SFU SCHOOL OF COMPUTING SCIENCE

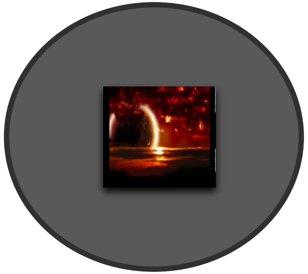
Dynamic Signal Processing (SIRE)

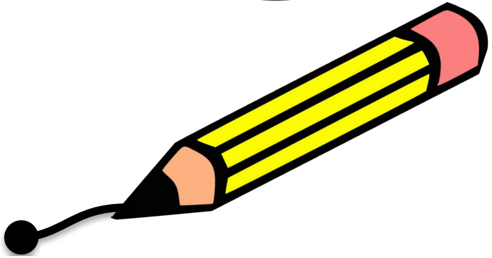
CMPT 419/983, Summer 2020

Dr. Angelica Lim

This lecture will be recorded and linked in Canvas.
You will be able to download it, but please don't post it anywhere. Thanks!

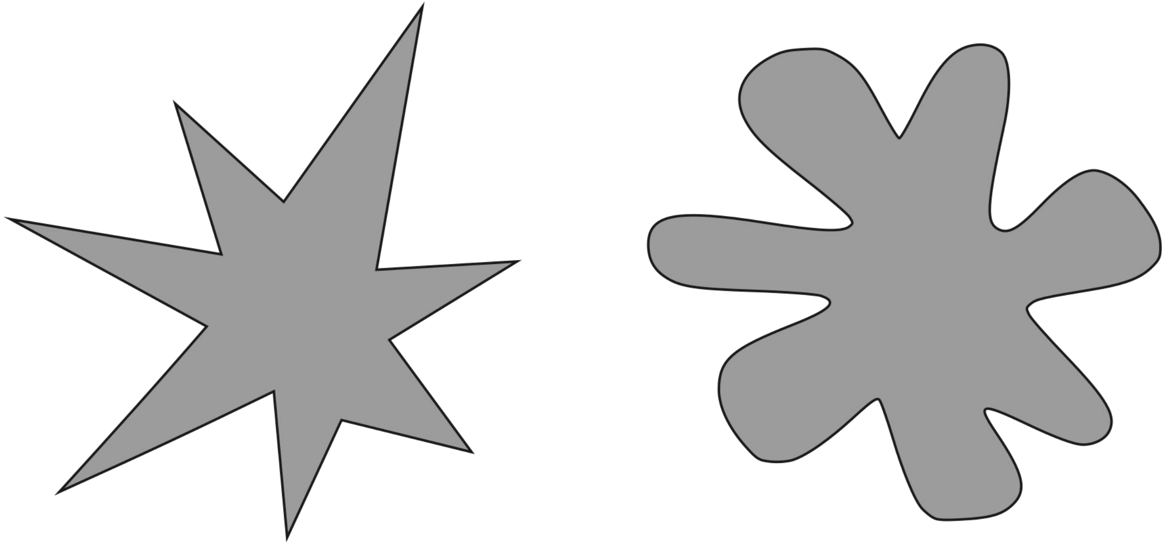
Let’s do a drawing exercise





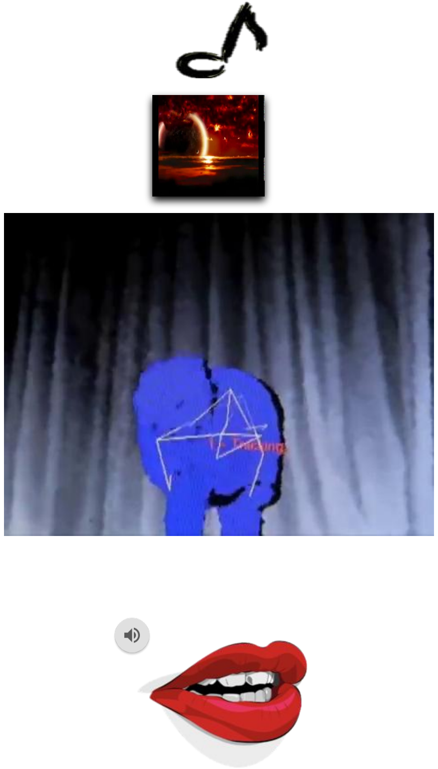
Using a pen or pencil, draw along with the music, without taking your pen off the paper.

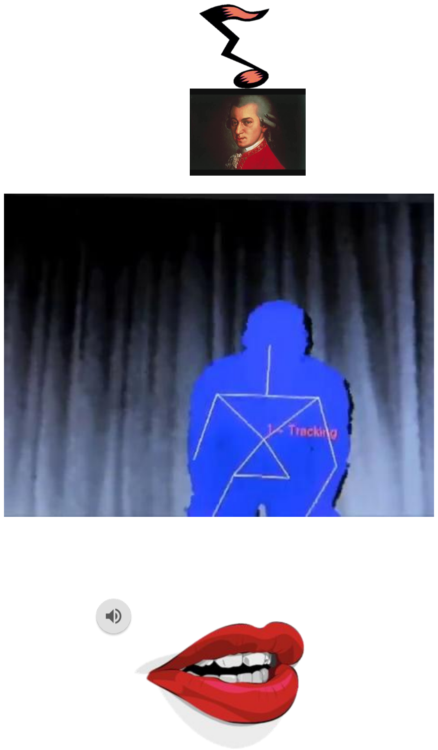
One of these shapes is called *Kiki* and the other is called *Bouba*.

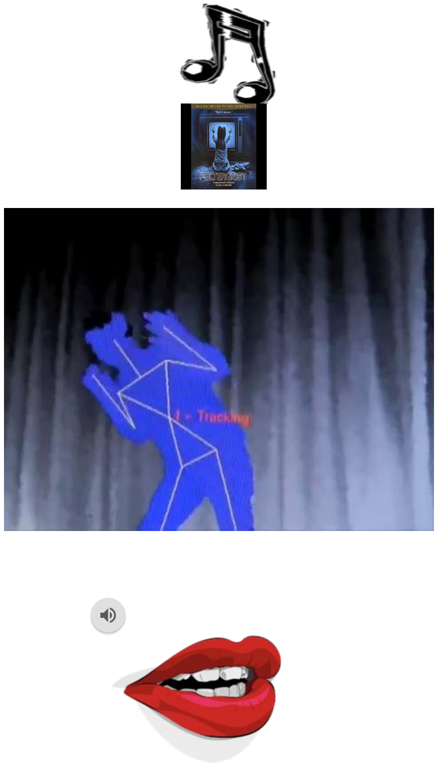


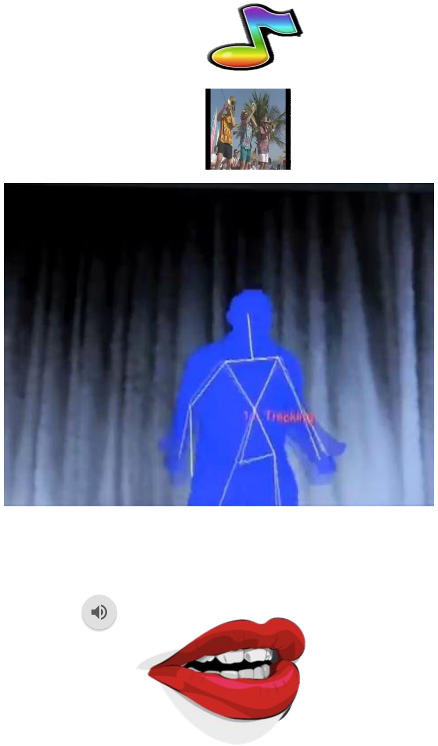
The shape of boubas: sound–shape correspondences in toddlers and adults (Maurer, 2006)

It’s long been suggested that **emotions** from **different modalities** have the same underlying ‘code’.
(Juslin & Laukka, 2003)







