) of the head, and $\kappa_1$ the extended posture. For the head, we find $\delta_{s1}$ and $\delta_{s2}$, 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.

**Algorithm 1** MAPSIRE $\left\langle \mathscr{G}_{\text{NAO,left arm}} \right\rangle$

**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: $g_{out}.t_0 = t_0$
4: $g_{out}.t_1 = \max((1-S) \cdot I \cdot t_1, \underline{m})$
5: $g_{out}.t_2 = \max((1-S) \cdot t_2, \underline{m})$
6: **return** $g_{out}$

**Algorithm 2** MAPSIRE $\left\langle \mathscr{G}_{\text{NAO,right arm}} \right\rangle$

**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 $\left\langle \mathscr{G}_{\text{NAO,head}} \right\rangle$

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

Figure 3.7: Arm base position ( [6], p. 5)

Figure 3.8: Arm extended posture ( [6], p. 5)

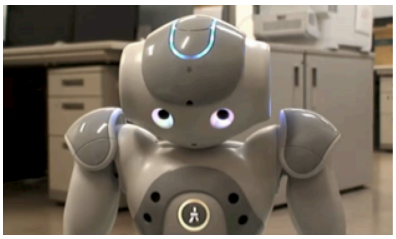


Figure 3.9: Head start position ( [6], p. 5)

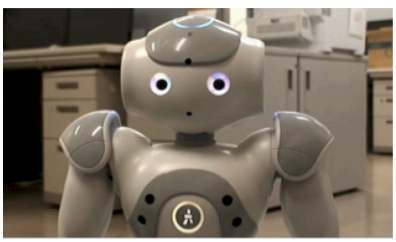


Figure 3.10: Head extended posture ( [6], p. 5)

### 3.3.3 Emotional music using SIRE

In this chapter, we also test the ability of SIRE to generate emotional music. Analysis of music has not been tested at the time of writing. Instead, we refer the reader to the analysis of expressive music in a programming by playing system [61] and give some insight into mappings with the analysis of a human thereminist playing in five expressive styles.

The target instrument for our music experiments is the theremin. The theremin is an electronic instrument with two antennas: a vertical one for pitch and a horizontal one for volume. A theremin player can change the pitch and volume by moving their hands closer to or farther from the antenna. Due to its non-linear pitch dynamics, it is a difficult instrument to master.

#### Analysis of music played by humans

We asked a semi-professional thereminist to play a scale in each of neutral, happy, sad, angry and fear styles. The characteristics of each style can be described qualitatively as follows.

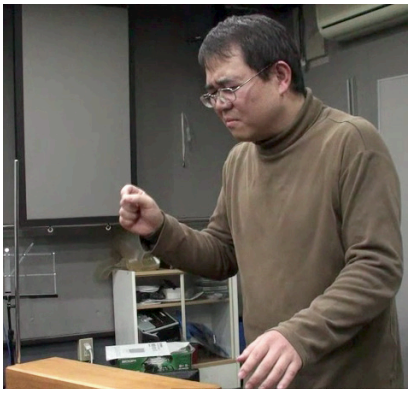
- Neutral: Even notes; the volume did not change (Figure 3.11 (a))
- Sad/wistful: A bit of vibrato, legato (Figure 3.11 (b))
- Happy/fun: Fast, lots of short volume bursts (Figure 3.11 (c))
- Angry: Plenty of vibrato and loud, attacked (Figure 3.11 (d))
- Fearful: Plenty of vibrato, quiet (Figure 3.11 (e))

Based on the thereminist's feedback, we propose mappings as follows:

- Speed: tempo
- Intensity: attack speed
- Regularity: vibrato rate
- Extent: variance of volume

(a) (b)

(c) (d)



(e)

Figure 3.11: (a) Neutral (b) Happy (c) Sad or wistful (d) Angry (e) Fearful

**Generation of emotional music played by robot thereminist**

We implement a theremin player module on the same NAO humanoid used in gesture experiments. The NAO robot first calibrates itself with the theremin's pitch by recording a sweep of points with its right arm (4-DOF) and performing pitch detection using subharmonic summation [96]. Along with a relatively long frame width of 8192 samples, we found the detection was robust to the noise from NAO's head-mounted microphones. Once pitch detection is complete, one of two interpolations methods [91] are used to retrieve the pose corresponding to a particular pitch: 1) parametric interpolation requiring approximately 12 recording points, or 2) a look-up table with linear interpolation between points, needing approximately 50-70 recording points.

In our experiments, we controlled the NAO's play using concepts from our previous implementations in gesture and voice.

*Speed* is modulated in a similar way to the IOI of Vocaloid in Section 3.3.1. The higher the speed, the shorter the IOI.

*Intensity* is implemented in the same way as intensity for the NAO gesturer as described in Section 3.3.2. The effector is the left (volume-controlling) hand. The base position is set to zero volume, that is, the hand is close to the volume antenna. The extended position is set to the maximum volume, approximately 20cm away from the antenna. The perceptual result is that higher intensity sounds like a faster rise in power.

*Regularity* is also modulated using the gestural concepts from Section 3.3.2 to temporally offset each gesture (in this case, one gesture is equivalent to a note onset). As described with the head effector, we determine the offset by defining a random variable taken from a normal distribution with variance inversely proportional to R. Instead of using this offset value for *spatial* offset, we use it to *temporally* offset the note before or after the onset time.

*Extent* changes the maximum volume of NAO's play, similar to extent in NAO gesturer. This affects the location of the extended position defined for intensity.

## 3.4 Experiments

### 3.4.1 The SIRE emotion transfer system

The SIRE model is best evaluated by developing an emotion transfer system. This means that SIRE parameters are extracted from one modality and are used to generate emotional expressions in another. In each of three experiments, we *extract* a SIRE from human portrayals of emotion, and use that SIRE to *generate* robot portrayals in a different modality. Both the source and generated portrayals are then evaluated by human raters. If both the source and generated portrayals are rated as the same emotion, then we can say that SIRE is sufficient to represent that emotion across the two modalities.

Using this emotion transfer system as a research paradigm, we wish to answer the following research questions:

- **Q1**: Does the same emotion in two different modalities have similar SIRE values?
- **Q2**: If so, what are the SIRE values and recognition rates for each emotion?
- **Q3**: How well do the implementations (described in Section 3.3) map to the SIRE parameters?
- **Q4**: What is the effect of using SIRE as a basis for expression in multiple, simultaneous modalities?

We performed a pilot study and four experiments to address these questions. They are as follows.

- **A pilot study: Gesture to voice via SIE.** A pilot experiment using only 3 parameters of speed, intensity, extent (SIE) from human gesture to synthesized voice **(Q1,Q2)**
- **Experiment 1: Evaluation of SIRE perceptual mappings.** Testing parameter mappings for gesture through perceptual surveys **(Q3)**
- **Experiment 2: Voice to gesture via SIRE.** Testing all 4 SIRE parameters from emotional voice to robot gesture **(Q1,Q2)**

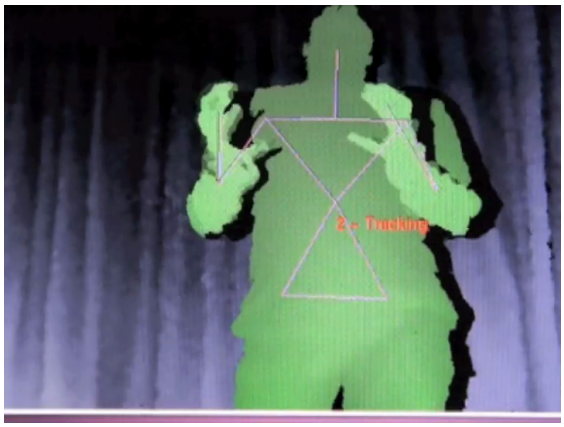



Figure 3.12: Body pose estimation using the Kinect 3D sensor.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Relative length* | 0.2 | 0.2 | 0.3 | 0.2 | 0.2 | 0.5 | 0.2 | 0.2 | 0.5 | 0.2 | 0.2 | 0.7 | |
| | I'm | go-ing | to | the | store, | do | you | want | a-ny-thing? | | | | |
| *Note/Octave* | G2 | G2 | G2 | G2 | G2 | G2 | G2 | G2 | A2 | B2 | B2 | C3 | |

Figure 3.13: Prosody for the utterance ( [5], p. 6)

- **Experiment 3: Voice to music via SIRE.** Testing all 4 SIRE parameters from emotional voice to theremin-playing robot **(Q1,Q2)**

- **Experiment 4: Multi-modal expression of emotion.** Testing the combination of voice and gesture with the same SIRE values **(Q4)**

### 3.4.2 A pilot study: Gesture to voice via SIE

We asked four naive students (3 male and 1 female) from Kyoto University to generate gestural portrayals of happiness, sadness, anger and fear in front of a 3D sensor. Each emotion was to be acted for around 5-10 seconds and their anonymized gestures recorded with a standard video camera (as in Figure 3.12). The participants were not professional actors, but scenarios were provided to help elicit a desired emotion (e.g., "You have just won the lottery. Convey your happiness to the robot").

The body pose in 3D was detected as in Figure 3.12, and the maximum speed, acceleration and extent of the participants' hands were extracted for each performance. Our program converted these values to SIE by linearly scaling them between 0 and 1 based on maximum speed, acceleration and distance, respectively. The minimum and maximum values were experimentally set prior to the experiment.

Table 3.2: Pilot study parameter mappings

| Gesture mapping | Parameter | Voice mapping |
|---|---|---|
| Hand Velocity | Speed | Tempo |
| Hand Acceleration | Intensity | Attack (onset delay) |
| Inter-hand Distance | Extent | Volume |

As output, the Vocaloid [95] synthesized male opera singer, Tonio was used. We chose a neutral utterance string: "I'm going to the store. Do you want anything?". The phrase was given the hand-made prosody as shown in Figure 3.13 to match the sentence semantics. Then, the extracted SIE triples were given as input to the voice module as per Table 3.2. The vocal utterances were recorded as videos with the robot head and shoulders in the frame, as in Figure 3.2.

The 16 human gesture videos and corresponding 16 robot voice videos were uploaded to the Internet in the form of an anonymous survey. Rating was performed in a forced-choice manner; according to [97], forced-choice judgments give results similar to free-labeling judgments when evaluators attempt to decode the intended emotional expression. After watching a video of either a human or speaking robot, the evaluator was asked to select the emotion most conveyed, among happiness, anger, sadness and fear. An average of 9 evaluations for each display of emotion, and 37 for each emotion class were collected for this pilot study.

**Results and discussion**

The results of the emotion evaluations can be visualized in the confusion matrices in Figure 3.14. The visualized confusion matrix here can be thought of as a distribution of perceived emotion for a given portrayal class. The intended emotion is shown in the titles, and the average percentage of raters that selected each emotion are given along the dimensional axes. For instance, Figure 3.14-1a shows that human portrayals of happiness through gesture were recognized on average by raters as happiness by 53% of raters, anger by 22%, sadness by 13% and fear by 13%. In this way, the graphs can also be interpreted as componential representations of emotion portrayals.

We look in particular for similar distribution shapes in each column – this would in-

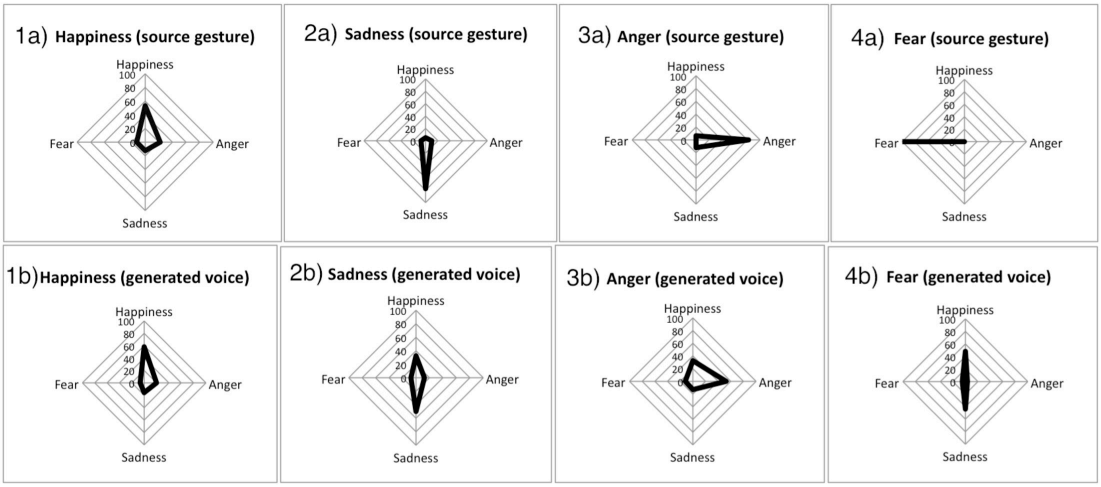



Figure 3.14: Pilot study: Visualization of confusion matrices for gesture and voice. Intended emotion is shown in the titles, and the average percentage of raters that selected each emotion are given along the dimensional axes. Pointed triangles indicate that the one emotion was greatly perceived on average. Similar shapes for a given number indicate similar perceived emotion for both input gesture and output voice ( [5], p. 9).

dicate a similar perceived emotion for both input gesture and output voice. For instance, the voice generated using the same SIE values as Figure 3.14-1a was rated as happiness by 58% of raters, anger by 20%, sadness by 16%, and fear by 7%, as shown in Figure 3.14-1b. This large overlap, plus the result significantly over chance (25%) suggests that SIE indeed was sufficient for transferring happiness from gesture to voice.

We give here some qualitative descriptions of the gestural portrayals to help interpret the recognition rates.

*Fear* was extremely well-recognized for human gesture portrayals, but not for transferred voice. Gestural portrayals included pulling back in fear sporadically, arms clutched to chest or to their sides. Two possibilities are possible: pose may have had a great effect on the understood emotion, which could not be transferred to the voice, or the SIE parameters are not sufficient for transferring this emotion.

*Anger* was portrayed by two participants in a prototypical manner – balled fists, gestures as if hitting a table, and approaching the camera. Interestingly, these were sometimes confused with sadness, presumably looking similar to outbursts of grief. On the other hand, one participant shook a finger at the camera, and this was recognized

by 100% of raters as anger. The next well-recognized was a portrayal that fluctuated between what would be considered "cold anger" (a 'stern' forward-facing pose with one arm crossed and the other hand to the chin), and outbursts of both arms forward in a open-handed, questioning pose.

*Sadness* portrayals contained prototypical hunched shoulders, hanging arms, and relatively low energy for two participants. The hand-to-chin gesture appeared again, which was perceived by 60% of raters as sad, but by 30% as anger. This gives an indication that sadness and anger may look similar when not distinguishing between hot and cold anger. Two of the more dramatic participants showed sadness with both hands to head, and bending at the waist in pain, which were more easily recognized by the raters.

*Happiness* was an interesting case, as only one participant made huge, prototypical "jumping for joy" portrayals of happiness. Another participant danced, another one made gestures to the sky in thankfulness, and the last shook her arms to the side in excitement while running in place. Interestingly, the happy dancing portrayal was most well recognized by raters, not the "jumping for joy" gesture. In fact, many raters watching the "jumping for joy" silhouette mistook it as anger. The gestures toward the sky were often perceived as grief or sadness; and shaking arms was interpreted by half of respondents as anger.

This discussion allows us to draw three observations:

1. Happiness, sadness and anger were transferred despite the varied gestural interpretations for each emotion (e.g., jumping for joy and dancing, or fist pounding and finger wagging). Their recognition rates were all greater than chance, with the exception of fear.

2. Fear was not well transferred through SIE. We note that the irregular, sporadic backwards movements in fear portrayals could not be captured solely through speed, intensity, and range, which is one reason why we add the Regularity parameter to the remaining experiments.

3. Source gestures (without contextual, facial, or vocal information) are not perfectly recognized, and for good reason as discussed in the qualitative analysis above.

Firstly, this underlines the importance of multimodal redundancy. Secondly, this suggests that studies should not aim at perceiving one "correct" transferred emotion at high rates, but also focus on the distribution of recognition, as in Figure 3.14. For instance, if a gesture is rated as 50% angry looking and 50% happy, the vocal output should best also be 50% angry and 50% happy. We touch further on the idea of emotion scores in Chapter 4.

### 3.4.3 Experiment 1: Evaluation of SIRE perceptual mappings

In this experiment, our goal was to verify that our SIRE mappings are supported by 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 extension movement of the arms.

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?

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.

| 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 |

Table 3.3: 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). ( [6], p. 5)

**Results and discussion**

As Table 3.3 shows, the mappings as described agree with raters' perceptions at more than 86% in all cases. 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. abrupt. The word "regular" was also understood as "normal" (i.e., not like a robot) by at least one evaluator. Like intensity, we suggest to use another description to evaluate regularity, like smooth vs. rough.