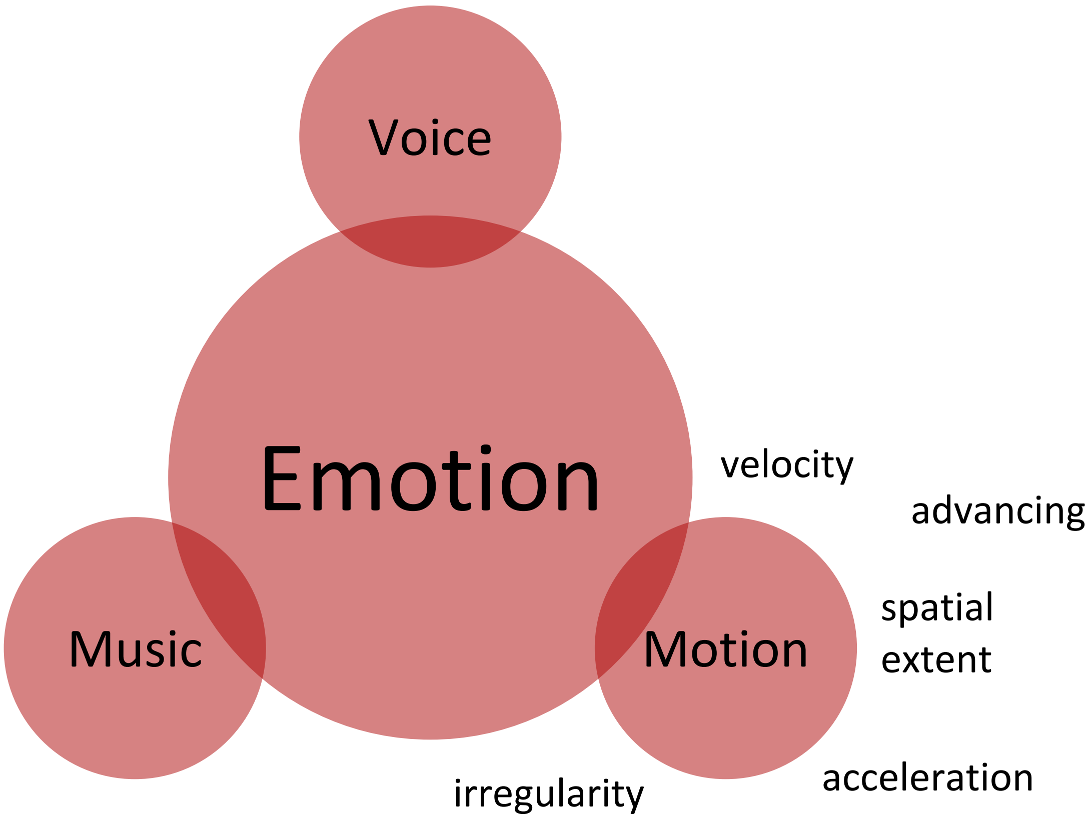


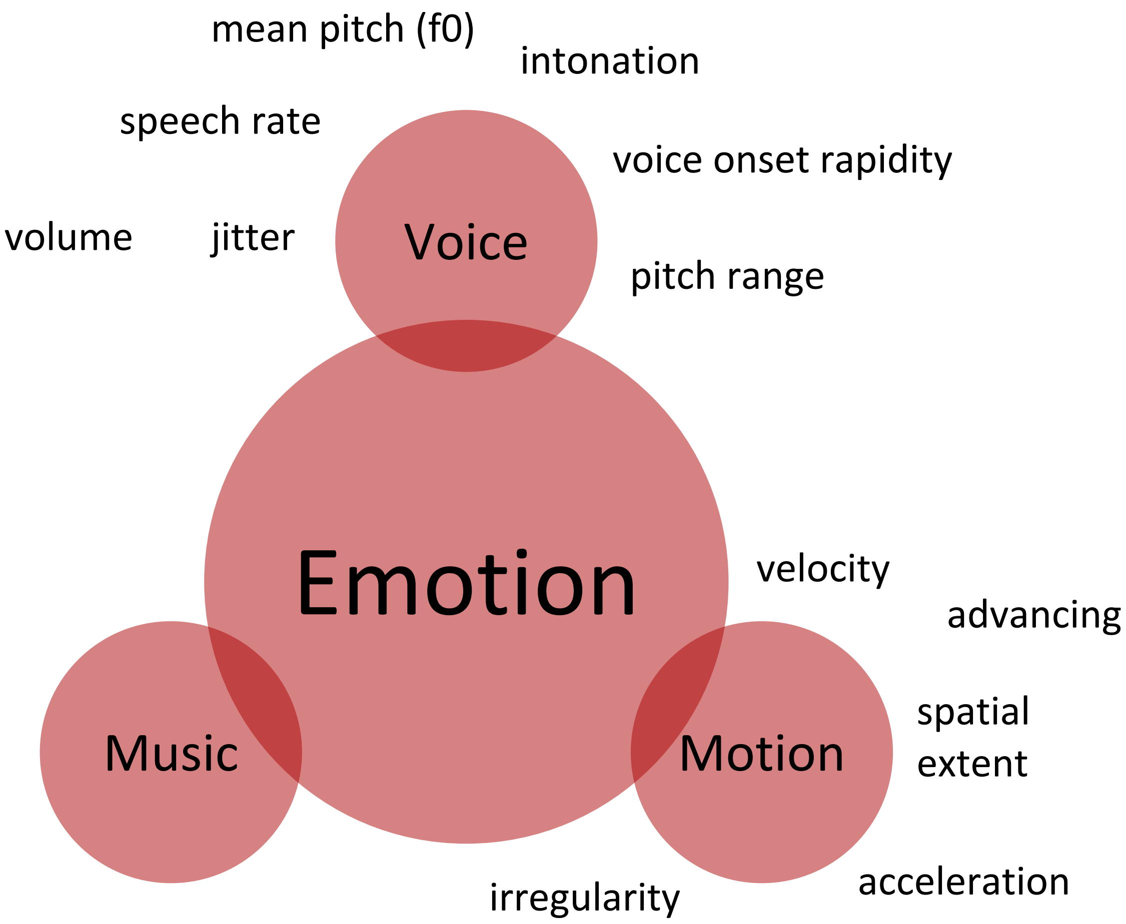
Dynamic Features

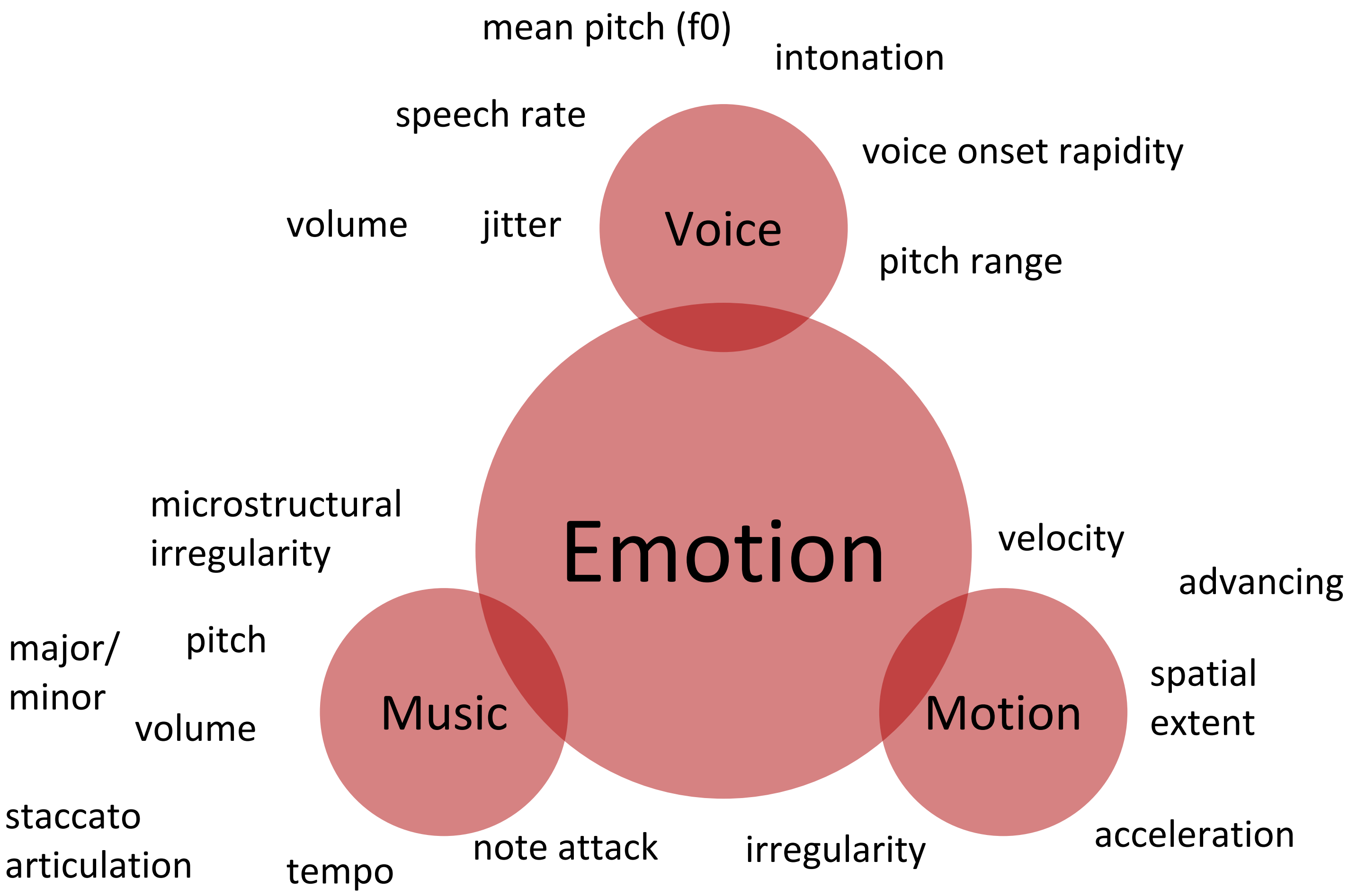
Dynamic Data

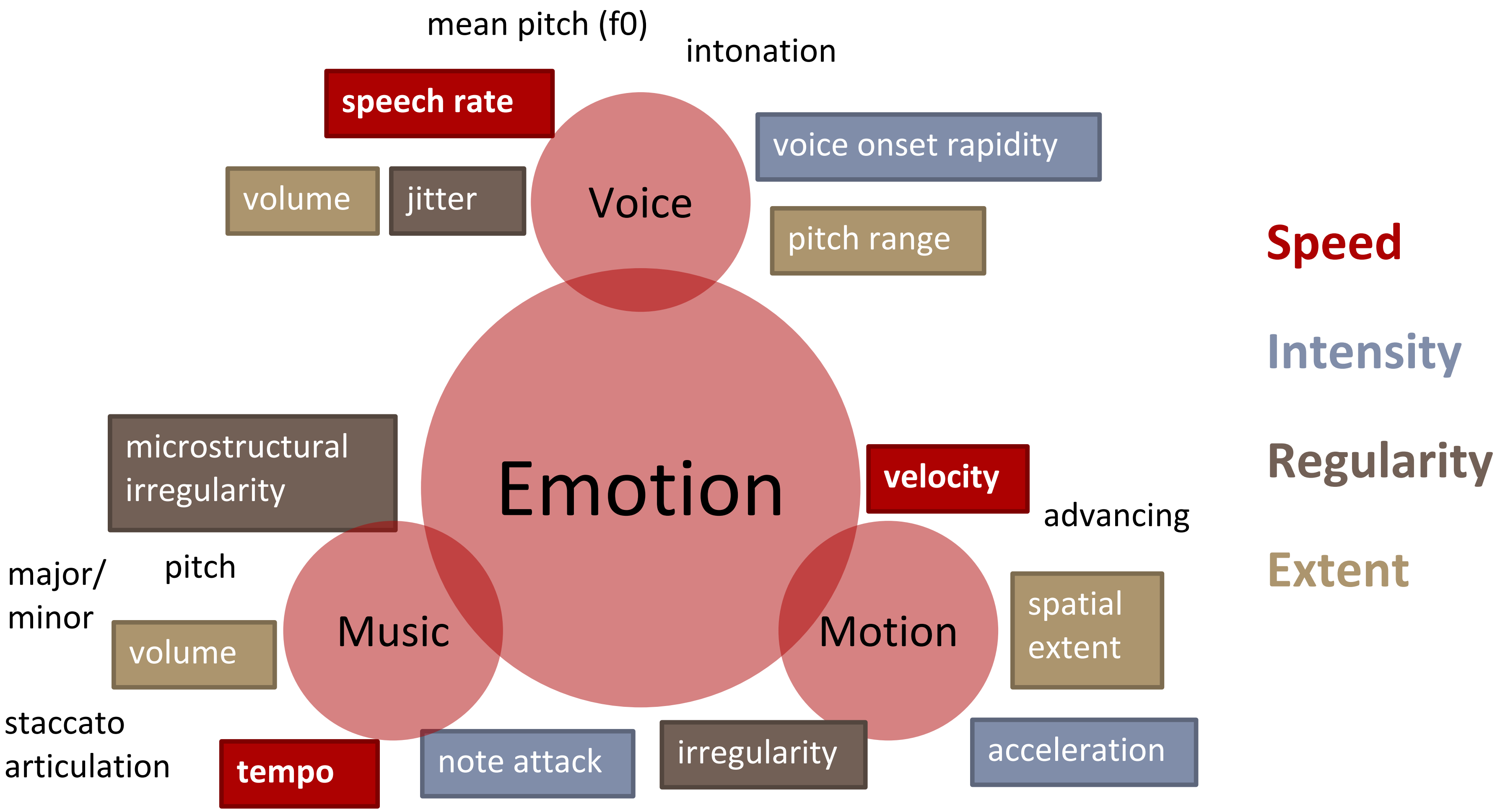
Temporal or dynamic data, also known as "time series" data, can be processed using:

- **Distance**-based methods ⇒ Compute the distance between pairs of time series (e.g. Dynamic Time Warping)
- **Feature**-based methods ⇒ Transform data into lower-dimensional feature vectors, before applying traditional classification techniques
- **Model**-based methods ⇒ Use a model such as Hidden Markov Model (HMM), Recurrent Neural Network (RNN), etc.


Voice
Emotion
Music
Motion
velocity
advancing
spatial
extent
acceleration
irregularity


mean pitch (f0)
intonation
speech rate
voice onset rapidity
volume
jitter
Voice
pitch range
Emotion
velocity
advancing
spatial
extent
Music
Motion
acceleration
irregularity

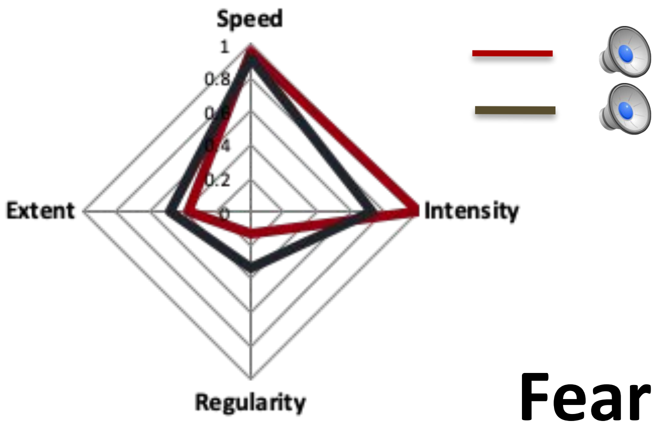

mean pitch (f0)
intonation
speech rate
voice onset rapidity
volume
jitter
Voice
pitch range
microstructural
irregularity
Emotion
velocity
advancing
major/
minor
pitch
volume
Music
Motion
spatial
extent
staccato
articulation
tempo
note attack
irregularity
acceleration



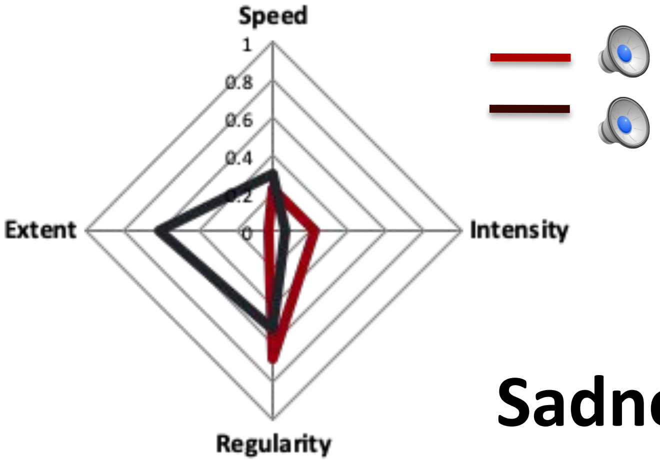
mean pitch (f0)
intonation
speech rate
voice onset rapidity
volume
jitter
Voice
pitch range
Speed
Intensity
Regularity
Extent
microstructural irregularity
Emotion
velocity
advancing
major/ minor
pitch
Music
Motion
spatial extent
volume
staccato articulation
tempo
note attack
irregularity
acceleration

Dynamic parameters for voice

Speed	Intensity	Regularity	Extent
Syllables per second	*Voice onset* $$p(k) = \sum_{i=0}^{n-1} x(k \cdot n + i)^2$$ $$\max_{k=1,\ldots,N/n} p(k) - p(k-1)$$	*(Inverse) jitter* $$\frac{1}{N-1} \sum_{t=1}^{N} \lvert x(t) - x(t-1) \rvert$$	*Pitch range*



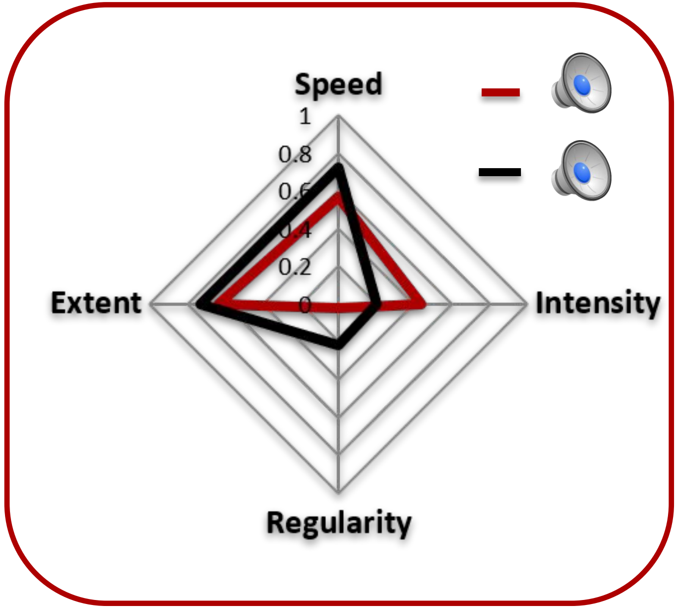
Fear



Sadness

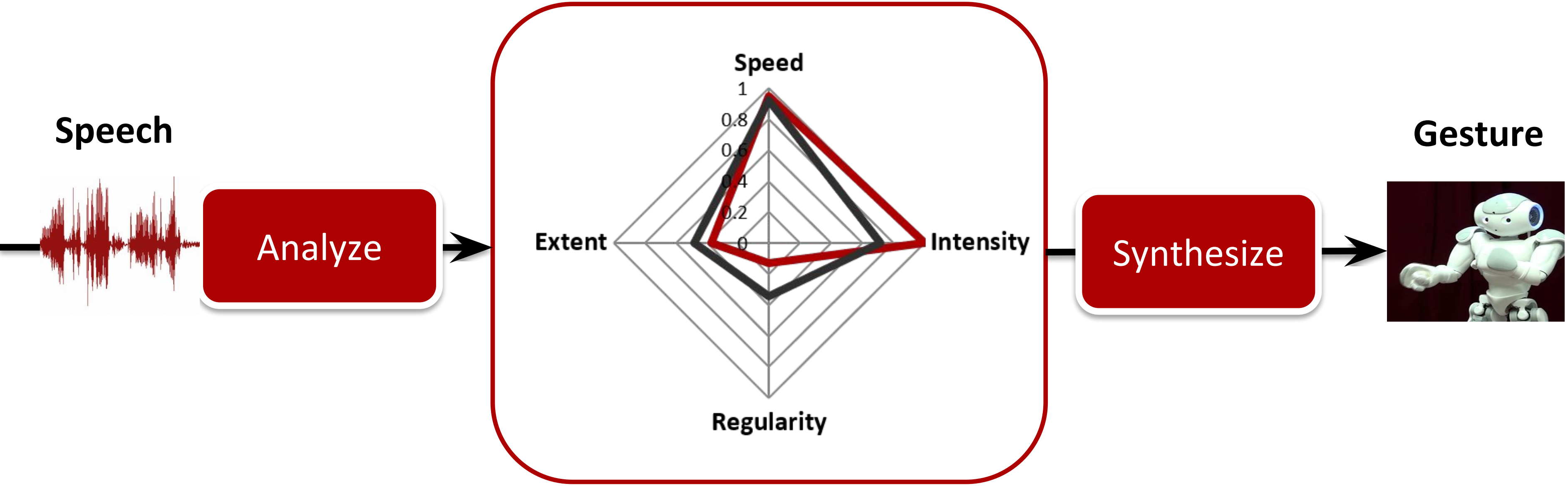
The SIRE Model

SIRE: A description of emotion using Speed, Intensity, Regularity, Extent. We represent an emotional expression using its dynamic features.