### 3.4.4 Experiment 2: Voice to gesture via SIRE

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 16 audio samples taken from the Berlin Database of Emotional Speech[3], which is a database of emotional speech recorded by professional German

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

Table 3.4: Experiment 2 parameter mappings

| Voice mapping | Parameter | Gesture mapping |
|---|---|---|
| Syllable rate | Speed | Arm velocity |
| Voice onset rapidity | Intensity | Arm acceleration |
| Jitter | Regularity | Inter-arm phase shift |
| Pitch range | Extent | Gesture extent |

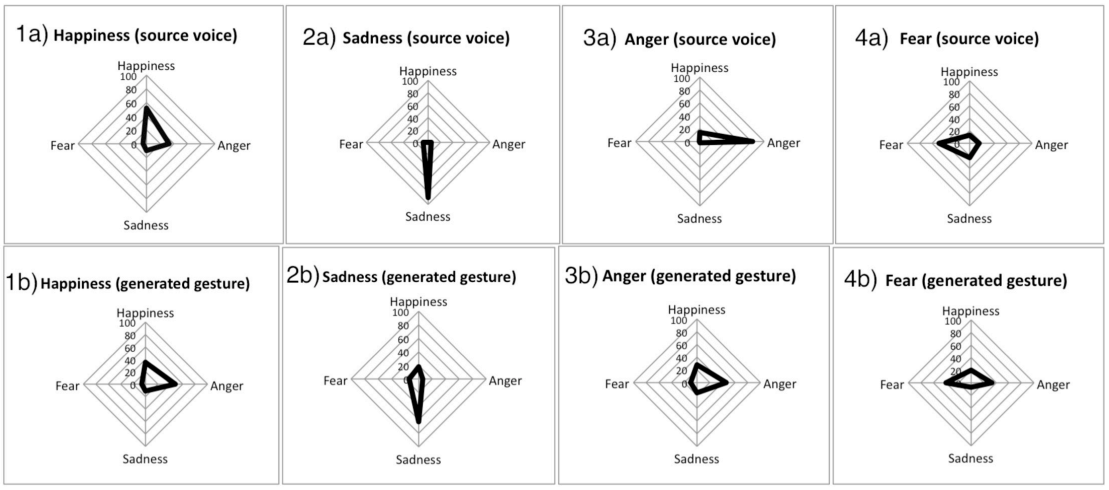



Figure 3.15: Experiment 2: Visualization of confusion matrices for voice and gesture. Similar shapes for a column indicate similar perceived emotion for both input voice and output gesture ( [5], p. 10).

actors. Each sample was a normalized wave file at 16kHz, 1.5 to 3.9 seconds long, all of the same sentence. Four samples each of happiness, sadness, fear, and anger were used, all with recognition rates of 80% or higher by German evaluators.

Given SIRE values extracted from these audio samples as per Table 3.4, we generated 16 movement sequences using a simulated NAO shown on a projected screen. A full description of implementation can be found in [6]. Only one type of gesture was used (an extension of both arms in front of the robot), repeated four times in series for each sequence. The sequences were shown in a random order to participants in a classroom. After each sequence, the participants were given 5 seconds to choose one of happiness, sadness, anger, or fear in a forced-choice questionnaire.

**Results and discussion**

Figure 3.15 shows the confusion matrices for emotional voice and gesture. Ratings of the German voices is taken from the result of a stationary, speaking robot outlined in [6]. We find that the recognition rates for all emotions are significantly greater than chance (25%), suggesting that the SIRE framework indeed converts the source vocal emotion to the same emotion in gesture. On the other hand, we can see that happiness (Figure 3.15-1b) was not clearly distinguished from anger. Further work in [6] suggested interaction with a pose cue: the immobile head of the robot. When compared with portrayals with a moving robot head, the staring, forward-facing head of the robot was significantly rated more often as anger.

### 3.4.5 Experiment 3: Voice to music via SIRE

Thirty-four participants were recruited over the Internet without respect to cultural or musical background. Average age was 33.2, stddev=12.2. Eight speech files (2 for each emotion) from the set of those used in Experiment 2 were used as input. Self-reported musical experience shown that 35% of raters had no musical experience, 38% were beginner level, 21% intermediate level, and 6% expert.

The output was generated by the NAO robot playing the theremin with the parameter mappings as shown in Table 3.5. The robot's right arm was set to control the pitch at 415Hz. To avoid bias based on song mode (e.g. major or minor), the robot played a simple sequence of quarter notes at the same pitch. This is a standard evaluation method used also in [87]. The left arm of the robot controlled the note volume, which started, shaped and ended the notes. The sounds of the theremin were recorded as sound files and uploaded to the internet in the form of anonymous survey.

Raters were first asked to calibrate their headphones or speakers so that they could hear the loudest and quietest samples comfortably. They were then asked to rate the sounds produced by the NAO thereminist in a forced-choice response. No image was provided.

Table 3.5: Experiment 3 parameter mappings

| Voice mapping | Parameter | Music mapping |
|---|---|---|
| Syllable rate | Speed | Tempo |
| Voice onset rapidity | Intensity | Note onset rapidity |
| Jitter | Regularity | Note timing offset |
| Pitch range | Extent | Maximum volume |

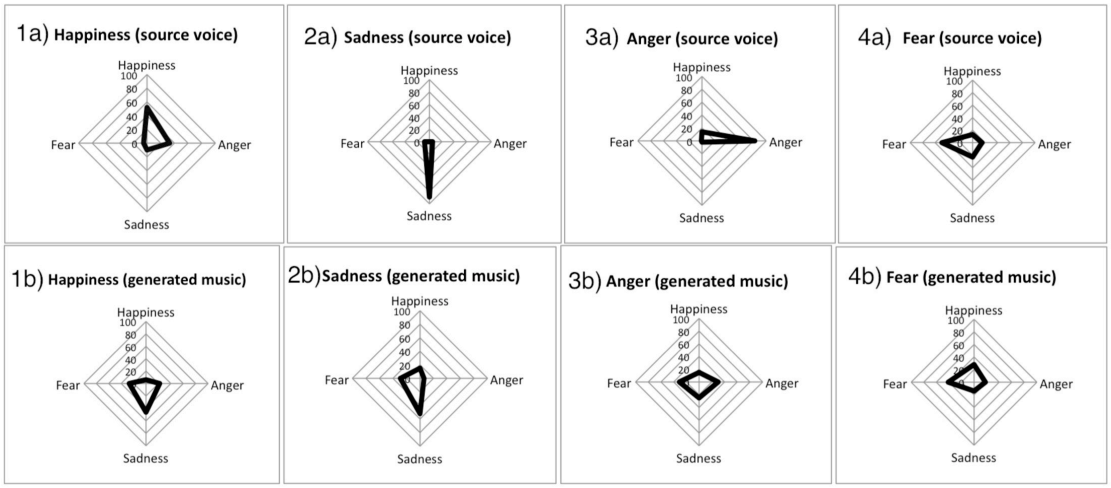



Figure 3.16: Experiment 3: Visualization of confusion matrices for voice and music. Similar shapes for a column indicate similar perceived emotion for both input voice and output music ( [5], p. 10).

**Results and discussion**

The results of the music experiment are shown in Figure 3.16 in the usual confusion matrix visualization format. We can see that the effectiveness of SIRE using the theremin modality is limited compared to speech and gesture. In particular, happiness and anger could not be reliably expressed. One reason for this may be the theremin sound itself. The theremin is often used for science fiction or horror films due to its eerie timbre, or for romantic, wistful songs such as Rachmaninoff's Vocalise. We find that overall, the evaluations of this modality were skewed towards 34% sadness and 32% fear, whereas only 16% and 19% of all portrayals were perceived as happiness or anger, respectively. It is possible that the maximum speed of the theremin was a limiting factor – unlike instruments such as piano or flute, the theremin cannot change notes quickly without the sounds becoming indistinct.

We found that there are certain SIRE parameters which allow the robot to play music that is perceived as sad: 62% of raters recognized as sadness with SIRE=(0.12, 0.44, 0.72, 0.42). In addition, SIRE=(0.95, 1.0, 0.13, 0.37) produced a performance recognized as fear by 53% of evaluators. In Experiment 2, these same SIRE parameters produced emotional gestures that were recognized as sadness at 76% and fear at 65%. These results, coupled with the fact that the source of these parameters were sad and fear voices, suggest that emotions can be captured through SIRE across three modalities. Further experiments with a more versatile musical instrument such as piano are needed to confirm the effectiveness for happiness and anger.

### 3.4.6 Experiment 4: Multi-modal expression of emotion

In this experiment, we assessed the usefulness of the SIRE system for multi-modal emotional expression. We compare the emotion recognition of 1) a humanoid playing a voice only with 2) a humanoid playing a voice *and* performing the associated SIRE motion. Our hypothesis is that adding motion will increase recognition of emotions, or make the impression of the emotion stronger than through a 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

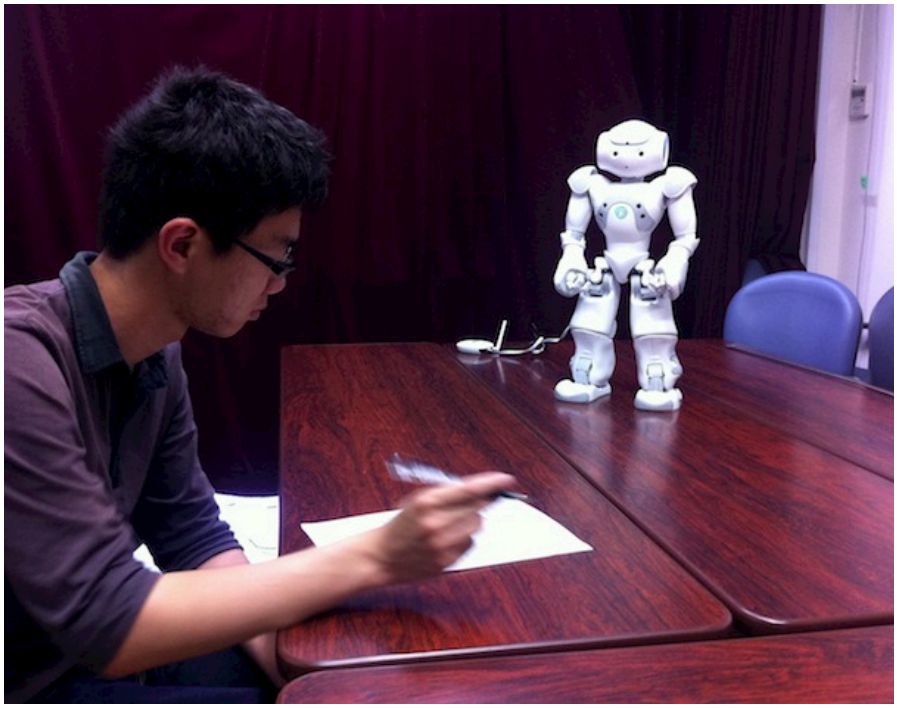


Figure 3.17: Experimental setup for experiment 4, position of robot during voice-only condition ( [6], p. 7)

with a NAO robot placed on a table as shown in Figure 3.17. 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 emotional voice utterances were the same used in Experiment 2, in German. The gesture dynamics were generated using the SIRE values extracted offline from the 16 utterances described in Experiment 2.

We presented the participants with two robot conditions.

- Condition 1: Voice only. The robot stayed still in a neutral position (Figure 3.17) 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.

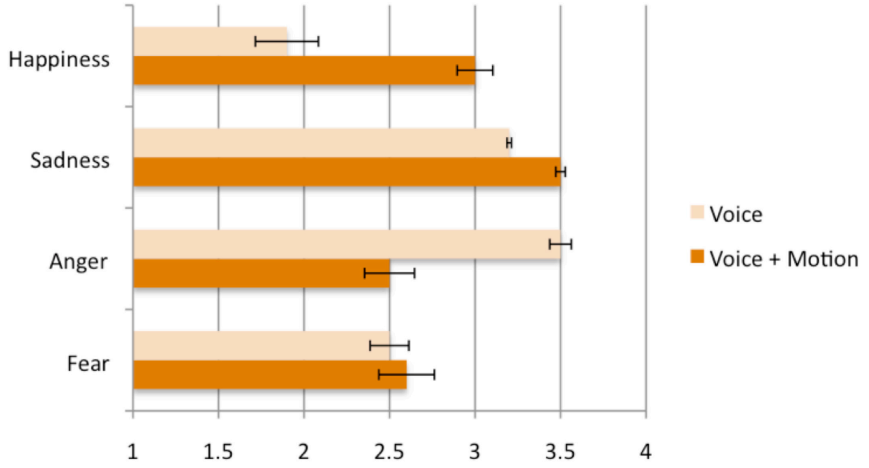



Figure 3.18: Experiment 4: Comparison of ease of understanding, from difficult (1) to easy (4), for correctly recognized samples ( [6], p. 7).

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 discussions**

In this experiment, we explore the result of expressing two modalities using the same SIRE. Since we saw in Experiment 2 that 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. 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%. This may be explained by the fact that, according to Japanese evaluator comments, purely vocal expressions of fear and sadness were easily confused between each other. However, we note 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 compounding modalities. For a given rater, when the sample was recognized correctly for both voice and voice+motion, we compared the rater's ease of understanding for the two sequences.

We see that in Figure 3.18 that both sadness and happiness were more clearly portrayed through voice+motion than through voice only. This suggests that the use of a SIRE-laden gesture may be most useful for humanoids to portray sadness and joy.

In Figure 3.18, we also notice the inverse effect for anger. 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 voice-only 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 [98] [99]. This suggests that to provide reliable anger movements the head should remain still (i.e., only arm gestures should be used). Experiment 2 gives further evidence to this hypothesis: recognition of anger was relatively high, and the samples only used arm and not head 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.

## 3.5 Summary

In this chapter, we saw that the SIRE paradigm showed promise in finding emotion "universals" across voice, movement, and music, as summarized in Table 3.6. This was tested by mapping high-level perceptual features to low-level features, such as speed to speech rate, arm velocity, or tempo. By extracting the SIRE dynamics from a voice and mapping it to a gesture, we found that, for instance, an expression of sadness is slow with low-medium intensity (0.12, 0.44, 0.72, 0.42), whether expressed in voice, gesture or music. Fear was fast and intense, with low regularity (0.95, 1.0, 0.13, 0.37). While SIRE was tested for particular values of speed, intensity, regularity 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. This will be addressed in the next chapter.

| Emotion | Human voice (%) | Robot gesture (%) | Robot music (%) | S | I | R | E |
|---|---|---|---|---|---|---|---|
| Happiness | 43 | 62 | 6 | 0.72 | 0.2 | 0.22 | 0.73 |
| Sadness | 95 | 76 | 76 | 0.12 | 0.44 | 0.72 | 0.42 |
| Anger | 95 | 86 | 27 | 0.71 | 0.46 | 0.04 | 0.73 |
| Fear | 33 | 43 | 53 | 0.95 | 1 | 0.13 | 0.37 |

Table 3.6: Emotional sequences with agreement among evaluators, and their corresponding SIRE values. Low scores for happiness and anger in music may be explained by the difficulty of the musical instrument (theremin) to express these emotions in general ( [5], p. 11).

# 4

# Multimodal Emotional Intelligence (MEI)

*"When you're curious, you find lots of interesting things to do."*

– Walt Disney

In this chapter, we extend the insights from the last chapter on expression of emotion through dynamics. How can a robot not just express, but gain a deep, multimodal emotional intelligence in voice, movement, and even music?

## 4.1 MEI based on the SIRE model

The aim of the Multimodal Emotional Intelligence (MEI) system is apply emotional intelligence to new situations. As mentioned in Chapter 3, humans have the ability to recognize emotion in 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., [100]), an emotion module for a novel language (e.g., cross-language emotion recognition [101] is a recent topic), an emotion module for the operatic singer (many emotion recognition systems exist for music [102], 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

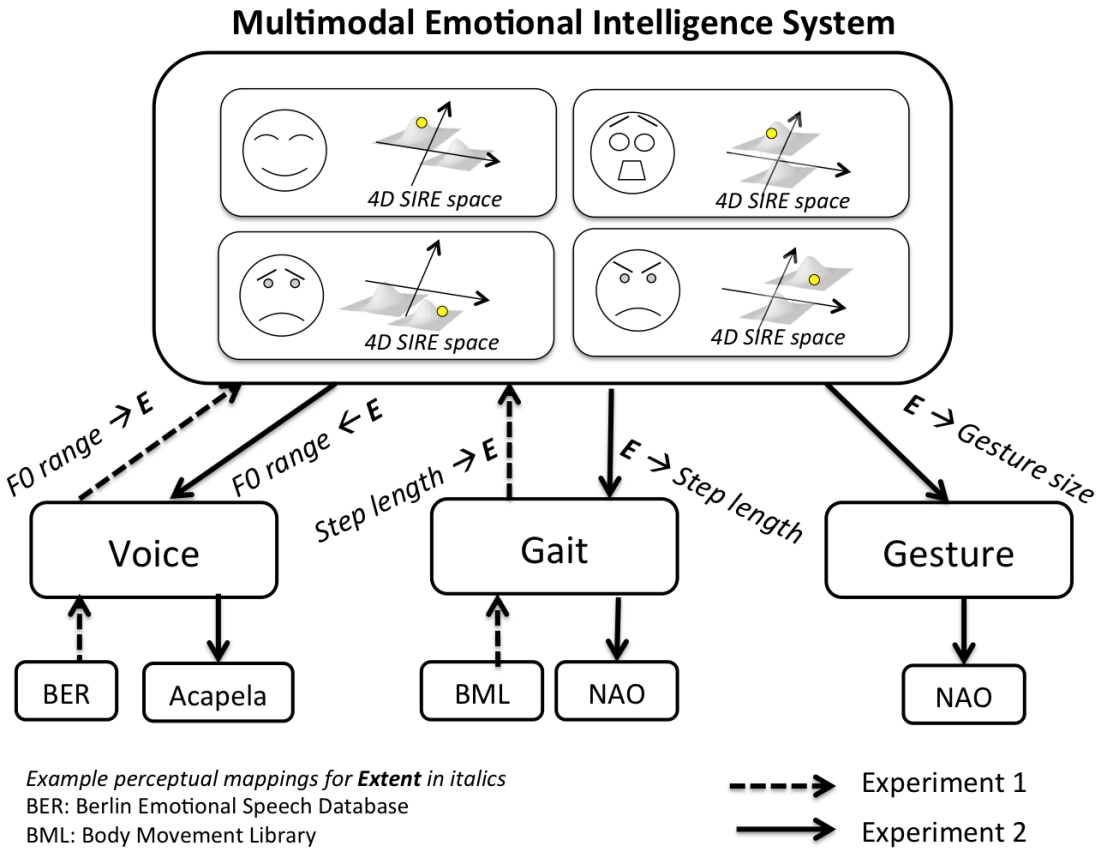



Figure 4.1: The present system performs cross-modal recognition and expression based on a GMM representation. In Experiment 1, we test how well the model trained with emotional voice can recognize emotional gait. In Experiment 2, we use the model to generate emotional voice, gait and gesture.

the emotional voice expression module [13]. This means that emotional voice input, though recognized, will never improve the way the robot's own emotions are expressed.