Lim et al. 2011

The SIRE Emotion Transfer System



Lim et al. 2011

Gesture Parameter Mappings

Speed

Joint speed

$$t_0 = t_0$$
$$t_1 = \max(S \cdot t_1, \underline{m})$$
$$t_2 = \max(S \cdot t_2, \underline{m})$$

Intensity

Joint acceleration

$$t_0 = t_0$$
$$t_1 = \max(I \cdot t_1, \underline{m})$$
$$t_2 = t_2$$

Regularity

Phase offset

$$\delta_t = (1 - R) \cdot \underline{r}$$
$$t_i = \delta_t + t_i$$

Extent

Gesture size

$$p_1 = p_0 + E \cdot (p_1 - p_0)$$

0




1




Converting emotional voice to gesture

Purpose: To verify how well can be used to represent emotion across voice and gesture

Materials:

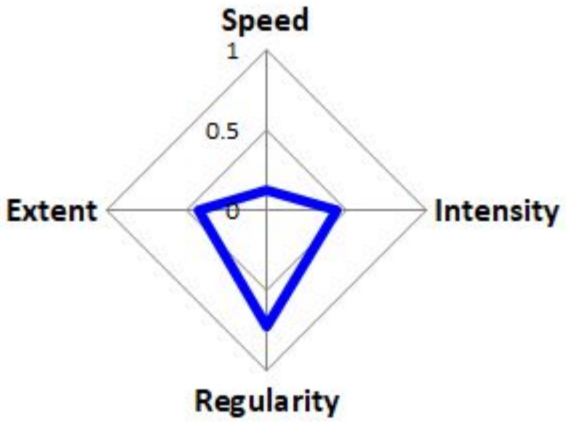
- 16 German voice samples of the same text
- 4 instances for each of **happiness**, **sadness**, **anger**, and **fear** (>80% agreement from German raters)

Apparatus: NAO simulated & embodied robot

Procedure:

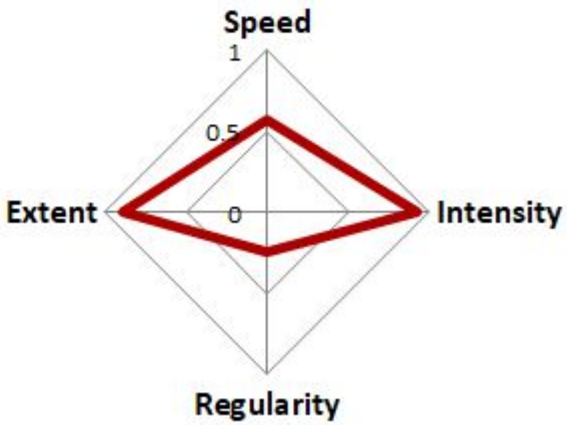
- Generated 16 motion sequences using 3 motion templates
- 3 conditions: voice only, motion only, motion + voice
- 21 Japanese participants selected one emotion the sequence best conveyed

Sadness



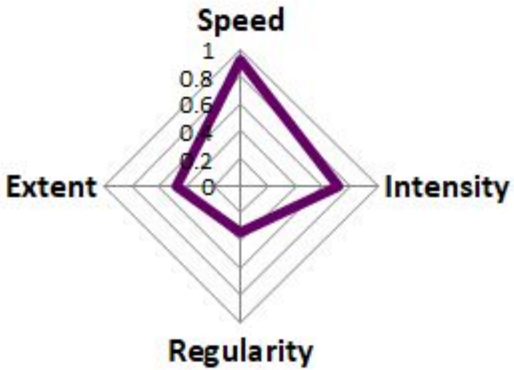
Agreement: 75%

Anger



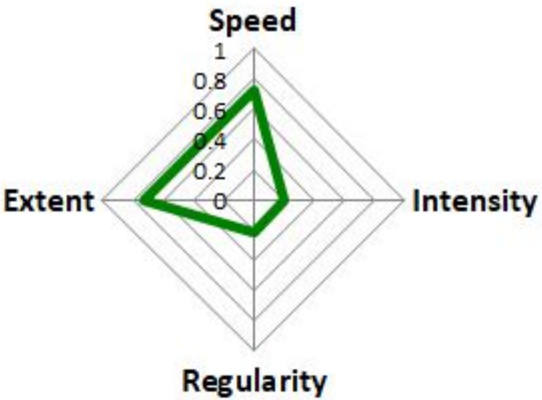
Agreement: 60%

Fear



Agreement: 65%

Happiness

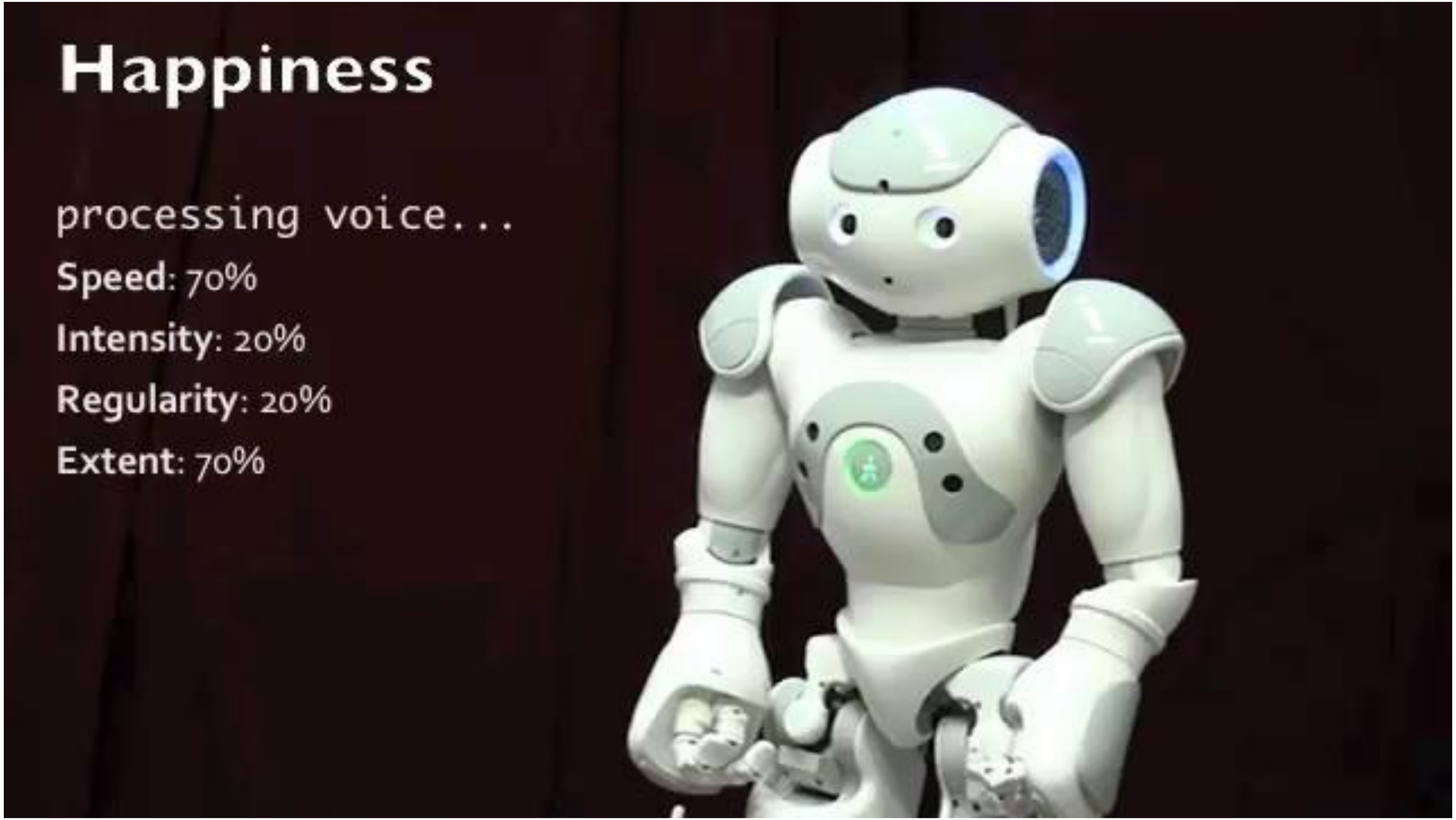


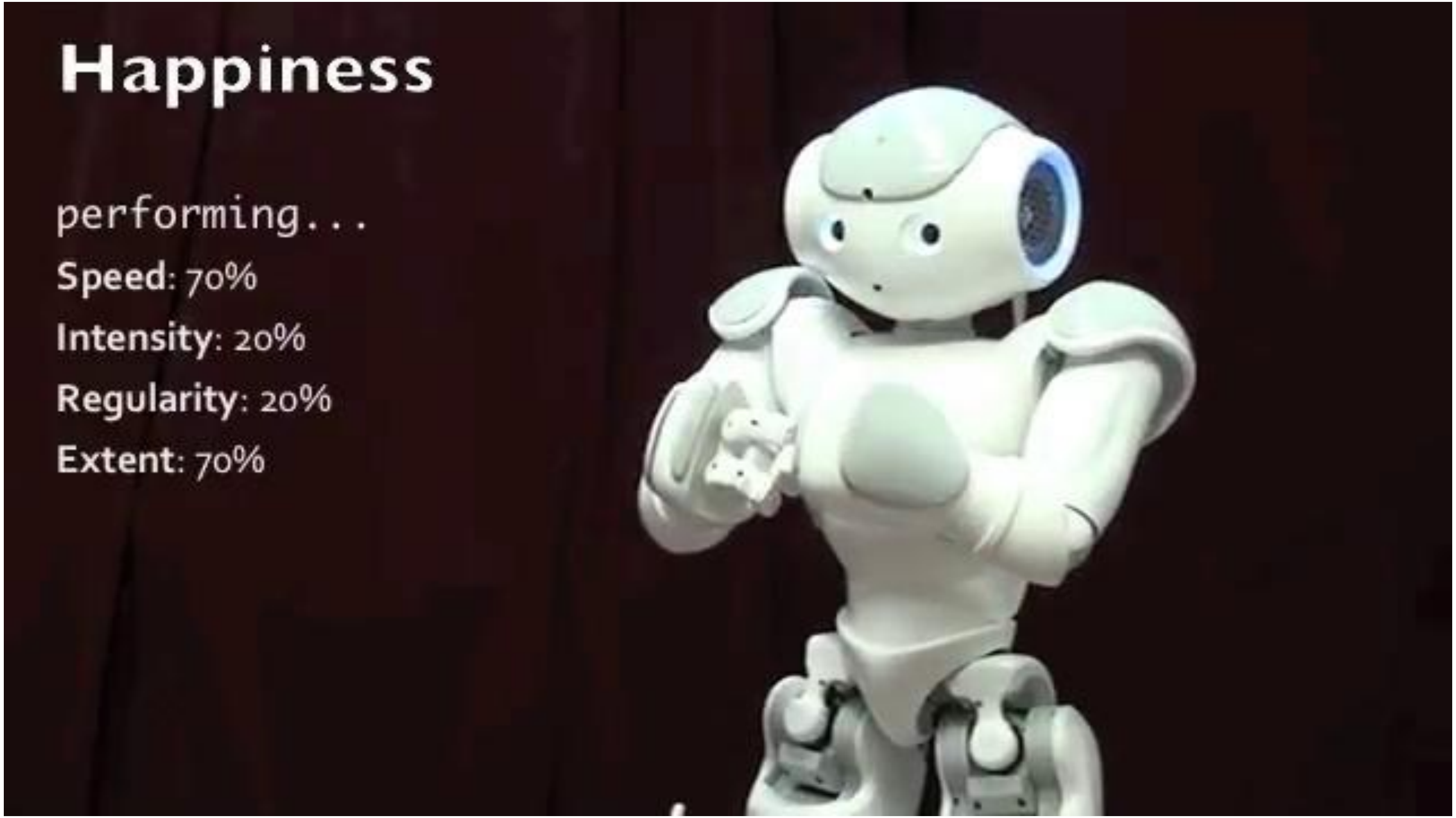
Agreement: 60%

Cross-modal SIRE values

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

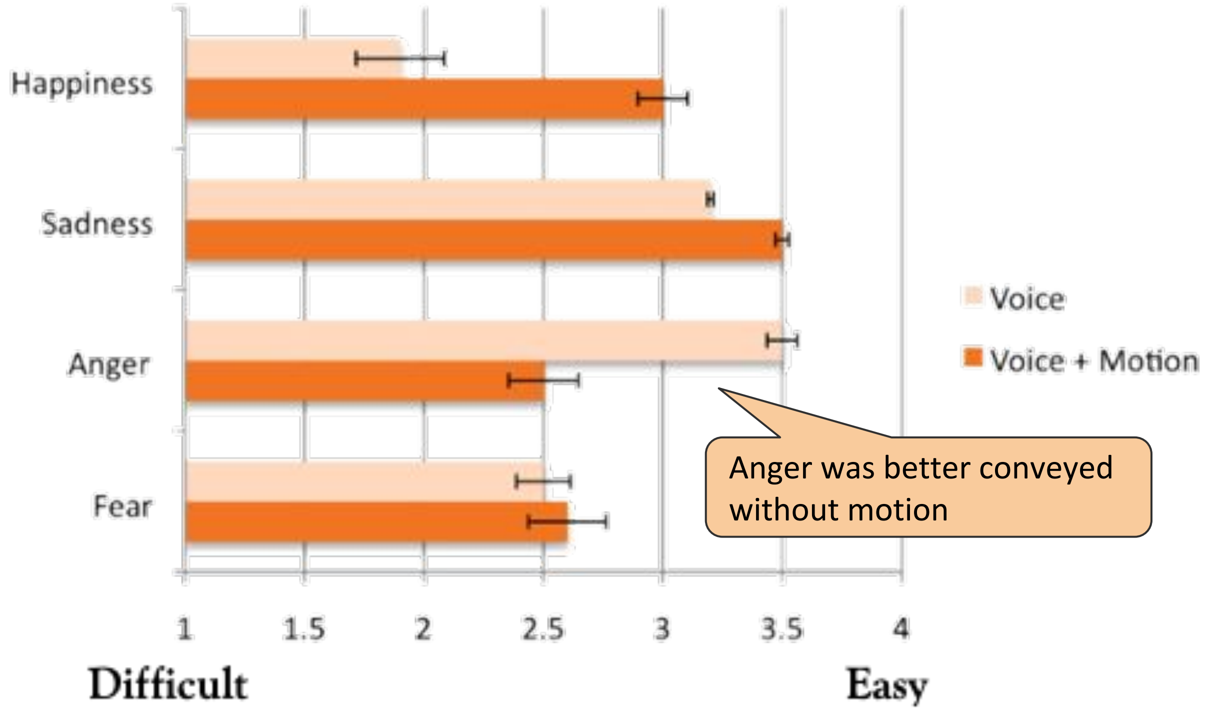
Lim et al., 2012

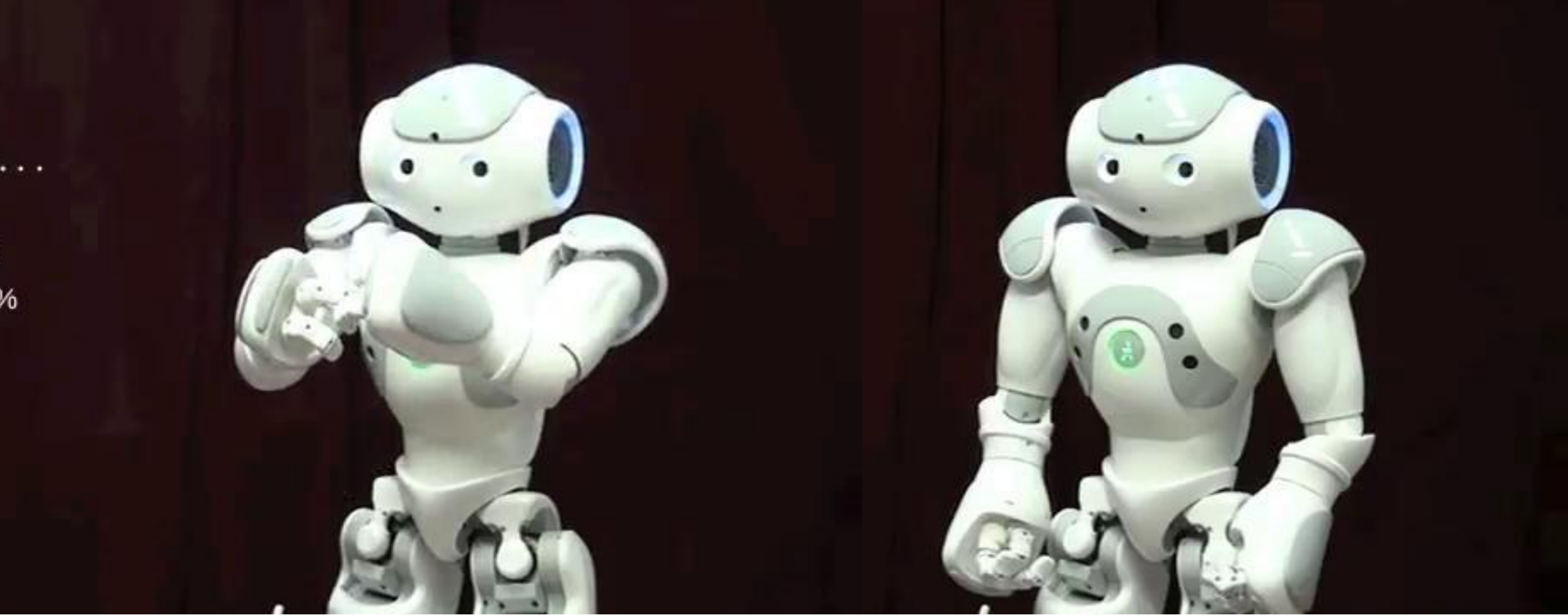

Happiness
processing voice...
Speed: 70%
Intensity: 20%
Regularity: 20%
Extent: 70%


Happiness
performing...
Speed: 70%
Intensity: 20%
Regularity: 20%
Extent: 70%

Multimodal Emotion

Adding SIRE motion to voice increased ease of understanding for most emotions:





A motionless head can look threatening.