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, (3) for analysis and synthesis. A model that fulfills these requirements remains an open problem according to a recent review of affect models [31]. To address this challenge, the MEI system uses a (1) 4-dimensional, (2) cross-modal SIRE [6] emotion paradigm, coupled with a statistical Gaussian Mixture Model (GMM) capable of both (3) recognition and expression.

In this chapter, SIRE stands for Speed, Intensity, irRegularity and Extent, where the tuple contains four values $(S,I,R,E)$ on $[0,1]$. While SIRE was tested for particular values of speed, intensity, irregularity and extent, we need to check whether the same results emerge with a large number of training samples. 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 Figure 4.1). For each emotion class $C$ of *happiness, sadness, anger* and *fear*, we define an $m$-mixture Gaussian in 4D SIRE space,

$$SIRE\_Emotion_c(X_c) = \sum_{k=1}^{m} \pi_k \mathcal{N}(X_c|\mu_k, \sigma_k) \tag{4.1}$$

where $X_c$ is a vector of SIRE tuples corresponding to the class C, and $m$ is the optimal number of components to minimize the Bayesian Information Criterion (BIC) over $X_c$ [103]. The above four emotion classes are the focus of the present study for two reasons. Firstly, we study emotions verified in infants less than one year old [35], as motivated in Chapter 2. 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 [104]. Secondly, 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 [18].

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 [105], K-means [106], or linear regression models [107], 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 chapter). And, a GMM trained on happy SIRE tuples can be sampled when the robot wishes to express joy (Experiment 2 of this chapter), while avoiding repetitious values. Secondly, a GMM provides interpretability. Like prototype methods [108], 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 Figure 4.8). 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, in Figure 4.8.) 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 [109] or contextual information) which could further differentiate between the top confusions.

## 4.2 Training MEI

An overview of the training of the MEI module is given in Figure 4.2. 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

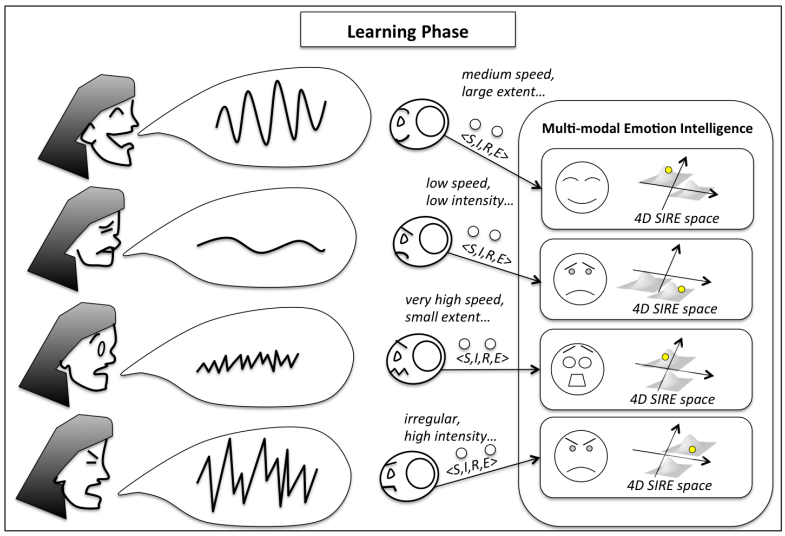



Figure 4.2: Overview of the learning phase of MEI. The robot (center) observes an emotional voice and extracts speed, intensity, irregularity and extent (SIRE) from its auditory input. This SIRE tuple is added to the relevant class model, strengthening the association between the class and vocal dynamics. In our experiments, the emotion is represented as a class tag, but it could be replaced other types of ground truth such as the output from a face recognition system, or the robot's internal state ( [7], p. 2) as discussed in Chapter 5.

| 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) |

Table 4.1: Low-level feature to SIRE mappings ( [7], p. 6).

into account that individuals may have varying speaking styles, for example.

3. **Training** the models in SIRE space using expectation-maximization.

### 4.2.1 Voice feature extraction

In this chapter, we select the features in Table 4.1 to map voice and gait to SIRE parameters. In general, we extract maximum values because they were shown to be highly relevant (more so than a mean value) in a cross-lingual emotion recognition task [110]. We also examine samples with a maximum length of 15 seconds, to roughly parallel the length of short-term memory [111].

We use the Snack Toolkit[1] to extract the following features from a given recorded utterance. In Chapter 5, we will explore the use of the HARK robot audition system [112] 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 (5-8kHz) 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.

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

### 4.2.2 Mapping 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. Non-linear 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 $-2\sigma$ or greater than $2\sigma$ are assigned to 0 and 1, respectively. Specifically,

$$x_{s'} = \begin{cases} 0 & \text{if } x_s \leq -2\sigma \\ 1 & \text{if } x_s \geq 2\sigma \\ 0.5 + \dfrac{x_s - \mu(X_s)}{4 * \sigma(X_s)} & \text{otherwise,} \end{cases} \tag{4.2}$$

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. 4.1). 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".

### 4.2.3 Training

The above mapping procedure results in a multi-speaker dataset $X_C$ which contains 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 [113]. In our experiments, the SciKit Learn Toolkit [114] 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 (Figure 4.4).

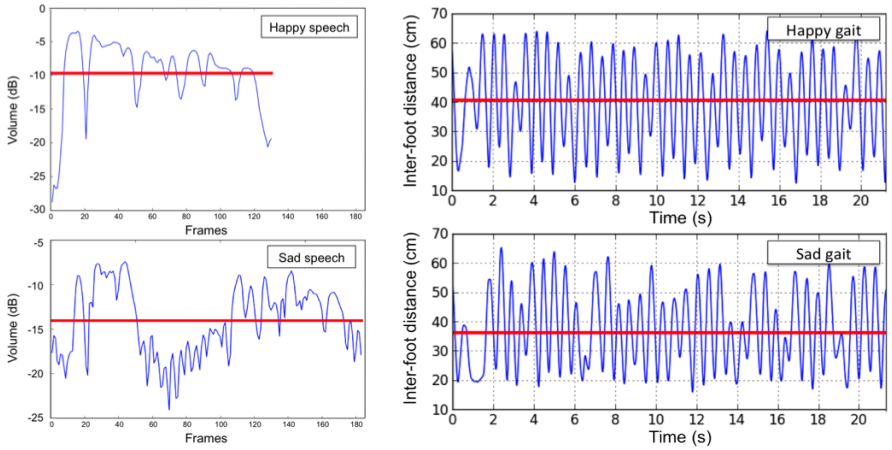



Figure 4.3: An example of volume trajectories of happy and sad speech (left) for the utterance "heute abend, könnte ich es ihm sagen" compared with foot distances of happy and sad gait (right). The red line is the average value, used here as a threshold for speaking and stepping, respectively ( [7], p. 6).

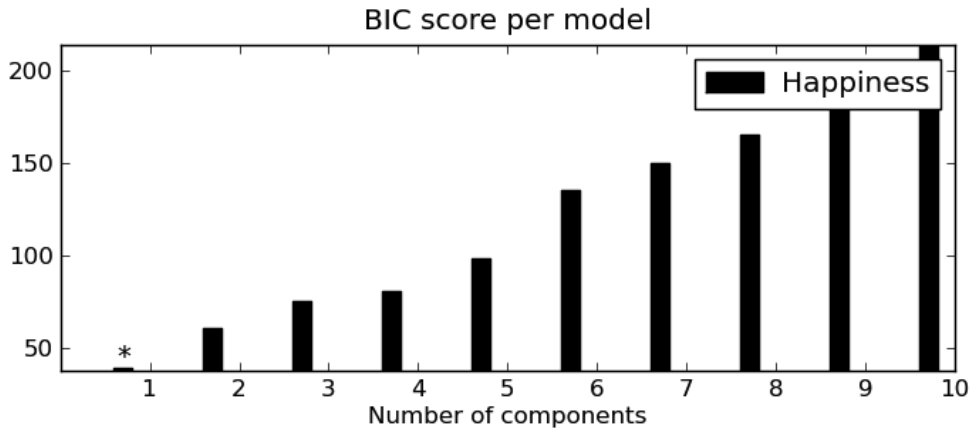



Figure 4.4: An example of the system selecting a 1-component GMM to model the happiness dataset of the Berlin database used in Experiment 1 ( [7], p. 5).

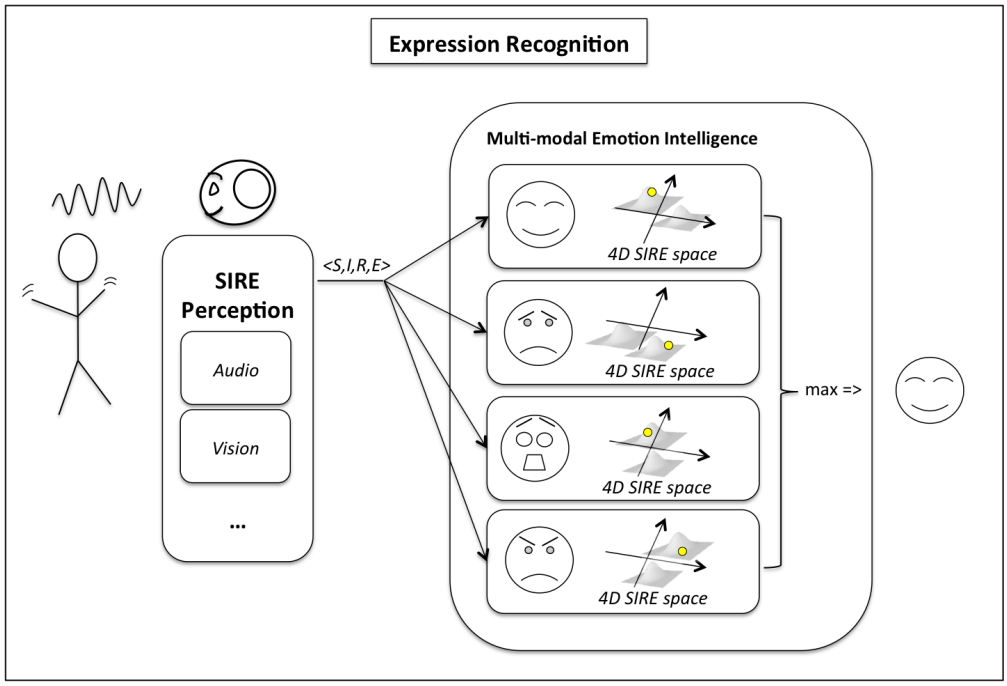



Figure 4.5: Overview of how MEI can perform recognition. The SIRE perception module extracts S,I,R,E parameters through audio or video, and evaluates the SIRE tuple to find the most likely emotion being portrayed. In the present experiment, we use offline data from motion capture, but in previous work a Kinect has been used to perceive emotional motion [5] ( [7], p. 5).

## 4.3 Recognizing emotions with MEI

We now describe how we can use a voice-trained MEI to recognize emotion in a modality different from voice: human gait.

### 4.3.1 Gait feature extraction

Gait studies such as [74] [115] 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 [75] and Janssen [116] considered the force of the steps, and Unuma et al. [117] took into account step-length and hip position. Montepare [118] 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 [119], which contains emotional walking by non-professionals, in neutral, happy, sad, angry, and a few samples of 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), assuming that feet are at their maximum horizontal distance when stepping. These centroids of these peaks determine the time of each step.

*Intensity*. Given our dataset, 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, this corresponds to the "heavy-footedness" of the steps.

*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.

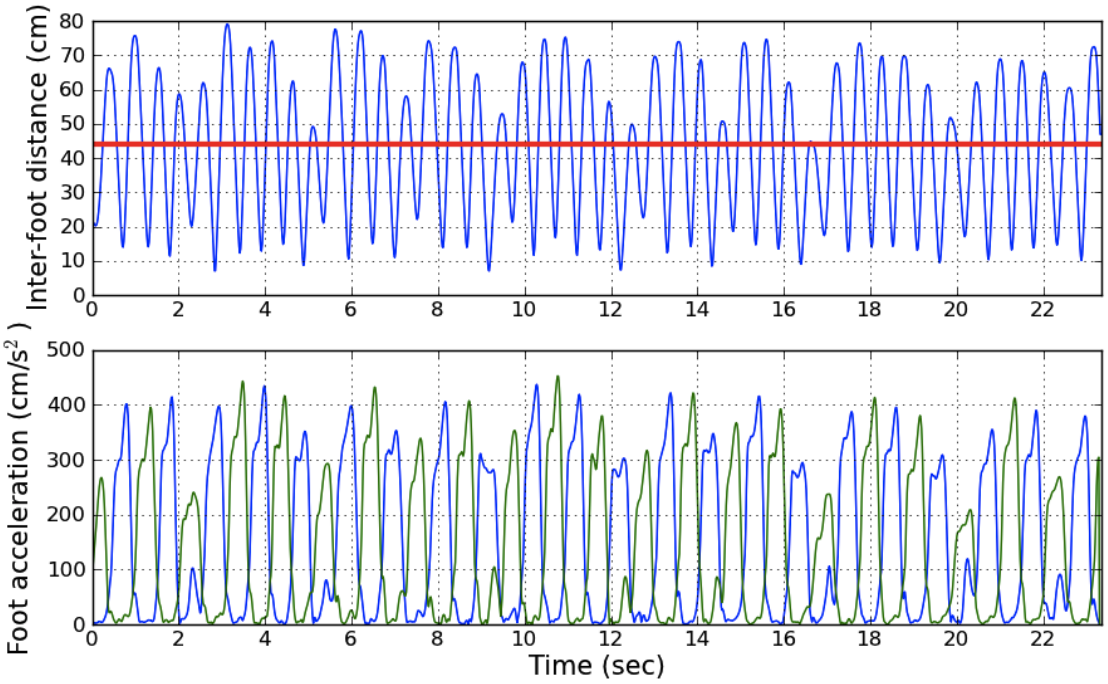
*Extent*. This is defined as the maximum step length in $x, y$ space.

### 4.3.2 Mapping to SIRE space

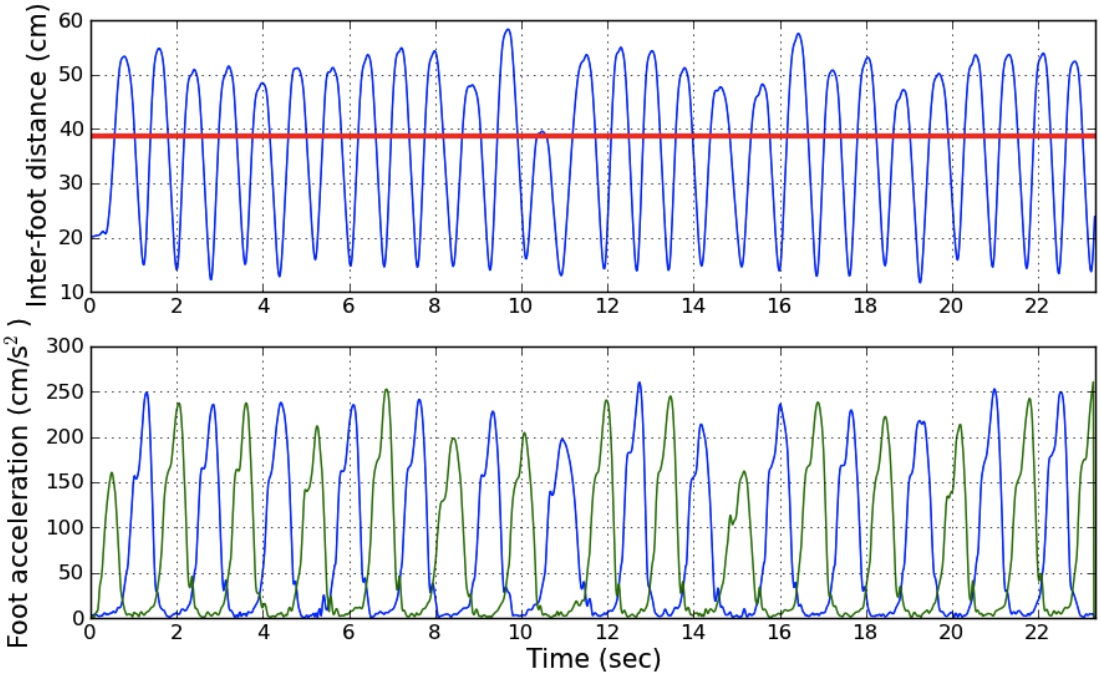
After the features are extracted, the next step of mapping the features to SIRE space is performed identically to the procedure in Section 4.2.2, using the new mean and standard deviations in the gait dataset.

### 4.3.3 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 (Figure 4.5).




(a) Example of angry gait of user "ale"




(b) Example of sad gait of user "ale"

Figure 4.6: Examples of gait analysis. The horizontal line indicates the threshold for peak-picking (mean value). For sad gaits, the step lengths (inter-foot distances) are shorter, and foot acceleration is lower ( [7], p. 7).