The role of irregularity

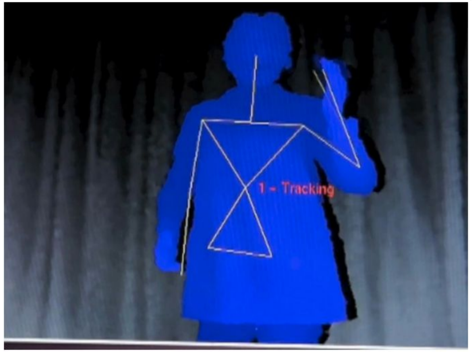


Figure 2 Body pose estimation using the Kinect 3D sensor to extract hand locations.

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

Gesture → SIE → Voice

Towards expressive musical robots: a cross-modal framework for emotional gesture, voice and music (2012)

Gesture → Voice (SIE)

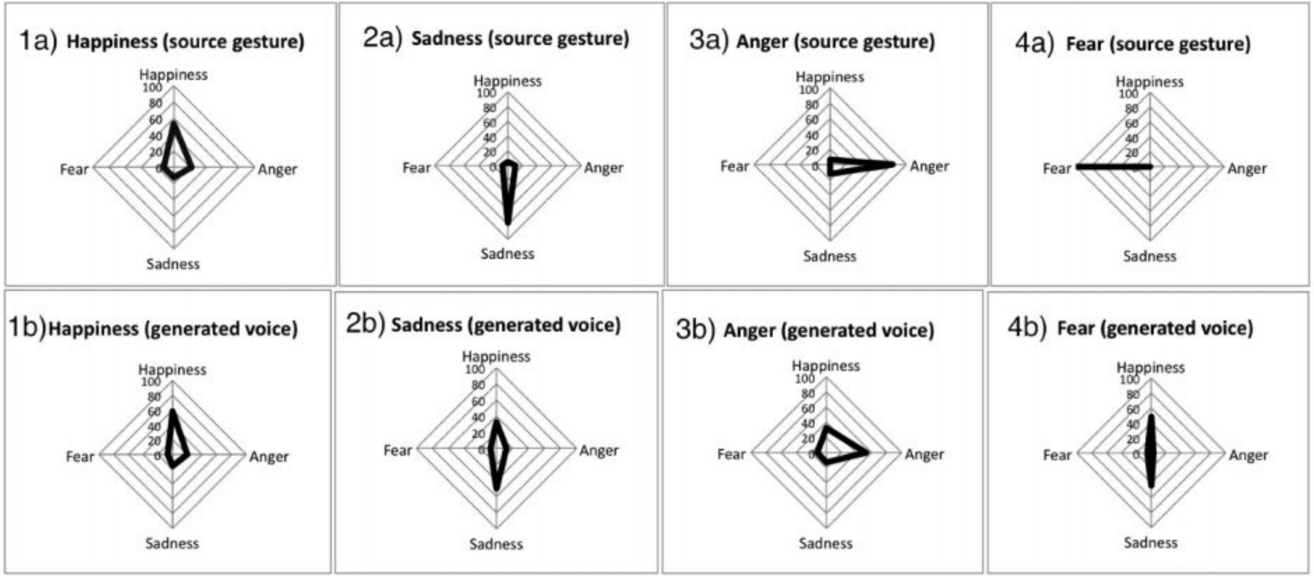


Figure 5 Experiment 1: Visualization of confusion matrices for gesture and voice. Intended emotion is shown in the titles, and the average percentage of raters that selected each emotion are given along the dimensional axes. Pointed triangles indicate that the one emotion was greatly perceived on average. Similar shapes for a given number indicate similar perceived emotion for both input gesture and output voice.

Happiness, sadness, and anger were transferred at greater than chance, despite the varied gestural interpretations for each emotion.

Fear was not well transferred. 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.

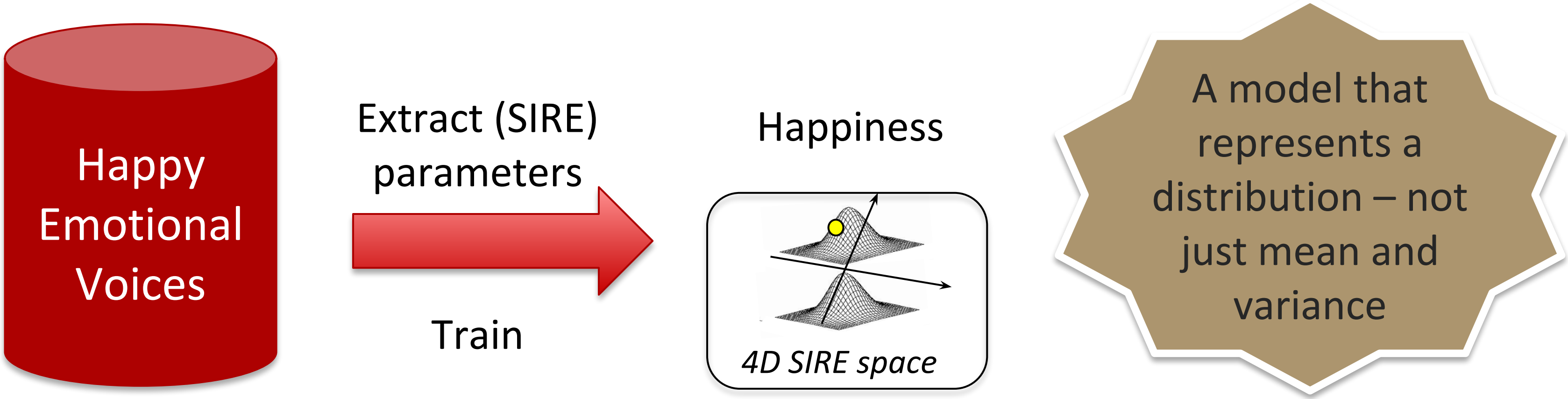
Learning with Dynamic Features

Modeling Emotions as a GMM for Statistical Learning

We are statistical learning machines.

The Gaussian Mixture Model (GMM)

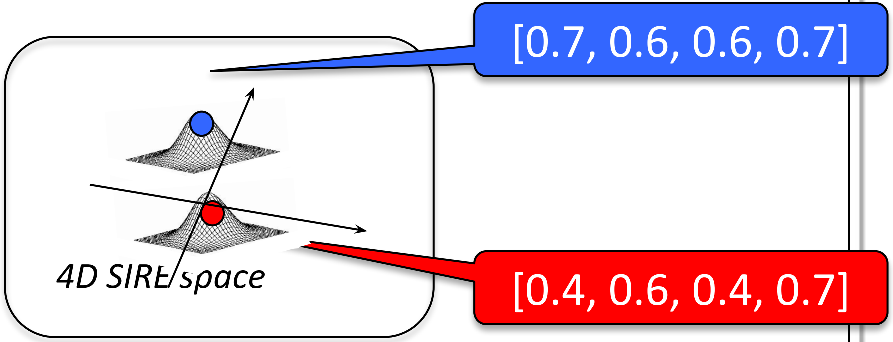
e.g.