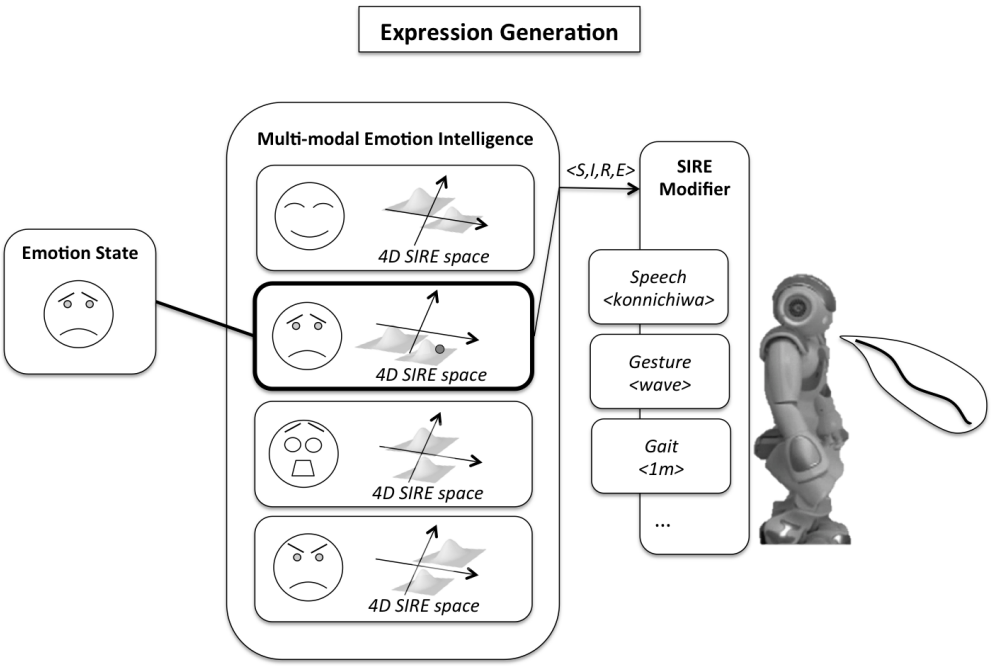



Figure 4.7: Overview of how MEI is used to generate emotionally colored speech and movements on the robot. The desired emotional state is used to select the relevant class model, which is then sampled to generate a SIRE tuple. The tuple is used to modify the speed, intensity, irregularity and extent of existing utterances and movements ( [7], p. 7).

## 4.4 Generating emotional expression using MEI

It is straightforward to generate an emotional expression using MEI (Figure 4.7). We first generate a SIRE tuple for a given emotion, then perform the mapping from the SIRE to the desired modalities. Modulating NAO's gesture from SIRE parameters was explored in Chapter 3. Here, we modulate the robot's speech, gesture and gait (Figure 4.7).

### 4.4.1 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 robots emotion class (happiness, sadness, anger, or fear). How to automatically decide a "current emotional state" is complex and outside the scope of this thesis. For more information, see for example [28] [120] on deriving an emotional state based on cognitive appraisal of one's goals and surroundings.

### 4.4.2 Mapping SIRE to speech

Once we have generated a SIRE sample from our desired emotion class, we map it to our output modalities, such as speech. In this chapter, the 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 [121]. We use the NAO's built-in Japanese TTS to generate an utterance $W$ composed of words $[w_0, ...w_n]$, using Acapela[2] 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.

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

*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.

*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%.

### 4.4.3 Mapping SIRE to gesture

We use the same approach as described in Chapter 3 to control gesture using SIRE parameters. 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 [5].

### 4.4.4 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 0.5cm and 4cm.

*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 3cm and 8cm.

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.

## 4.5 Experiment 1: Cross-modal emotion recognition

### 4.5.1 Purpose

Cross-language emotion recognition has been explored with limited success [110] (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.

### 4.5.2 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 [122] used in Chapter 3. Up to ten different sentences in four styles were used: happy (71 samples), sad (62 samples), angry (127 samples) and fear (69 samples). We used this data to train MEI's four SIRE emotion models as described in Section 4.2.

As test data, we used foot motion capture data from 28 subjects from the Body Movement Library [119]. 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.

Recognition was performed offline, and $p$-values were calculated using a chi-square test with a null hypothesis of 25% (uniform) recognition distributed over the four categories.

### 4.5.3 Results and discussion

How well can MEI recognize emotion in a new context: gait? In Tables 4.2-4.5, 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 4.2 illustrates that cross-modal

| **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 |

Table 4.2: 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%) ( [7], p. 7).

| **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 4.3: 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%) ( [7], p. 8).

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 (Table 4.3). 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 [11] and 75% here (Tables 4.5 & 4.4). This suggests that cross-modal recognition can be achieved by first abstracting data features to a higher-level perceptual space, such as SIRE.

This result is also comparable to human performance. Consider that human emotion recognition in a new context is also low: in [123], participants from 9 countries and 3 continents rated emotional German voice samples over five emotions. The recognition accuracy ranged from a maximum of 74% by native Germans participants, to 52% by

| **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 4.4: 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%) ( [7], p. 8).

| **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 |

Table 4.5: Intra-modal recognition (Eigenwalkers method [11]): Recognition of emotional gait input trained in 20 dimensions (Accuracy: 72%) ( [7], p. 8).

Indonesian participants. There was variability between emotions, too; the Dutch participants rated, for example, German "joy" portrayals with an accuracy of 30%.

Next, we give possible explanations for confusions by analyzing the structure of the voice and gait GMMs. In Figure 4.8, two-component GMMs are plotted for both voice and gait for ease of comparison.

In Table 4.3, 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 sub-types of fear have been shown to differ greatly [8]. Indeed, upon comparing the voice means and gait means (Figure 4.8), it appears that the voice dataset contained fast (terrified) voices, while the gait dataset contained slow, irregular (anxious) walks. According to [122], 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

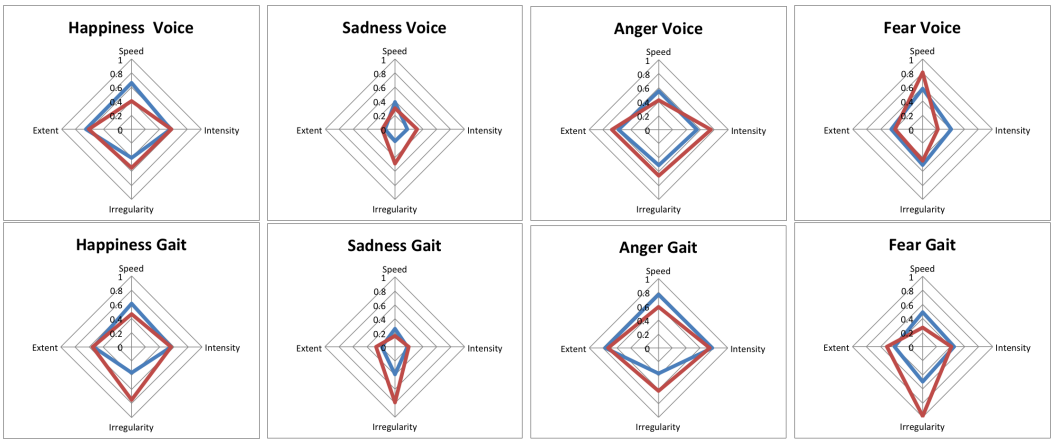



Figure 4.8: Comparison of voice and gait means of GMMs trained with the full voice dataset (>62 samples per emotion) and full gait dataset (>42 samples per emotion). Red and blue lines correspond to the two 4-dimensional components per GMM, which were fixed at 2-components for visualization purposes. 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 [8] ( [7], p. 9).

misclassifications, MEI consistently output high probabilities of both anger and happiness. Why such confusion with happiness? According to an experiment with human evaluators of voice data in [8], "elation was relatively often confused with despair, hot anger, and panic fear, which differ strongly in quality but are similar in intensity". Inspection of Figure 4.8 supports this; we can notice that the dynamics of anger and happiness are relatively similar. How to overcome this confusion is discussed in Chapter 6, for example by including another modality such as face to overcome the difference in valence.

## 4.6 Experiment 2: Cross-modal emotion expression

### 4.6.1 Purpose

In this experiment, we wish to test whether a robot trained with emotional voice can express emotions through speaking, gesturing and walking, as shown in Figure 4.7. As mentioned in Chapter 2, expression itself is a particularly difficult challenge, because the robot a) does not use an expressive face (as in [13] [49]), b) does not use any custom

emotion animations (such as weeping for sadness) [50] and c) does not use hand-defined parameters to control its movement [124]. Importantly, we are also testing whether emotion parameters learned from voice data could be a basis for expression in multiple modalities.

### 4.6.2 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 Chapter 3. Similarly, looking away can also express embarrassment or social disinterest, according to gaze studies [125]. 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 [6]) and even a one second latency could imply negative hesitation.

Based on our previous experiments, there is a plethora of cues to consider, so now attempt to control for some of 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 many studies test emotional expression in a independent (i.e., solitary) context, humans use many cues, including social context, to decide the emotion of a person. For instance, [126] showed that point-light displays of love and joy were understood in dyads but not the single person condition.

For these reasons, and also due to the fact that the NAO platform is small and child-like, we design our experiment to evaluate MEI 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 (Figure 4.9). Four interactions were created:

1. The human said "Konnichiwa" (Hello) while waving at the robot.

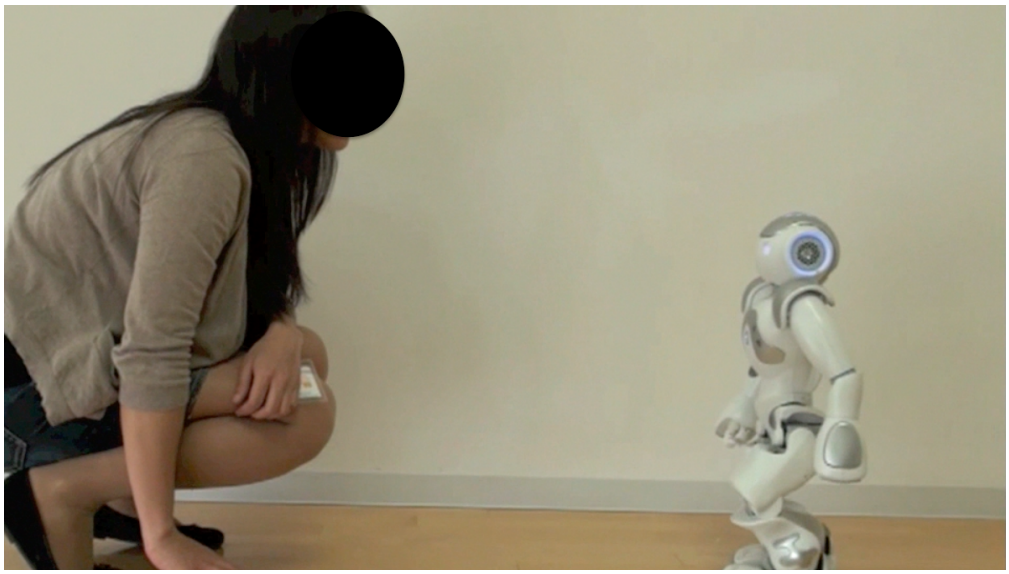


Figure 4.9: Stimulus used in Experiment 2 of robot interacting with human with various emotions. The robot spoke, gestured, then walked toward the human in all stimuli ( [7], p. 9).

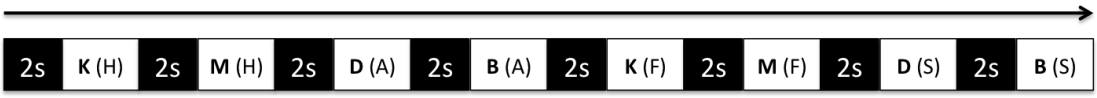



Figure 4.10: Order of presented stimuli for all subjects. The letter in bold corresponds to the interaction utterances: K–Konnichiwa, M–Mite, D–Dame, B–Baibai. The letter in parentheses is the robot's emotional SIRE modification: H–Happiness, S–Sadness, A–Anger, F–Fear ( [7], p. 9).

| **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 4.6: Interactions between human and robot, and SIRE modifications used in Experiment 2. JP=Japanese language ( [7], p. 10).

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 4.6, then walked toward the human. The gesture contained 2 movements used in Chapter 3, starting with the robot's hands close together and 1) the hands moved apart to either side, and 2) one hand moved upwards and the other moved downwards. As shown in Table 4.6, for each interaction, the robot's speech, gesture and gait were subject to one of two emotional modifications, depending on the stimuli (Figure 4.10). 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 Figure 4.10, separated by 2 second black frames, for a total of 2min 10s. 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 of sadness and fear, portraying a "reserved" robot. We chose to use these logical progressions because pilot trials with a random order

| P | A | D | Emotion terms from [9] | 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 |

Table 4.7: Our expected PAD values for happiness, sadness, anger and fear portrayals in Experiment 2, based on emotion terms provided in [9] ( [7], p. 10).

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 SAM (Self-Assessment Manikin) Measurement Scale for Japanese called REM [127] to rate the pleasure, arousal and dominance (PAD) of the robot in each scene [9].

In Table 4.7, we show the expected positive/negative PAD values for the four emotional classes used in our study. We used the table from [9], 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 that 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. 4.10.

3. The participant watched the video again, this time while choosing how they believed the robot felt during the scene, on each of the PAD scales. The participant was given as much time as desired after each scene before proceeding to the next.

### 4.6.3 Results and discussion

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

According to the ratings shown in Figure 4.11, happiness and sadness were well expressed. We find that the portrayals of happiness had +P, +A, +D, (.53, .44, .28) and (.48, .46, .24). Both portrayals of sadness also were shown to have -P, -A, -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 -P, +A, -D, was not well captured in our scenarios. The assessments as +P, +A, +/-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 [128]. 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 -P,+A,+D, but participants rated the expressions as -P,+A,-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 (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 D = -0.15, compared to sadness, which was more submissive at D = -0.3. It may be interesting to check the effect of robot size relative to the human; with a robot that was equal or larger in size to a human, this dominance dimension may possibly be pushed to +D, making the robot look angry using our technique.

There are limitations to these results. Firstly, Experiment 2 cannot ascertain how much effect 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 [129].

## 4.7 Summary

In this chapter, we used the SIRE paradigm from Chapter 3 as a basis for training a multimodal emotional intelligence (MEI). This MEI contained SIRE GMMs for each of happiness, sadness, anger and fear emotion classes to *recognize*, *represent*, and *express* each of them. We showed MEI's cross-modal generalization ability: emotional gait recognition with a voice-trained model gave results almost as good as training with gait


Figure 4.11: Results of user evaluations, where P=pleasure, A=arousal, D=dominance. Happy and sad emotional expressions conform to expected values PAD values from [9]. 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. ( [7], p. 12).

samples. MEI could also reliably express robot happiness and sadness in a multimodal way. In the next chapter, we perform training in a realistic human-robot interaction loop. Whereas we have been using offline, annotated emotion databases until now, we will face the question: how are humans able to develop multimodal emotional intelligence without these explicit tags, and how do we associate SIRE dynamics with feelings?

# 5

# Infant-inspired Emotional Development

*"Some of the most revolutionary ideas in brain science are coming from cribs and nurseries."*

– Patricia Kuhl

## 5.1 Introduction

Studies in infant development can provide a good clue on how to develop robots with an adaptive emotional system. Primary emotion capabilities are often regarded as "innate" [130], yet–like the acquisition of language–it is likely that the mechanism is more subtle.

In this chapter, we describe an emotional robot system that can be trained with human, caregiver-like interaction. First, we will first examine the universal phenomenon of infant-directed speech, or "motherese", as a tool for the development of primary emotions in a robot. Next, we describe an interactive motherese robot system where the emotional dynamics in voice (and possibly movement and music, based on results from Chapter 3 and 4) are learned through associative learning with a robot's internal physical state. Finally, we show the results of an experiment with the system, with naive participants as caregivers. We show that the motherese-trained models show basic capabilities of recognizing adult emotional expressions.

## 5.2 The case for emotion and motherese

In Chapter 2, we noticed that humans undergo a rapid growth in emotional intelligence between the ages of 0 and 1. Let us consider a possible mechanism for this growth [45], 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 [131]. The emotional speech of motherese often co-occurs with exaggerated emotional facial expressions [40]. 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., [132] [42]). Although most studies of ID speech concentrate on language acquisition (e.g., [133]), ID speech and its role on the comprehension of prosody has received little attention [134] [45]. Soken and Pick [40] suggest an important role played by motherese for learning about affective events:

> It has been shown that infants are attracted by and attend to motherese, which is characterized by more exaggerated intonation and higher pick 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

[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

> movement). [...] Child-centered displays may serve as opportunities for learning about affective events.

Lewis [135] 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 [134] 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 'Dont touch that!' are spoken with low pitch and high intensity."

## 5.3 A robot that develops emotions through interaction

We have covered some of the emotional characteristics of motherese (ID speech) itself, but what are the characteristics of a motherese-like interaction? According to Gleason [136], it is not a one-way communication – it requires two active participants: when a caregiver speaks to an infant, the infant's *reactions* shape the interaction. In fact, Fernald has shown that mothers cannot reliably produce motherese it in front of a microphone [137]. *Turn-taking* is also observed early vocalizations between a mother and infant [138], which has been described as "mutual entrainment between mothers an infants during early social interactions" [139]. According to some studies, this may even involve correlations of melody types [45]. Inspired by these findings from infant psychology, we design a robot system that allows a back-and-forth interaction, where the robot takes the place of the infant. We use the SIRE paradigm of Chapter 3 as a basis for coordinating the "melody types" between the caregiver and the robot, and design an emotional feedback loop to allow human-robot entrainment.

### 5.3.1 Design of an emotional human-robot feedback loop

How do we design an emotional feedback loop? Let us consider three possible human-robot interaction configurations, as shown in Figure 5.1. The first possibility is a pure imitation scheme, where the robot mimics the dynamics of the motherese utterance. In

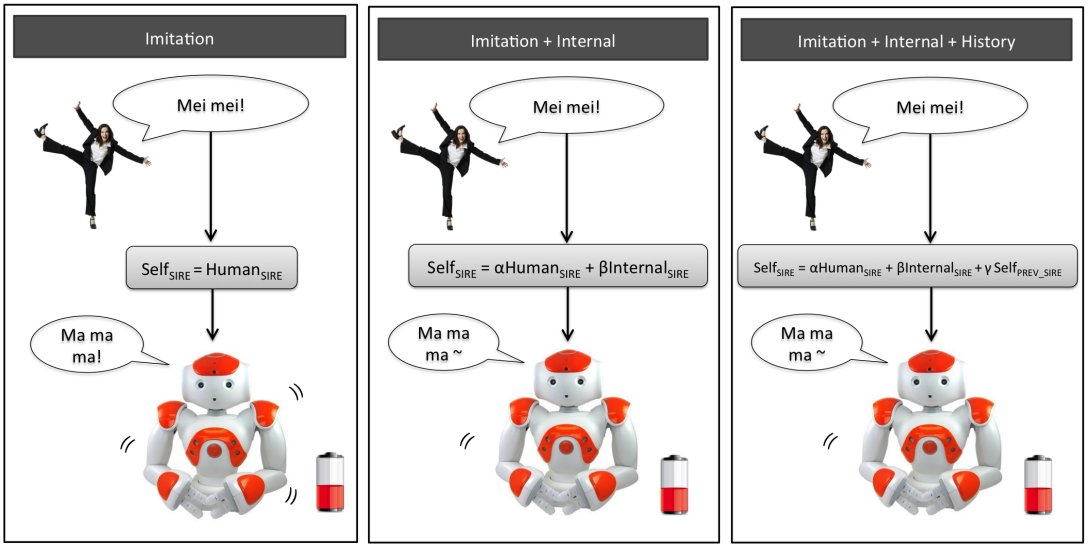



Figure 5.1: Possible human-robot interaction configurations for an emotional feedback loop. In this chapter, the right-most scheme is used. (Left) An imitation scheme: The robot simply extracts SIRE parameters from the human and reproduces them in gesture and speech, similar to the transfer system in Chapter 3. (Middle) The robot expresses a combination of observed $Human_{SIRE}$ and $Internal_{SIRE}$, a SIRE it associates with its current internal state. (Right) Similar to Imitation + Internal scheme, but effects are dampened through time.

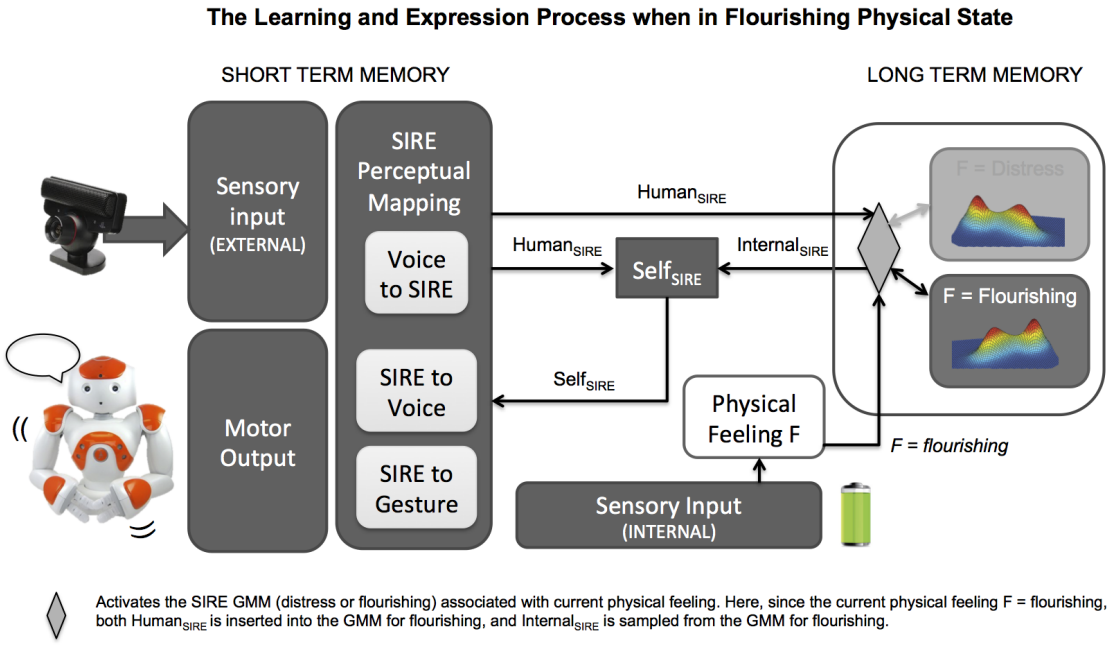



Figure 5.2: An overview of the system when the robot is in a flourishing state

this scheme, the robot extracts SIRE parameters from the human's vocalizations, and reproduces them in gesture and speech, similar to the transfer system in Chapter 3. In some ways, this could be considered as simple mimicry–a complete mirroring of the observed dynamics. In the second configuration, the robot expresses a combination of a) what it sees and b) its own internal emotional state. Specifically, it is a linear combination of observed $Human_{SIRE}$ and $Internal_{SIRE}$, the latter of which is a SIRE 4-tuplet it has learned to associate with its current internal state (described shortly in Section 5.3.3). A third option is similar to the second, but the effects are dampened through time by deviating from the previous SIRE expression. In this chapter, we implement the third schema, but we must consider how the robot's internal state might be defined.

### 5.3.2 Robot physical feeling

At birth, the infant is equipped with the most innate of emotional expressions: crying. Indeed, at this point in development, the infant is in one of two physical states: homeostasis, or not (e.g., extreme heat or cold, empty stomach). This in-built distress signal

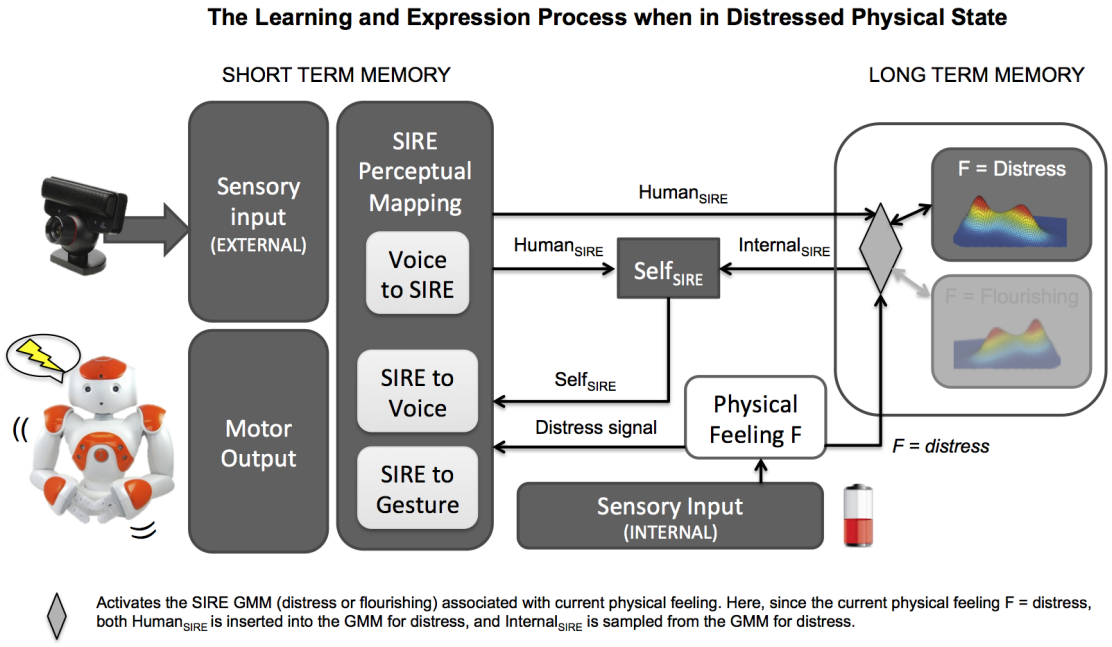



Figure 5.3: An overview of the system when the robot is in a distress state

of crying alerts the caregiver to a lack of homeostasis.

Inspired by newborns, we define a robot's most basic level of physical "feeling" based on these two states. Ortony calls the most innate emotional level the *reactive level* which "assigns along two output dimensions, one of which we call "positive" and the other "negative"" [3]. Similarly, Damasio defines feelings as "the expression of human flourishing or human distress, as they occur in mind and body." [140]. Therefore, we define two "innate" physical feeling states for our robot, which we will call *flourishing* and *distress*. In this chapter, these two states are represented symbolically, but in future work they should be tied to a robot's physical state, for instance *flourishing* corresponding to full battery and CPU/motor temperatures within working limits, and *distress* corresponding to a near-empty battery and/or hot motors.

This physical feeling $F = (flourishing, distress)$ has 2 important functions. The first is to cause a distress signal to be emitted when the robot is in *distress* state. The second is to serve as a switch (cf. the diamond symbol in Figure 5.2) for two sub-functions: a) storing information into long-term memory based on the value of $F$ and b) retrieving information from long-term memory based on the value of $F$. For example, if the robot is in a *flourishing* state when a caregiver is smiling and speaking to it in a

happy voice, the system will store this "happy voice" information in the $F = flourishing$ state (Figure 5.2). If the robot is in a *distress* state and the caregiver tries to comfort it with a soothing tone, the robot will store this "comforting" vocal information into the $F = distress$ state (Figure 5.3).

### 5.3.3 Using SIRE GMM as emotional long-term memory

How is information stored into long-term memory? Here, we use our SIRE GMM scheme from Chapter 4 as a model for flourishing and distress. This scheme is similar to how we used four GMMs to learn the SIRE distributions for happiness, sadness, anger and fear, but in this case, we will only use two GMMs (flourishing and distress).

**Learning** To illustrate, let us assume that the robot is in a flourishing (i.e., full-battery) state. A human, following the intuitive parenting paradigm described earlier, may begin to speak to the robot in a happy way. In this case, since the robot's physical feeling $F = flourishing$, the human's observed SIRE values $Human_{SIRE}$ will be added to the training data for the *flourishing* SIRE GMM.

**Expressing** The robot's expressed gesture and vocal dynamics $Self_{SIRE}$ depend both on the human's SIRE dynamics and the robot's $Internal_{SIRE}$. The value of $Internal_{SIRE}$ is produced by sampling from the SIRE GMM corresponding to the current physical feeling $F$. For example, when the robot's physical feeling $F = distress$ (i.e., low battery), we sample from the *distress* SIRE GMM.

The voice and movements of the robot are thus modified using $Self_{SIRE}$, a vector of four values on $[0,1]$, where

$$Self_{SIRE} = \alpha Human_{SIRE} + \beta Internal_{SIRE}. \tag{5.1}$$

and

$$\alpha + \beta = 1, 0 \leq \alpha, \beta \leq 1 \tag{5.2}$$

**Empathy: The ratio of imitation and internal state**. How do we decide the values of $\alpha$ and $\beta$? Consider that if $\alpha = 1$, the robot shows pure mimicry. If $\beta = 1$, then the human in front of the robot has no immediate effect on the robot's expressions, and

the robot simply expresses based on its internal physical feeling. These values can be considered a kind of empathy setting for the robot. For example, if the human is expressing sadness, then a robot with $\alpha = 1$ would immediately portray similar sad vocal and gestural dynamics (high empathy). On the other hand, a robot with $\beta = 1$ and a full-battery state would simply convey what it has learned to associate with it's own flourishing physical feeling, ignoring the sad expressions of the human (low empathy). In our experiments, we generally set $\alpha$ and $\beta$ to be equal, but it would be interesting in future work to test the impression of the robot when changing these parameters.

**Entrainment** Entrainment is a term used to designate synchronizing with and adapting to the interaction partner (p. 134, [141]). As shown in Figure 5.1, we deduce our current SIRE state based on the previous emotional state, to designate a temporal entrainment between the caregiver and the robot.

$$Self_{SIRE} = \alpha Human_{SIRE} + \beta Internal_{SIRE} + \gamma Self_{PREV\_SIRE}. \tag{5.3}$$

and

$$\alpha + \beta + \gamma = 1, 0 \leq \alpha, \beta, \gamma \leq 1 \tag{5.4}$$

In our experiments, we set $\alpha = \beta = \gamma = 1/3$, and future work should test other configurations for these parameters.

### 5.3.4 Real-time implementation of SIRE audio processing

We implemented the system described in Figures 5.2 and 5.3 using the Aldebaran NAO robot and HARK (HRI-JP Audition for Robots with Kyoto University)[2] real-time robot audition system. A Playstation Eye was used as a microphone input, to avoid mixture with the robot's speaker output. The speech recognition system Julius, trained with an English acoustic model, was used to detect the words spoken by the user, in order to count the number of syllables. The mappings for SIRE voice and gesture, and SIRE GMM learning mechanism remained the same as defined in Chapter 4.

[2] http://www.hark.jp

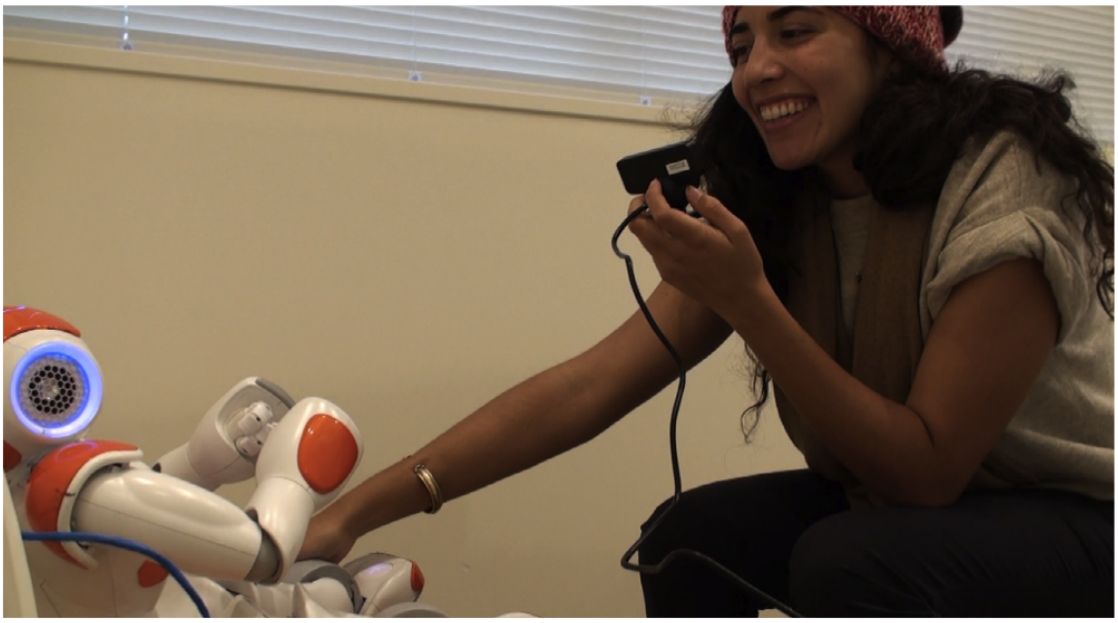


Figure 5.4: A participant interacts with the robot by speaking into a microphone.

## 5.4 Experiment 1: Online collection of motherese

### 5.4.1 Purpose

The goal of this experiment was to test the system in an online manner with naive experiments, and visualize the SIRE dynamics of their utterances.

### 5.4.2 Materials and procedure

We used the robot motherese system to collect infant-directed samples of *praise*, *comfort*, *prohibition* and *attention* from human participants, four categories of motherese as defined in [134].

We recruited 6 fluent English speakers from Western countries (3 female, 3 male, mean age=29.8 years, std=3.9): 3 from USA, 1 Australia, 1 Madagascar, 1 Chile. The participants were first introduced to the robot through a written introduction, and told that the robot's name was "Mei Mei". The robot was introduced as "young and continuously learning", and that the participants were the robot's caregiver for the duration of the experiment. The participants were also told that the robot, because it is young, could not understand the content of their words, only the way in which they say it.

The participants were then instructed to interact with the robot in four different situations, by speaking into the microphone and saying the robot's name, as in Figure 5.4.

1. **Attention**: Get Mei Mei's attention by saying her name.

2. **Prohibition**: Mei Mei is crying. You will try two different ways to stop her from crying. First, you will prohibit her from crying by saying "Mei Mei".

3. **Comfort**: Mei Mei is crying again. Your goal is to soothe and comfort her by saying "Mei Mei".

4. **Praise**: Your goal is to praise Mei Mei because she is no longer crying, and make her feel that she is loved.

For the purposes of the experiment, the physical feeling F was set manually to correspond to the situation. In Situation 1 and 4, $F = flourishing$, and in Situation 2 and 3, $F = distress$. In all four situations, the robot gestured to convey its affect using $Self_{SIRE}$. In the 2nd and 3rd situation, to simulate a distress signal, the robot also vocalized. The "cry" was produced by repetition of the syllables *"ma ma ma"*. This vocalization was also subject to $Self_{SIRE}$. In the distress situations (2 and 3), $Self_{SIRE}$ was initially to [0.9, 0.9, 0.9, 0.9]. In the flourishing situations (1 and 4), $Self_{SIRE}$ was initialized to [0.1, 0.1, 0.1, 0.1]. The interactions, of course, modified the robot's internal state continuously, based on Equations 5.3 and 5.4. All initial settings were identical for each participant.

The interactions were recorded with a standard Sony Exmor R video camera, as well as through the robot's own front-facing camera.