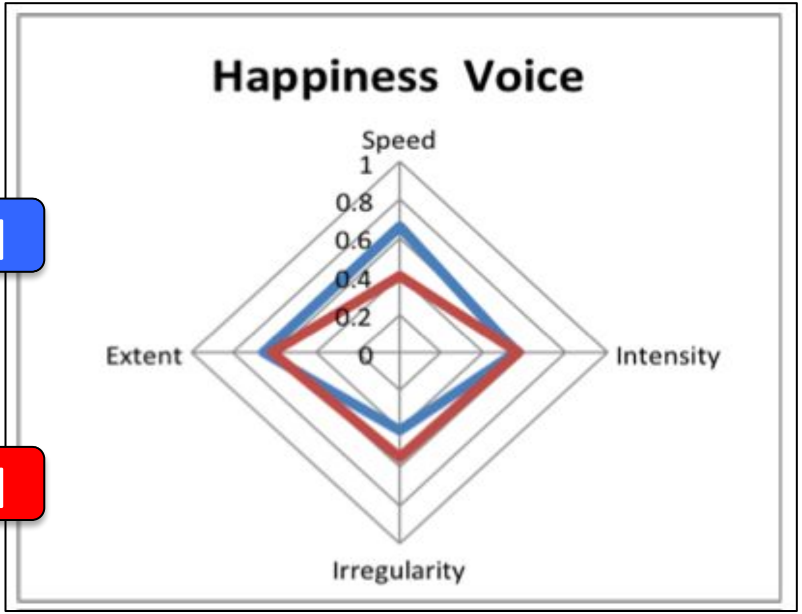


1. The GMM Represents the Knowledge

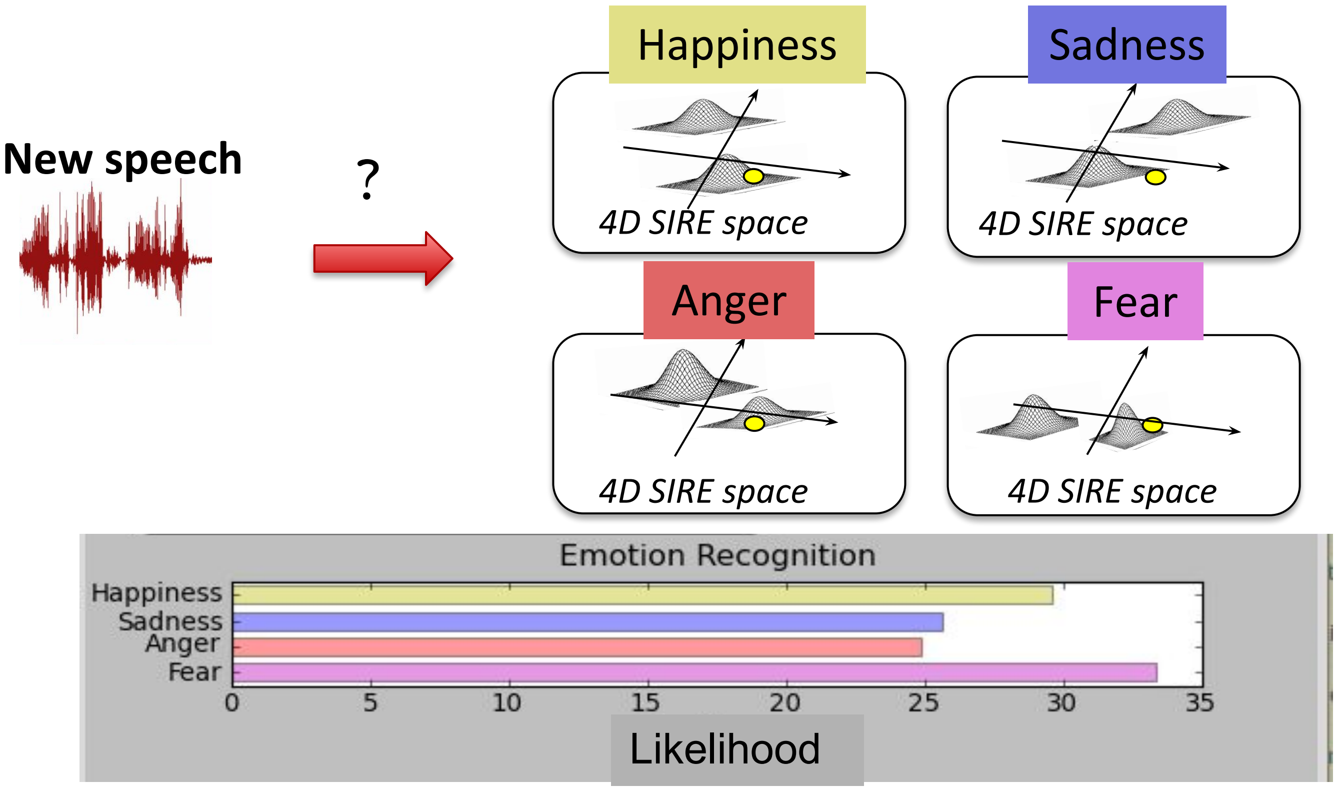
How do we understand the trained model? We can do this by inspecting the GMM means:

Happiness GMM

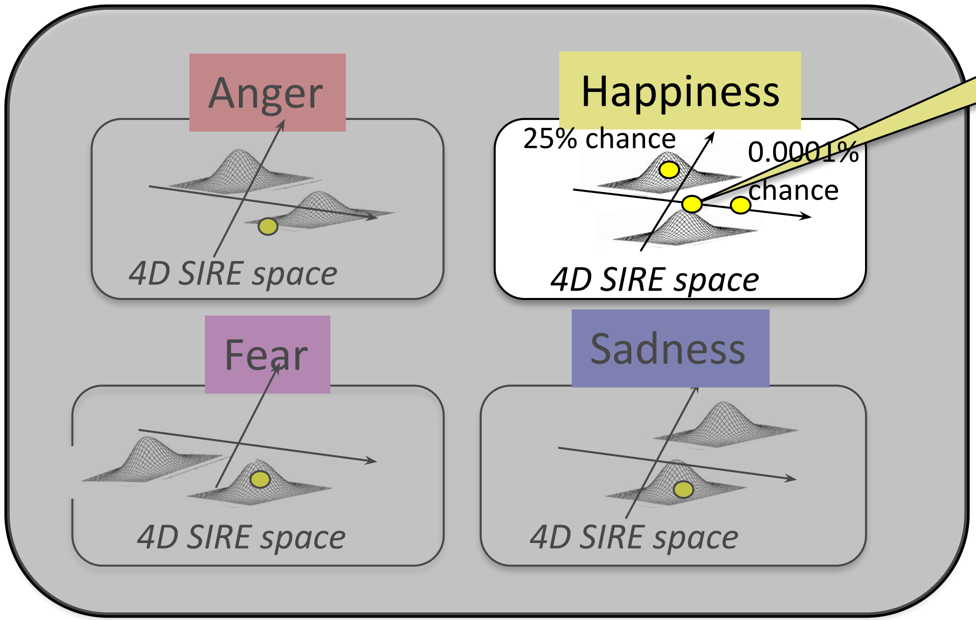


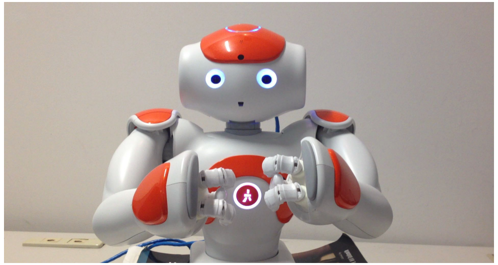


2. The GMM Recognizes



3. The GMM allows statistically probable expression





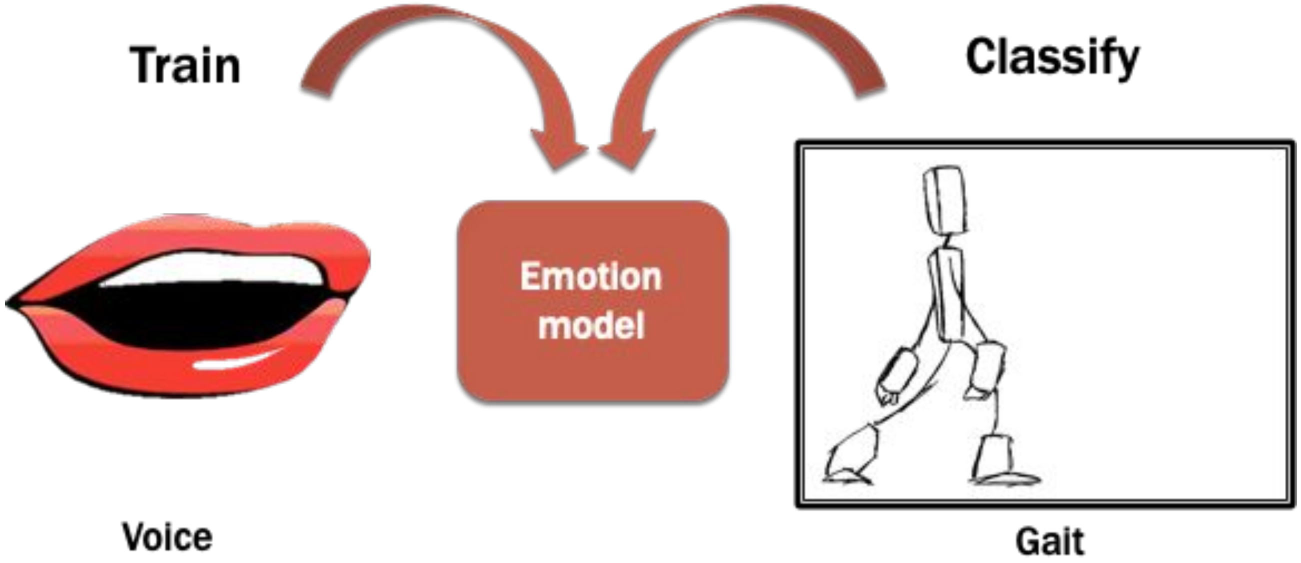
- Statistical expression based on model of observations
- Expression “rules” are implicit