### 5.4.3 Results and discussion

The interactions resulted in 510 motherese utterances in total (128 praise, 114 comfort, 123 prohibition, 145 attention). We plot the means of the resulting GMMs in Figure 5.5, and show example captures from the robot's camera during vocalizations in each condition in Figure 5.6. The variation across motherese types and facial expressions gives qualitative weight to our hypothesis that our robot system can elicit motherese and facial

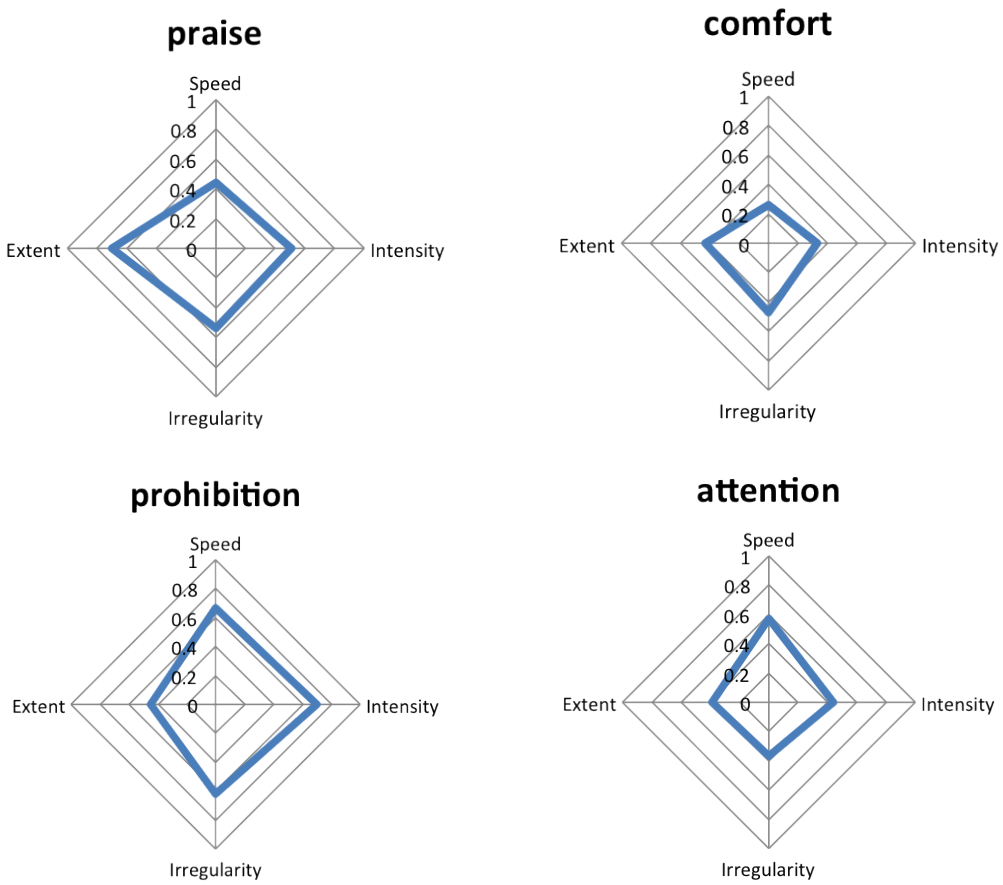



Figure 5.5: Plotting the SIRE means of 1-mixture GMMs trained in each condition.

Figure 5.6: Visual input accompanying the different kinds of “Mei Mei” vocalizations: praise (top left), comfort (top right), prohibition (bottom left), attention (bottom right). Images captured from robot’s camera during each condition, mid-utterance.

expression that are differentiable across interaction types (praise, comfort, prohibition and attention).

## 5.5 Experiment 2: Training with motherese, testing with emotional voice

### 5.5.1 Purpose

The goal of this experiment was to test if motherese could be an adequate training mechanism for learning emotion dynamics. We do this by performing two analyses.

First, we test whether the robot could associate *happy* dynamics with a physical *flourishing* state, and *sad* dynamics with a physical *distress* state. Indeed, if a robot listens to a sad voice and associates it with its own experience of distress, this could give evidence that our model provides a means for robot empathy. When the participants praised and comforted the robot, they spoke to the robot while it was in a physical state of flourishing and distress, respectively. We therefore train two SIRE GMM models on the caregiver input during the *praise* and *comfort* interactions and check how they respond to happy and sad voices.

Secondly, we hypothesize that all the utterances in motherese (praise, comfort, prohibition and attention) have correlates with adult-directed expressions of emotions. If that is true, then our robot could learn to associate emotional voices to situations in which it heard similar motherese utterances.

### 5.5.2 Materials and procedure

The utterances captured in Experiment 1 were used to train two different MEI (see Chapter 4).

- The first MEI contained two GMMs, one for each of the conditions *flourishing* and *distress*. The flourishing GMM was trained with samples collected during the praise condition, and the distress GMM was trained with samples from the comfort condition.

| **Detected** Input | **Flourishing** # associations | **Distress** # associations |
|---|---|---|
| Happiness | **64** (90%) | 7 (10%) |
| Sadness | 10 (16%) | **52** (84%) |

Table 5.1: Emotional voice association rates on a model trained on comfort and praise motherese

- The second MEI contained four GMMs, one for each of the conditions praise, comfort, prohibition and attention.

Next, the emotional voice samples from EmoDB (see Chapter 4 for details) were tested against these MEI, where the GMM which output the highest probability was selected as the best match. *P*-values were calculated using a chi-square test with a null hypothesis of 25% (uniform) recognition distributed over the four categories.

### 5.5.3 Results and discussion

**Association of happy voices with flourishing, and sad voices with distress**

In Table 5.1, we can see that happy voices from the German database were associated with the flourishing state 90% of the time. Sad voices were associated with the distress state 84% of the time. This suggests that a robot could develop a physically grounded association in response to happy and sad voices, by exposing the robot to comforting and praise motherese when it is in low-battery or high-battery states, respectively. In fact, this limited exposure to praise and comfort mimics the sequence of motherese directed at very young infants: at 3 months, infants prefer and receive more comfort vocalizations. Then, at 6 months, they prefer and receive praise vocalizations [142].

**Associations of happy, sad, angry and scared emotional voices**

What happens when a robot has been exposed not only to two types of motherese vocalizations, but four? Table 5.2 shows the output of the robot's recognition system which was trained with motherese and analyses adult-directed emotional voice. Our hypothesis is that a) happy voices are associated most strongly with praise b) sad voices associated with comfort c) anger voices associated most strongly with prohibition. As

| **Detected** Input | **Praise** (%) | **Comfort** (%) | **Prohibition** (%) | **Attention** (%) | *p-value* |
|---|---|---|---|---|---|
| Happiness | **54** | 10 | 24 | 13 | <0.0001 |
| Sadness | 0 | **65** | 2 | 34 | <0.0001 |
| Anger | 47 | 8 | **38** | 7 | <0.0001 |
| Fear | 12 | 13 | 13 | **62** | <0.0001 |

Table 5.2: Emotional voice association rates on a motherese-trained model

a preliminary hypothesis, it is not clear that fear (a negative emotion) would correlate with attention voices, because "attention bids" were described as a positive, playful interaction for caregivers and infants in [39].

In Table 5.2, we first notice that recognition rates of happiness and sadness drop, but that sadness is still associated with comfort at 65%, and happiness associated with praise at 54%, both at levels significantly higher than chance (Table 5.2). However, happiness is sometimes associated with the prohibition condition, at 24%. Anger is not well associated, being confused with happiness. Surprisingly, fear voices are associated with attention motherese, to a high degree.

It is interesting to note that infants prefer motherese vocalizations in a preset order: at 3 months, they prefer comfort vocalizations. At 6 months, they prefer approval (praise) vocalizations. Lastly, at 9 months, they prefer directive vocalizations [142]. The association rates of first comfort, then praise, and lastly prohibition parallel this infant progression.

In Figure 5.7, we can visually inspect the relationship between emotional voice and motherese-trained models in SIRE space. Here, the GMM means for the 1-mixture models for both motherese and emotional voice are plotted for comparison purposes. In Table 5.3, we show the same data in more detail, by calculating the Euclidean distance between the GMM means. We can see that happiness and praise reside very close to each other, as do fear and attention.

We can conclude that a robot that interacts in praise, comfort and attention conditions can be primed to associate happy, sad and fear voices with very relevant correlates (e.g., a fear voice means that the human is trying to attract their attention to something).

Although we did not predict that the attention motherese condition would have any

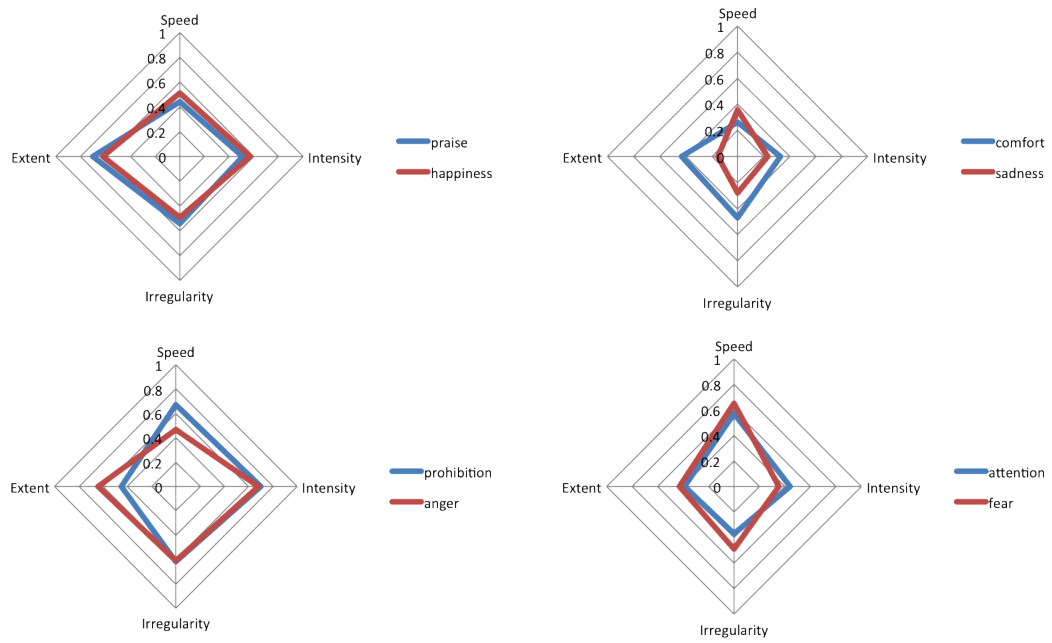



Figure 5.7: Plotting the SIRE means of 1-mixture GMMs trained in each condition.

correlate in emotional voices, the fear and attention SIRE profiles appeared to be very similar (Fig. 5.7 and Table 5.3). This could make sense, since during fear vocalizations by a caregiver, it would be necessary to attract the infants' attention of the danger.

Together, the results suggest a "Primary Emotion Recognition" utility: a MEI trained in comfort, praise, prohibition and attention conditions can provide output as a basis for detecting happiness, sadness and fear, while serving as a scaffold for the detection of anger. Since primary emotions are not supposed to involve any rational appraisal processes, it may be natural that happiness and anger are easily confused; a robot's

| **Motherese** Emotional Voice | **Praise** | **Comfort** | **Prohibition** | **Attention** |
|---|---|---|---|---|
| Happiness | **.13** | .67 | .20 | .39 |
| Sadness | .39 | **.36** | .47 | .39 |
| Anger | .30 | .72 | **.28** | .37 |
| Fear | .30 | .39 | .43 | **.18** |

Table 5.3: Euclidean distance between SIRE means of 1-mixture GMMs for motherese and emotional voice classes. Lower values indicate that the two classes are more similar. Distances in bold show the closest motherese profile for a given emotion in voice.

situation must be enriched contextual information such as goal-state or current action.

## 5.6 Summary

In this chapter, we proposed a way that the multimodal emotional intelligence of Chapter 4 could be trained through human-like interaction, by associating SIRE expressions of emotion with low-level physical feelings. We found that motherese interactions corresponding to praise and comfort trained a robot to later recognize happiness and sadness in novel voices. Further interactions in motherese prohibition and attention conditions also appear to provide a primary emotion intelligence for happiness, sadness, and fear, as well as a basic scaffold for anger that could be later refined in secondary appraisal.

# 6

# Discussion

*"Any emotion, if it is sincere, is involuntary."*

– Mark Twain

In this chapter, we will summarize some of the results and interesting conclusions of the research presented thus far, and their implications.

## 6.1 Observations

### 6.1.1 On anger and happiness

We observed that anger and happiness were mistaken for one another in many of our experiments. For example, when the MEI module of Chapter 4 was presented with an angry gait, it output high scores for both happiness and anger. This suggests the need for an additional piece of information such as high-level appraisal. In other words, robot emotion systems should not aim to distinguish the "basic" emotions at the same level, but rather work in a stratified manner. According to Ortony in Chapter 2, we must remember that happiness, excitement, fear and sadness do not require cognitive appraisal, but anger does.

Interestingly, we can find anger-happiness confusions in infant development, too. Charlotte Buhler in 1927 observed the emotional recognition skills of infants less than one year old [143]. She found that at 8 or 9 months, "children sometimes misinterpret an angry facial expression as a joke or a bit of kidding – perhaps because they are unable to imagine any ground for blame on their side". Perhaps context, for infants too as

with robots, is needed to clarify the emotional expression. A robot that was stopped in mid-action (i.e., could indeed imagine ground for blame), for example, may have more reason to distinguish an expression as angry.

### 6.1.2 More cues: embodied factors of emotion

When a robot is mobile, it can express confusing emotional cues for fear (e.g. Chapter 4, Experiment 2). In one experiment, the robot expressed fast "terror" fear towards an object held by the human. This was incongruent, as terror fear should be directed away from the object of fear. One possible explanation for this is that in addition to SIRE dynamics, a fifth parameter *direction* could be necessary. Another interpretation is that cognitive appraisals are needed, with a defined object necessary in the expression of fear.

We also found in Chapter 3 that an immobile robot was perceived as more aggressive than a robot that gestured with angry dynamics. One possible suggestion for this is that staring can be perceived as a sign of threat, even among animals [98]. This could also have a cultural explanation: staring is considered impolite or aggressive in Japan (where the experiment took place), whereas among Americans it is considered a sign of honesty to make continued eye contact with the person [144]. This reminds us that cultural cues in expression of emotion are important; development of emotion in a culture-specific interaction loop may be a natural way to accommodate for these particularities.

## 6.2 General discussion and remaining work

### 6.2.1 The primacy of multimodal emotions

What makes the multimodal emotions explored here so special? What drives the common link between music, voice and motion? I suggest that it is what Picard called the "bodily" component of primary emotions that we addressed in this thesis. For example, when a heart beats faster in fear, arms and legs can move faster. The primary emotions studied here may be linked to the bodily, physiological states that induce them.

Music is an ideal lens by which to study emotions at the bodily level. After all, without any cognition or context, we can feel and express emotion through music. And—

whether playing a drum, violin or piano— the music is the result of body movements. Simple movements and sounds are the forms of communication in infants less than one year old, that do not have higher forms of communication (e.g., language.). For these reasons, the bodily component appears to be a fundamental basis for primary emotions.

### 6.2.2 What is a robot emotion?

In previous systems, a robot emotion usually took one of two forms: a) a state label such as happy or sad, or b) a feeling point in three-dimensional such as PAD [13].

A future direction for work could be to consider emotions as a conjunction of simultaneously experienced embodied features. For example, a primary emotion could be the combination of a robot's bodily expression (dynamics through SIRE, e.g. [0.1, 0.3, 0.2, 0.5] and facial combination, e.g., upturned eyebrows) + its internal physical state (e.g., distress). The addition of cognitive components, such as action tendencies, could additionally distinguish between the secondary emotion classes typically known as happy and angry. This is shown in Fig. 6.1 as components C. and D., and is further described in the following section on secondary emotions.

### 6.2.3 Extension to secondary emotions

An important direction for future work is to integrate the input from a robot's primary emotion system into a secondary emotion system. For example, when a MEI gives high scores for happiness and anger, a robot should make a further distinction between happiness and anger by examining its current action tendencies. This corresponds to the multilevel processing shown in Scherer's 2009 appraisal theory [10] shown in Figure 6.1. In the diagram, our SIRE MEI would correspond to (B.), and together with goals and action tendencies (A.), a final set of features would be received in a central representation (C.). This central representation can be given a natural-language label such as happy, fearful, etc (D.).

Infant development could explain how these components (action tendencies, motor expression, etc.) are associated together as emotions in the first place. 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

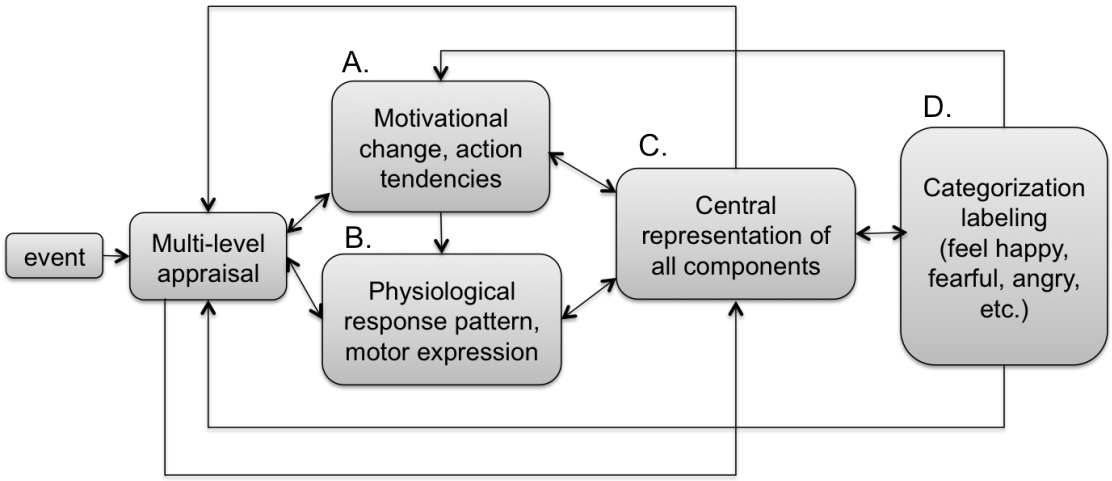



Figure 6.1: Appraisal Model proposed by Scherer [10] and reproduced here.

whose goal has been thwarted may express that which he has associated with such a situation: an angry face and an angry voice. Indeed, anger is often defined by theorists as "negatively-valenced affect that arises from the blockage of movement toward a desired goal" [145]. Or, consider an infant that is hungry and crying: they will be comforted by the empathetic parent with a sad face and sad voice. The infant may learn to associate these quiet, low intensity sounds with both a negative internal state and the positive interactions of their beloved parent. In fact, it has been reported by Kawakami et al. that sad music stirs up both sad feelings and romantic feelings [146]. It is possible that naturalistic interactions with a robot could also result in deeper associations of context, motor expression, and feelings.

### 6.2.4 Moving past the four emotion categories

In this thesis, we focused on four emotion categories: happiness, sadness, anger and fear. We concentrated on these emotions for two reasons. First, they are part of those considered "six basic emotions" by Ekman (happiness, sadness, anger, fear, surprise and disgust) [20]. We excluded surprise and disgust because they are not readily studied in music literature. Secondly is the availability of emotion data. Although dimensional emotion data exists (e.g., voice samples annotated in pleasure-arousal dimensions), data for emotional music, voice, motion and gait were consistently classified in the four categories above. However, we do not claim that these four categories are an ideal

representation for emotions.

For instance, what about subcategories of these four emotions, such as elation or anxious fear? In future work, one possibility is to create emotion classes automatically, depending on the desired granularity. For example, from the two happiness and distress GMMs, an algorithm could be constructed to automatically split the model when the complexity becomes too high (e.g., based on BIC score) or when other co-occuring features (such as a toothy smile for elation) are added so as to make one component clearly separable.

Finally, the neutral expression was not considered here. One possibility is to consider, for example during recognition, that neutrality is equivalent to a low score for all of happiness, sadness, anger and fear. In future work, we could search for an appropriate threshold.

### 6.2.5 The advantages of an Occam's razor approach

The concise, low-dimensional SIRE emotion representation is unique for an emotion system, and we found that it provided at least two advantages. First, it let us have an intrinsic understanding of the models that were created via machine learning. This is not a common case when using GMMs, because high-dimensional feature vectors and anywhere from 8 to 32 or more components are typically used. Factor analysis using techniques like Principal Component Analysis (PCA) is then required as an additional step for visualization of data, but the visualization is no longer representative of the trained model. In our experiments, we could easily plot the 4 SIRE parameters and 1 to 3 components of the trained model. This allowed us, for example to discover that our voice model contained mostly terror fear, and the gait model contained anxious fear (Chapter 4). The second advantage is that it provides a practical approach to emotion synthesis, described in the next section.

### 6.2.6 How to express emotion on an arbitrary robot

This thesis provided a straightforward technique to create emotional expressions, e.g., for gait, gesture, and voice. We used the following two-step process: 1) define the speed, intensity, irregularity and extent mappings and 2) apply the learned SIRE parameters in

Figure 5.7. We can compare this to other emotion synthesis systems which use a more involved procedure. For instance, for emotional gait, complex methods involving motion capture and PCA on entire body poses are common approaches for humanoids, but it could be more difficult to extend this work to emotional gait on a hexapod robot, for example. Our approach can be a practical technique that many roboticists and virtual agent creators can employ. It should be noted, though, that a different, optimal representation may exist (although we tested SIRE and SIE, we did not test other combinations, such as SIR, etc.)

### 6.2.7 Integration of face and touch

This thesis disregarded a major source of affective information in humans; the face. In future work, we suggest two ways to integrate the face. The first might be to, along with SIRE, store *static* facial feature configuration in long-term memory. In this case, for example, a robot would associate distress with both a) upturned eyebrows and b) the SIRE dynamics of a comforting voice. A second way to store the *dynamics* of the face is to consider the fact that during speech, the mouth moves simultaneously. As a caregiver speaks to the robot, the visual movement of the face and auditory dynamics of the voice may be associated.

Indeed, integration of the face would be interesting because psychology has shown that each modality has its advantages. Whereas voice provides *activation* and is weak for valence [147], face readily provides *valence* information. 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.

This thesis also disregarded the sense of touch. However, a major source of emotional communication with infants is the sense of touch [148], and Clynes has shown that emotional touch or 'sentic' curves may even be universal [72]. It is possible that along with auditory and visual information, tactile information is also integrated, for example stroking an infant softly while comforting or bouncing its arms during play. In our human-robot experiment in Chapter 5, one subject stated that he wanted to "comfort the robot by touching him – that's how I communicate with my dog," and multiple

participants did touch the robot. Clearly, tactile sensors could be an important sense to integrate.

This simultaneous integration of all of these modalities should be considered for future work. Recent hypotheses in psychology called Neonatal Synaesthesia Hypothesis suggest that before the age of 4 months, infants are "synesthetic" [149]. That is, the specialized auditory and visual parts of their brain are not yet fully separated. A model for robots that incorporate this initial phase of tight modality coupling could also be an interesting direction for future work.

### 6.2.8 How does this explain how a robot might be moved by music?

We believe that an interesting thought experiment for testing whether a robot "has emotions", is whether it could be "moved" when it listens to music. As a result of this thesis, we may suggest a mechanism by which a robot could be moved by music. In particular, we noted that a) vocal SIRE dynamics and music SIRE dynamics are similar (Chapter 3) and b) a robot could be trained to associate SIRE voices with a physical flourishing or distressed state (Chapter 5). Therefore, in future work we could test whether emotional music could also activate deep, human-like feelings in the robot.

### 6.2.9 Emotion, cognition and language

Turning towards emotion's role in artificial intelligence, it would be very interesting to consider the relation between emotion and language development. Consider that emotion processes appear before language is acquired. Emotional vocalizations are distinguishable in the 5th or 7th month [36]. Yet, only around the age of three do babies acquire the ability to speak full sentences [42]. A recent literature review by Saint-Georges et al. [45] describes the nature of motherese and its links to emotion, cognition and language development: roboticists could use this as a guide for considering emotional processes as a scaffold for developing language and meaning.

### 6.2.10 Taking a cue from emotional development in humans

The philosophy of this thesis was to to use human development as a guide for creating an emotional robot. Here, we provide a final suggestion on how artificial intelligence

research could take inspiration from methods used to develop social skills in children with autism.

There are two main approaches to developing social skills among autistic children. The traditional approach to autism therapy is called Applied Behavior Analysis (ABA), which teaches children social interactions in a Skinner-like, behavioralist paradigm. For example, when a trainer says "I have done something nice yesterday," the child is taught to ask a follow-up question [150]. A second, growing trend is a recent autism therapy called "FloorTime"[1], which asks caregivers to help their child achieve developmental milestones through close, emotional interactions: "adults follow the child's lead utilizing affectively toned interactions through gestures and words" [151]. This work is based on the developmental, individual-difference, relationship-based (DIR) model.

In some ways, we could imagine that emotional robotics research until now has engineered robots following the first, behavioralist model – the robot is programmed to express itself following a pattern-recognition or rule-based approach, with a focus on functional interactions. Future work in robotics could take inspiration from this new DIR model, in which social interactions and theory of mind are emergent from a deep understanding rooted in emotions.

### 6.2.11 Do we really need real robot empathy?

Finally, some researchers in robotics are proponents of a functional robotic system (e.g., Brooks and Breazeal [3]). The major argument is that emotions do not actually have to be real [152], just to appear human-like and believable. Additionally, Reeve's and Nass' Media Equation [153] states that humans will project traits like emotion to machines. For example, simply by changing the words used by a computer, participants saw one computer as more "dominant", and the other as more "submissive".

The problem here is the same that we have for artificial intelligence in general – believability has been shown for short-term interactions, but it is not clear at what point the believability will break down. Consider a robot that interacts on a long-term, such as in a retirement home or in a user's house. At some point, a functionally emotional robot may say, "I'm sorry that you had a bad day,' followed by the human asking, "Are you truly

[1] http://www.stanleygreenspan.com/

sorry, or are you programmed to say that?" Indeed, even among fellow humans, showing "fake emotion" is considered a manipulative trait. For these reasons, theorists such as Sherry Turkle strongly oppose machines that show an emotional façade [154]. Therefore, "deep" artificial empathy is a challenge that must be addressed. A human-like development process for an emotional robot (e.g., trained to "feel" through caregiver-like interaction, associating emotions with its physical state), may be one way to improve believability and positive evaluations on sincerity.

# 7

# Conclusion

*"The purpose of our lives is to be happy."*

– Dalai Lama

The goal of this thesis was to design and implement a multimodal emotion system for robots at the primary emotion level. We aimed to complement artificial emotion systems which focus on secondary, or appraisal type models. We tackled this problem with a unique multimodal approach called SIRE (speed, intensity, irregularity, and extent), which we showed could explain the quick, emotional impact that voice, movement, and music have on us. We then introduced a model called multimodal emotional intelligence (MEI), a collection of Gaussian Mixture Models that allowed statistical learning of emotional expressions. The design of this MEI allowed it to serve as a basis for both affect recognition and expression across modalities, which had not been seen in previous emotion models. Finally, we showed how a MEI could be trained in a way similar to human infants. We placed the MEI system into a real-time interactive robot, to develop dynamic emotions through an empathetic loop with a caregiver. The resulting trained MEI was a low-level emotion system that could express and recognize emotions with grounded feelings, and serve as an informative foundation for secondary appraisal.

We emphasized the use of clues from infant development, and hope that this thesis helps to build the new area of developmental emotional robotics – one where robots are not trained through hand-annotated databases, but by continually making associations through natural interactions with humans, towards the goal of adaptive, emotional, empathetic robots.

# Bibliography


[1] Klaus R Scherer, Tanja Bänziger, and Etienne Roesch. *A Blueprint for Affective Computing: A sourcebook and manual*. Oxford University Press, 2010.

[2] Antonio Damasio. Descartes' error: Emotion, reason and the human mind. *New York: Grossett/Putnam*, 1994.

[3] Jean-Marc Fellous and Michael A Arbib. *Who needs emotions?: The brain meets the robot*. Oxford University Press Oxford, 2005.

[4] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[5] Angelica Lim, Tetsuya Ogata, and Hiroshi G Okuno. Towards expressive musical robots: a cross-modal framework for emotional gesture, voice and music. *EURASIP Journal on Audio, Speech, and Music Processing*, 2012(1):1–12, 2012.

[6] Angelica Lim, Tetsuya Ogata, and Hiroshi G Okuno. Converting emotional voice to motion for robot telepresence. In *Humanoid Robots (Humanoids), 2011 11th IEEE-RAS International Conference on*, pages 472–479. IEEE, 2011.

[7] Angelica Lim and Hiroshi G Okuno. The mei robot: Towards using motherese to develop multimodal emotional intelligence. *IEEE Transactions on Autonomous Mental Development*, (In press).

[8] Rainer Banse and Klaus R Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614, 1996.

[9] Albert Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*, 121(3):339–361, 1995.

[10] Klaus R Scherer. Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3459–3474, 2009.

[11] Michelle Karg, Robert Jenke, Kolja Kühnlenz, Martin Buss, et al. A two-fold pca-approach for inter-individual recognition of emotions in natural walking. In *MLDM Posters*, pages 51–61, 2009.

[12] Chiemi Onishi, Kyoko Yuasa, Masako Sei, Ashraf A Ewis, Takuro Nakano, Hokuma Munakata, and Yutaka Nakahori. Determinants of life satisfaction among japanese elderly women attending health care and welfare service facilities. *Journal of Medical Investigation*, 57:69–80, 2010.

[13] Cynthia L Breazeal. *Designing Sociable Robots*. MIT press, 2004.

[14] Dolores Cañamero. Modeling motivations and emotions as a basis for intelligent behavior. In *Proceedings of the first international conference on Autonomous agents*, pages 148–155. ACM, 1997.

[15] Juan D Velásquez. Modeling emotions and other motivations in synthetic agents. In *AAAI/IAAI*, pages 10–15. AAAI Press, 1997.

[16] Stein Bråten. *On being moved: From mirror neurons to empathy*. John Benjamins Publishing, 2007.

[17] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[18] Patrik N Juslin and Petri Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin*, 129(5):770, 2003.

[19] W Parrott. *Emotions in social psychology: Essential readings.* Psychology Press, 2001.

[20] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[21] Silvan S Tomkins. Affect theory. *Approaches to emotion*, 163:195, 1984.

[22] Keith Oatley and Philip N Johnson-Laird. Towards a cognitive theory of emotions. *Cognition and emotion*, 1(1):29–50, 1987.

[23] Jordi Vallverdu. Ekman's paradox and a naturalistic strategy to escape from it. *International Journal of Synthetic Emotions*, 4(2):1–8, 2013.

[24] Patrik N Juslin and John A Sloboda. *Handbook of music and emotion: Theory, research, applications*. Oxford University Press, 2010.

[25] Minoru Asada, Yukie Nagai, and Hisashi Ishihara. Why not artificial sympathy? In *Social Robotics*, pages 278–287. Springer, 2012.

[26] Elaine Hatfield, John T Cacioppo, and Richard L Rapson. Emotional contagion. *Current Directions in Psychological Science*, 2(3):96–99, 1993.

[27] Phoebe C Ellsworth and Klaus R Scherer. Appraisal processes in emotion. *Handbook of affective sciences*, pages 572–595, 2003.

[28] Andrew Ortony. *The cognitive structure of emotions*. Cambridge university press, 1990.

[29] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[30] James A Russell, Maria Lewicka, and Toomas Niit. A cross-cultural study of a circumplex model of affect. *Journal of personality and social psychology*, 57(5):848, 1989.

[31] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 827–834. IEEE, 2011.

[32] Paul Ekman and Wallace V Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, pages 49–98, 1969.

[33] Arlene S Walker-Andrews. Infants' perception of expressive behaviors: differentiation of multimodal information. *Psychological bulletin*, 121(3):437, 1997.

[34] Carol K Sigelman and Elizabeth A Rider. *Life span human development*. CengageBrain, 2010.

[35] Tobias Grossmann. The development of emotion perception in face and voice during infancy. *Restorative neurology and neuroscience*, 28(2):219–236, 2010.

[36] Elisabeth Scheiner, Kurt Hammerschmidt, Uwe Jürgens, and Petra Zwirner. Acoustic analyses of developmental changes and emotional expression in the preverbal vocalizations of infants. *Journal of Voice*, 16(4):509–529, 2002.

[37] Tobias Grossmann, Tricia Striano, and Angela D Friederici. Crossmodal integration of emotional information from face and voice in the infant brain. *Developmental Science*, 9(3):309–315, 2006.

[38] Joseph J Campos, David I Anderson, Marianne A Barbu-Roth, Edward M Hubbard, Matthew J Hertenstein, and David Witherington. Travel broadens the mind. *Infancy*, 1(2):149–219, 2000.

[39] Anne Fernald. Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages. *Child development*, 64(3):657–674, 1993.

[40] Nelson H Soken and Anne D Pick. Intermodal perception of happy and angry expressive behaviors by seven-month-old infants. *Child development*, 63(4):787–795, 1992.

[41] Ross Flom, Douglas A Gentile, and Anne D Pick. Infants discrimination of happy and sad music. *Infant Behavior and Development*, 31(4):716–728, 2008.

[42] Patricia K Kuhl. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11):831–843, 2004.

[43] Ayako Watanabe, Masaki Ogino, and Minoru Asada. Mapping facial expression to internal states based on intuitive parenting. *Journal of Robotics and Mechatronics*, 19(3):315, 2007.

[44] Sofiane Boucenna, Philippe Gaussier, Pierre Andry, and Laurence Hafemeister. Imitation as a communication tool for online facial expression learning and recognition. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 5323–5328. IEEE, 2010.

[45] Catherine Saint-Georges, Mohamed Chetouani, Raquel Cassel, Fabio Apicella, Ammar Mahdhaoui, Filippo Muratori, Marie-Christine Laznik, and David Cohen. Motherese in interaction: At the cross-road of emotion and cognition?(a systematic review). *PloS one*, 8(10):e78103, 2013.

[46] Toyoaki Nishida, Lakhmi C Jain, and Colette Faucher. *Modeling Machine Emotions for Realizing Intelligence*. Springer-Verlag Berlin Heidelberg, 2010.

[47] Catherine Pelachaud. Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630–639, 2009.

[48] Dongwoon Choi, Dong-Wook Lee, Duk Yeon Lee, Ho Seok Ahn, and Hogil Lee. Design of an android robot head for stage performances. *Artificial Life and Robotics*, 16(3):315–317, 2011.

[49] Massimiliano Zecca, Yu Mizoguchi, Keita Endo, Fumiya Iida, Yousuke Kawabata, Nobutsuna Endo, Kazuko Itoh, and Atsuo Takanishi. Whole body emotion expressions for kobian humanoid robot – preliminary experiments with different emotional patterns. In *Robot and Human Interactive Communication*, pages 381–386. IEEE, 2009.

[50] Aryel Beck, Antoine Hiolle, Alexandre Mazel, and Lola Cañamero. Interpretation of emotional body language displayed by robots. In *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, pages 37–42. ACM, 2010.

[51] Aryel Beck, Lola Cañamero, and Kim A Bard. Towards an affect space for robots to display emotional body language. In *RO-MAN, 2010 IEEE*, pages 464–469. IEEE, 2010.

[52] Mark L Knapp. *Nonverbal communication in human interaction*. Cengage Learning, 2012.

[53] Toru Nakata, Taketoshi Mori, and Tomomasa Sato. Quantitative analysis of impression of robot bodily expression based on laban movement theory. *Journal of the Robotics Society of Japan*, 19(2):104–111, 2001.

[54] Martin Saerbeck and Christoph Bartneck. Perception of affect elicited by robot motion. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 53–60. IEEE Press, 2010.

[55] Judith M Kessens, Mark A Neerincx, Rosemarijn Looije, Melanie Kroes, and Gerrit Bloothooft. Facial and vocal emotion expression of a personal computer assistant to engage, educate and motivate children. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE, 2009.

[56] Iain R Murray and John L Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93:1097, 1993.

[57] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80, 2001.

[58] Tetsuya Ogata, Akitoshi Shimura, Koji Shibuya, and Shigeki Sugano. A violin playing algorithm considering the change of phrase impression. In *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, volume 2, pages 1342–1347. IEEE, 2000.

[59] Jorge Solis, Kei Suefuji, Koichi Taniguchi, Takeshi Ninomiya, Maki Maeda, and Atsuo Takanishi. Implementation of expressive performance rules on the wf-4riii by modeling a professional flutist performance using nn. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 2552–2557. IEEE, 2007.

[60] Tomoyasu Nakano and Masataka Goto. Vocalistener: A singing-to-singing synthesis system based on iterative parameter estimation. *Proc. SMC 2009*, pages 343–348, 2009.

[61] Angelica Lim, Takeshi Mizumoto, Toru Takahashi, Tetsuya Ogata, and Hiroshi G Okuno. Programming by playing and approaches for expressive robot performances. In *IROS Workshop on Robots and Musical Expressions*, 2010.

[62] Gil Weinberg, Aparna Raman, and Trishul Mallikarjuna. Interactive jamming with shimon: a social robotic musician. In *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*, pages 233–234. IEEE, 2009.

[63] Yoshihiro Kusuda. Toyota's violin-playing robot. *Industrial Robot: An International Journal*, 35(6):504–506, 2008.

[64] Toyotoshi Yamada, Hideki Hashimoto, and Naoko Tosa. Pattern recognition of emotion with neural network. In *Industrial Electronics, Control, and Instrumentation, 1995., Proceedings of the 1995 IEEE IECON 21st International Conference on*, volume 1, pages 183–187. IEEE, 1995.

[65] Maurizio Mancini and Ginevra Castellano. Real-time analysis and synthesis of emotional gesture expressivity. In *Proc. of the Doctoral Consortium of Intl. Conf. on Affective Computing and Intelligent Interaction*. Citeseer, 2007.

[66] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[67] Raul Fernandez and Rosalind W Picard. Classical and novel discriminant features for affect recognition from speech. In *INTERSPEECH*, pages 473–476, 2005.

[68] Marie Tahon, Agnes Delaborde, and Laurence Devillers. Real-life emotion detection from speech in human-robot interaction: Experiments across diverse corpora with child and adult voices. In *INTERSPEECH*, pages 3121–3124, 2011.

[69] Michael A Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, 2008.

[70] Jean-Marc Fellous. From human emotions to robot emotions. *Architectures for Modeling Emotion: Cross-Disciplinary Foundations, American Association for Artificial Intelligence*, pages 39–46, 2004.

[71] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.

[72] Manfred Clynes. *Sentics: The touch of emotions*. Anchor Press Garden City, NY, 1977.

[73] Frank E Pollick, Helena M Paterson, Armin Bruderlin, and Anthony J Sanford. Perceiving affect from arm movement. *Cognition*, 82(2):B51–B61, 2001.

[74] Claire L Roether, Lars Omlor, Andrea Christensen, and Martin A Giese. Critical features for the perception of emotion from gait. *Journal of Vision*, 9(6), 2009.

[75] Joann M Montepare, Sabra B Goldstein, and Annmarie Clausen. The identification of emotions from gait information. *Journal of Nonverbal Behavior*, 11(1):33–42, 1987.

[76] Renée Van Bezooijen, Stanley A Otto, and Thomas A Heenan. Recognition of vocal expressions of emotion a three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology*, 14(4):387–406, 1983.

[77] Klaus R Scherer. Vocal affect expression: a review and a model for future research. *Psychological bulletin*, 99(2):143, 1986.

[78] Charles T Snowdon. Expression of emotion in non-human animals. *Handbook of affective sciences*, pages 457–480, 2003.

[79] Laura-Lee Balkwill and William Forde Thompson. A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music perception*, pages 43–64, 1999.

[80] Laura-Lee Balkwill, William Forde Thompson, and Rie Matsunaga. Recognition of emotion in japanese, western, and hindustani music by japanese listeners1. *Japanese Psychological Research*, 46(4):337–349, 2004.

[81] Joseph G Cunningham and Rebecca S Sterling. Developmental change in the understanding of affective meaning in music. *Motivation and emotion*, 12(4):399–413, 1988.

[82] Herbert Spencer. The origin and function of music. *Frasers Magazine*, 56:396–408, 1857.

[83] Beau Sievers, Larry Polansky, Michael Casey, and Thalia Wheatley. Music and movement share a dynamic structure that supports universal expressions of emotion. *Proceedings of the National Academy of Sciences*, 110(1):70–75, 2013.

[84] Steven R Livingstone, Ralf Muhlberger, Andrew R Brown, and William F Thompson. Changing musical emotion: A computational rule system for modifying score and performance. *Computer Music Journal*, 34(1):41–64, 2010.

[85] Antonio Camurri, Gualtiero Volpe, Giovanni De Poli, and Marc Leman. Communicating expressiveness and affect in multimodal interactive systems. *Multimedia, IEEE*, 12(1):43–53, 2005.

[86] Peggy E Gallaher. Individual differences in nonverbal behavior: Dimensions of style. *Journal of Personality and Social Psychology*, 63(1):133, 1992.

[87] Luca Mion and Giovanni De Poli. Score-independent audio features for description of music expression. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):458–466, 2008.

[88] Harald G Wallbott. Bodily expression of emotion. *European journal of social psychology*, 28(6):879–896, 1998.

[89] Kenji Amaya, Armin Bruderlin, and Tom Calvert. Emotion from motion. In *Graphics interface*, volume 96, pages 222–229. Citeseer, 1996.

[90] Angelica Lim, Takeshi Mizumoto, Louis-Kenzo Cahier, Takuma Otsuka, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. Robot musical accompaniment: integrating audio and visual cues for real-time synchronization with a human flutist. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 1964–1969. IEEE, 2010.

[91] Takeshi Mizumoto, Hiroshi Tsujino, Toru Takahashi, Tetsuya Ogata, and Hiroshi G Okuno. Thereminist robot: development of a robot theremin player with feedforward and feedback arm control based on a theremin's pitch model. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 2297–2302. IEEE, 2009.

[92] Dagen Wang and Shrikanth S Narayanan. Robust speech rate estimation for spontaneous speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(8):2190–2201, 2007.

[93] Philippe H Dejonckere, Marc Remacle, Elisabeth Fresnel-Elbaz, Virginie Woisard, Lise Crevier-Buchman, and Benoîte Millet. Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Revue de laryngologie-otologie-rhinologie*, 117(3):219, 1996.

[94] Myron Ross, Harry Shaffer, Andrew Cohen, Richard Freudberg, and Harold Manley. Average magnitude difference function pitch extractor. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 22(5):353–362, 1974.

[95] Hideki Kenmochi and Hayato Ohshita. Vocaloid-commercial singing synthesizer based on sample concatenation. In *INTERSPEECH*, pages 4009–4010, 2007.

[96] Dik J Hermes. Measurement of pitch by subharmonic summation. *The journal of the acoustical society of America*, 83:257, 1988.

[97] Patrik N Juslin. Can results from studies of perceived expression in musical performances be generalized across response formats? *Psychomusicology: A Journal of Research in Music Cognition*, 16(1-2):77, 1997.

[98] Phoebe Ellsworth and J Merrill Carlsmith. Eye contact and gaze aversion in an aggressive encounter. *Journal of Personality and Social Psychology*, 28(2):280, 1973.

[99] RA Hinde and TE Rowell. Communication by postures and facial expressions in the rhesus monkey (macaca mulatta). *Proceedings of the Zoological Society of London*, 138(1):1–21, 1962.

[100] Ginevra Castellano, Santiago D Villalba, and Antonio Camurri. Recognising human emotions from body movement and gesture dynamics. In *Affective computing and intelligent interaction*, pages 71–82. Springer, 2007.

[101] Manav Bhaykar, Jainath Yadav, and K Sreenivasa Rao. Speaker dependent, speaker independent and cross language emotion recognition from speech using gmm and hmm. In *Communications (NCC), 2013 National Conference on*, pages 1–5. IEEE, 2013.

[102] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer, 2010.

[103] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[104] Michael Lewis. Self-conscious emotions. *Handbook of emotions*, 2:623–636, 2000.

[105] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[106] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):881–892, 2002.

[107] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 1. Springer, New York, 2006.

[108] Panu Somervuo and Teuvo Kohonen. Self-organizing maps and learning vector quantization for feature sequences. *Neural Processing Letters*, 10(2):151–159, 1999.

[109] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE, 2011.

[110] Tim Polzehl, Alexander Schmitt, and Florian Metze. Approaching multi-lingual emotion recognition from speech-on language dependency of acoustic/prosodic features for anger detection. In *Proceedings of the Fifth International Conference on Speech Prosody*, 2010.

[111] Lloyd Peterson and Margaret Jean Peterson. Short-term retention of individual verbal items. *Journal of experimental psychology*, 58(3):193, 1959.

[112] Kazuhiro Nakadai, Toru Takahashi, Hiroshi G Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. Design and implementation of robot audition system'hark'open source software for listening to three simultaneous speakers. *Advanced Robotics*, 24(5-6):739–761, 2010.

[113] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[114] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron

Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[115] Michelle Karg, Kolja Kuhnlenz, and Martin Buss. Recognition of affect based on gait patterns. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 40(4):1050–1061, 2010.

[116] Daniel Janssen, Wolfgang I Schöllhorn, Jessica Lubienetzki, Karina Fölling, Henrike Kokenge, and Keith Davids. Recognition of emotions in gait patterns by means of artificial neural nets. *Journal of Nonverbal Behavior*, 32(2):79–92, 2008.

[117] Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi. Fourier principles for emotion-based human figure animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 91–96. ACM, 1995.

[118] Joann Montepare, Elissa Koff, Deborah Zaitchik, and Marilyn Albert. The use of body movements and gestures as cues to emotions in younger and older adults. *Journal of Nonverbal Behavior*, 23(2):133–152, 1999.

[119] Yingliang Ma, Helena M Paterson, and Frank E Pollick. A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior research methods*, 38(1):134–141, 2006.

[120] Michael Sellers. Toward a comprehensive theory of emotion for biological and artificial agents. *Biologically Inspired Cognitive Architectures*, 2013.

[121] Pierre-yves Oudeyer. The synthesis of cartoon emotional speech. In *Speech Prosody 2002, International Conference*, 2002.

[122] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Interspeech*, pages 1517–1520, 2005.

[123] Klaus R Scherer. A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology. In *INTERSPEECH*, pages 379–382, 2000.

[124] Hun-ok Lim, Akinori Ishii, and Atsuo Takanishi. Emotion-based biped walking. *Robotica*, 22(5):577–586, 2004.

[125] Jung Ju Choi, Yunkyung Kim, and Sonya S Kwak. Have you ever lied?: the impacts of gaze avoidance on people's perception of a robot. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 105–106. IEEE Press, 2013.

[126] Tanya J Clarke, Mark F Bradshaw, David T Field, Sarah E Hampson, David Rose, et al. The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception*, 34(10):1171–1180, 2005.

[127] Katsuo Kuji. Applicability of "rem", revised emotions measurement, to marketing analysis. *Advances in Consumer Studies*, 15:57–76, 2009.

[128] Edmund T Rolls. Precis of the brain and emotion. *Behavioral and brain sciences*, 23(2):177–191, 2000.

[129] Ferrinne Spector and Daphne Maurer. Synesthesia: a new approach to understanding the development of perception. *Developmental psychology*, 45(1):175, 2009.

[130] Christian Becker-Asano and Ipke Wachsmuth. Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems*, 20(1):32–49, 2010.

[131] Laurel J Trainor, Caren M Austin, and Renée N Desjardins. Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological science*, 11(3):188–195, 2000.

[132] Anne Fernald, Traute Taeschner, Judy Dunn, Mechthild Papousek, Bénédicte de Boysson-Bardies, and Ikuko Fukui. A cross-language study of prosodic mod-

ifications in mothers and fathers speech to preverbal infants. *Journal of child language*, 16(3):477–501, 1989.

[133] Deborah G Kemler Nelson, Kathy Hirsh-Pasek, Peter W Jusczyk, and Kimberly Wright Cassidy. How the prosodic cues in motherese might assist language learning. *Journal of child Language*, 16(01):55–68, 1989.

[134] Anne Fernald. Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child development*, pages 1497–1510, 1989.

[135] Morris Michael Lewis. *Infant speech; a study of the beginnings of language.* Harcourt, Brace, 1936.

[136] Jean Berko Gleason. Talking to children: Some notes on feedback. *Talking to Children*, pages 199–205, 1977.

[137] Anne Fernald and Thomas Simon. Expanded intonation contours in mothers' speech to newborns. *Developmental psychology*, 20(1):104, 1984.

[138] Nadja Reissland and Terence Stephenson. Turn-taking in early vocal interaction: a comparison of premature and term infants vocal interaction with their mothers. *Child: care, health and development*, 25(6):447–456, 1999.

[139] Catherine T Best. Accommodation in mean f, during mother-infant and father-infant vocal interactions: a longitudinal case study. *J. Child Lang*, 4(7):9–736, 1997.

[140] Antonio R Damasio. *Looking for Spinoza: Joy, sorrow and the feeling brain*. Random House, 2004.

[141] Björn Schuller and Anton Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, 2013.

[142] Christine Kitamura and Christa Lam. Age-specific preferences for infant-directed affective intent. *Infancy*, 14(1):77–100, 2009.

[143] Bénédicte de Boysson-Bardies and Malcolm B DeBevoise. *How language comes to children: From birth to two years*. MIT Press Cambridge, 1999.

[144] Joseph A DeVito, Susan O'Rourke, and Linda O'Neill. *Human communication*. Addison Wesley, 2000.

[145] Charles S Carver and Eddie Harmon-Jones. Anger is an approach-related affect: evidence and implications. *Psychological bulletin*, 135(2):183, 2009.

[146] Ai Kawakami, Kiyoshi Furukawa, Kentaro Katahira, and Kazuo Okanoya. Sad music induces pleasant emotion. *Frontiers in psychology*, 4, 2013.

[147] James A Russell, Jo-Anne Bachorowski, and José-Miguel Fernández-Dols. Facial and vocal expressions of emotion. *Annual review of psychology*, 54(1):329–349, 2003.

[148] Matthew J Hertenstein. Touch: Its communicative functions in infancy. *Human Development*, 45(2):70–94, 2002.

[149] Simon Baron-Cohen. Is there a normal phase of synaesthesia in development. *Psyche*, 2(27):223–228, 1996.

[150] Bibi Huskens, Rianne Verschuur, Jan Gillesen, Robert Didden, and Emilia Barakova. Promoting question-asking in school-aged children with autism spectrum disorders: Effectiveness of a robot intervention compared to a human-trainer intervention. *Developmental neurorehabilitation*, 16(5):345–356, 2013.

[151] Serena Wieder and Stanley I Greenspan. Climbing the symbolic ladder in the dir model through floor time/interactive play. *Autism*, 7(4):425–435, 2003.

[152] Luisa Damiano, Paul Dumouchel, and Hagen Lehmann. Should empathic social robots have interiority? In *Social Robotics*, pages 268–277. Springer, 2012.

[153] Byron Reeves and C Nass. *The Media equation: how people treat computers, television, and new media*. Cambridge University Press, 1997.

[154] Sherry Turkle. *Alone together: Why we expect more from technology and less from each other*. Basic Books, 2012.

# List of Publications

## Journal Papers

1) **Angelica Lim** and Hiroshi G. Okuno: The MEI Robot: Towards Using Motherese to Develop Multimodal Emotional Intelligence, IEEE Transactions on Autonomous Mental Development, Accepted with minor revisions, Jan. 21, 2014. → **Chapter 4**.

2) **Angelica Lim**, Tetsuya Ogata, and Hiroshi G. Okuno: Towards expressive musical robots: A cross-modal framework for emotional gesture, voice and music, EURASIP Journal on Audio, Speech, and Music Processing, 2012:3. DOI: 10.1186/1687-4722-2012-3 → **Chapter 3**.

3) **Angelica Lim**, Takeshi Mizumoto, Tetsuya Ogata, and Hiroshi G. Okuno: A musical robot that synchronizes with a co-player using non-verbal cues, Advanced Robotics, Special Issue on Cutting Edge of Robotics in Japan, Vol. 26, pp.363–381 (2012)

## International Conference (Peer-reviewed)

1) **Angelica Lim** and Hiroshi G. Okuno: Using speech data to recognize emotion in human gait, *Proceedings of IEEE/RSJ HBU Workshop*, Portugal, LNCS Vol. 7559, pp.52–64 (2012). → **Chapter 4**.

2) **Angelica Lim**, Tetsuya Ogata, and Hiroshi G. Okuno: Converting emotional voice to motion for robot telepresence, *Proceedings of IEEE/RSJ Humanoids*, pp. 472–479, Bled, Slovenia, Oct. 2011. → **Chapter 3**.

3) **Angelica Lim**, Takeshi Mizumoto, Louis-Kenzo Cahier, Takuma Otsuka, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno: Robot Musical Accompaniment: Integrating Audio and Visual Cues for Real-time Synchronization with a Human Flutist, *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.1964-1969, Taipei, Oct. 2010.

4) **Angelica Lim**, Takeshi Mizumoto, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno: Programming by Playing and Approaches for Expressive Robot Performances, *Workshop on Robots and Musical Expression*, Taipei, Oct. 2010.