Learning with SIRE

Cross-*domain* Generalization

Evgenia Obraztsova
Principal Dancer
The Bolshoi Ballet

Cross-modal Generalization



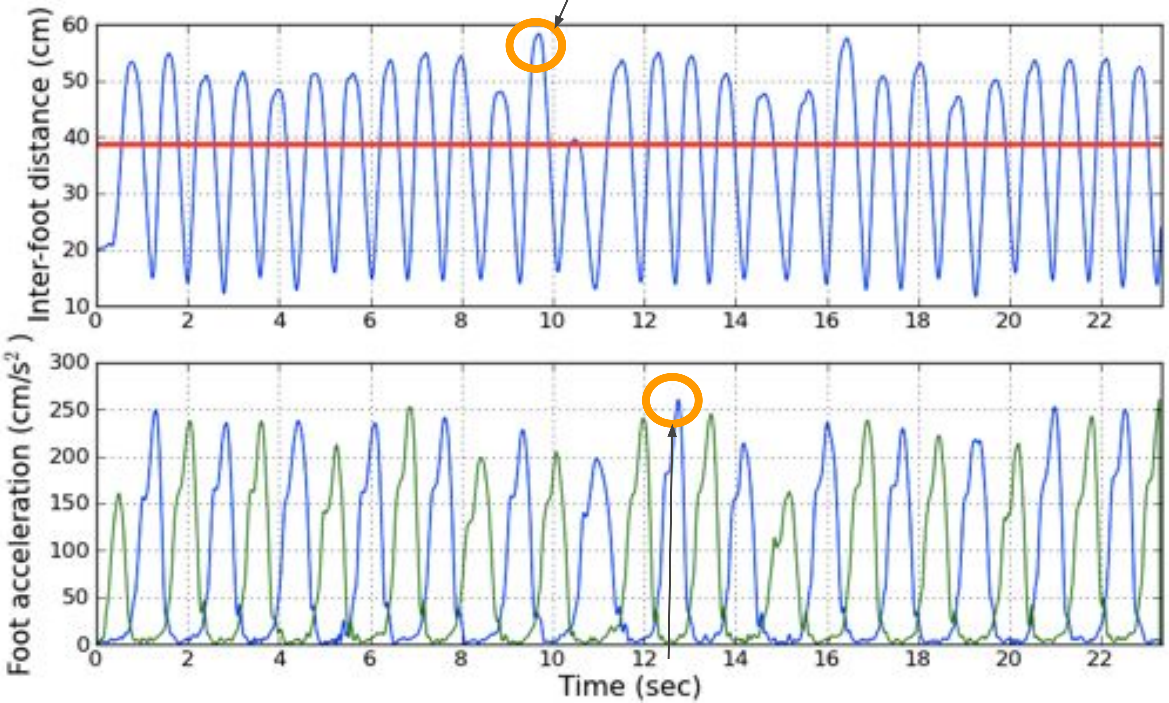
1. Feature Extraction

Voice feature	Parameter	Gait feature
Speech rate (syllables/sec)	Speed	Walking speed (steps/min)
Voice onset rapidity (dB/sec^2)	Intensity	Maximum foot acceleration (cm/sec^2)
Jitter (dB/sample)	irRegularity	Step timing variance (sec)
Pitch range (Hz)	Extent	Maximum step length (m)

Table 1: **Low-level feature to SIRE mappings**

Sad gait example

Extent: Maximum step length (x,y)



Intensity: Maximum acceleration in (x,y,z)

irRegularity: Standard deviation in step lengths

Speed: Average number of steps per minute

2. Mapping features to SIRE space

e.g. sad gait sample

Walking speed: 76 steps/min
Foot acceleration: 272 cm/sec^2
Step timing variance: 77 sec
Step length: 56 cm

↔

Speed = ?
Intensity = ?
irRegularity = ?
Extent = ?

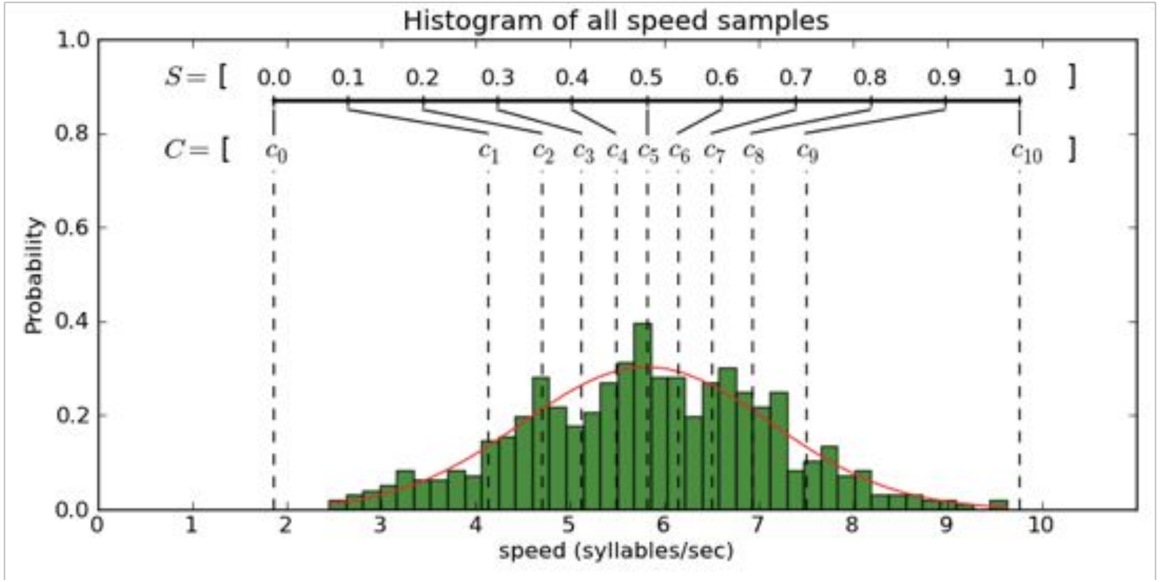
- How do we map our samples to [0,1] SIRE space?

Feature	μ	σ
Walking speed (steps/min)	91.75	16.76
Maximum foot acceleration (cm/sec^2)	341.22	68.88
Step timing variance (sec)	0.07	0.06
Maximum step length (cm)	63.21	8.08

2. Mapping to [0,1] SIRE space

- Assume single Gaussian distribution of samples
- Find mapping array $C(k), k = 0,\ldots 9$

$$0.1 = cdf(x_{k+1}) - cdf(x_k)$$



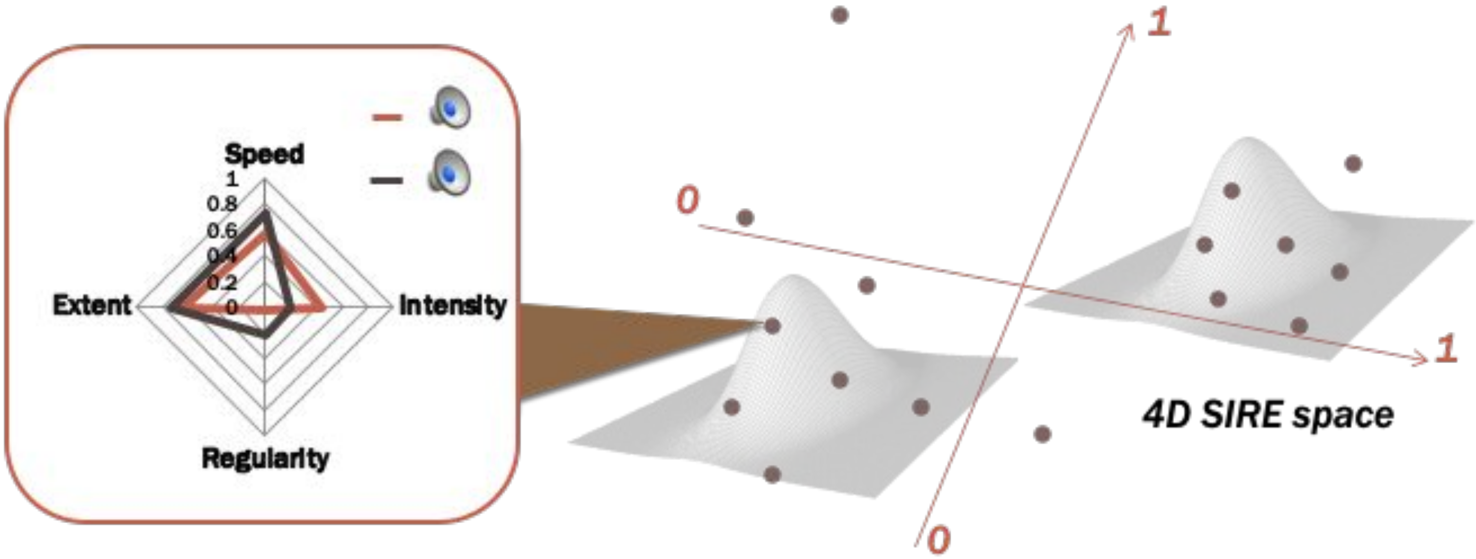
3. Personalization

- **Idea:** We should take into account the difference in, for example, step length of a tall person vs. short person.
- **Approach:** For each subject P's emotion samples, add a personalized bias (difference between subject's average values and group's average)

$$P_{f,k} = P_{f,k} + b(P_f)$$

4. Training GMM

Train Gaussian Mixture Model using Expectation Maximization on our emotional sample set



Learning with SIRE

Experiments and Results

Research Questions

1. What are the real-world values defining emotion in speech and gait?
2. What are the SIRE values defining emotions in speech and gait, and are they similar?
3. What is the effect of using SIRE mapping and personalization on emotion training and recognition?
4. Can an emotion classifier be trained with one modality and tested with another?

Experiments

Materials

Databases containing:

- Happiness
- Sadness
- Anger
- Fear
- Neutral

Procedure

- Sci-kit-learn toolkit
- 5-component Gaussian Mixture Model EM
- 10-fold cross validation

Berlin Emotional Speech Database

- Wave files
- 10 subjects
- Up to 10 sentences per emotion
- Total: 408 voice samples

Body Movement Library

- Feet position data
- 28 subjects
- Up to 2 samples per subject, per emotion
- Total: 236 gait samples

Results

1. What are the average real-world values defining emotion in speech and gait?

Feature	Speech rate (syll/sec)	Voice onset rapidity (dB/sample^2)	Jitter (dB/sample)	Pitch range (Hz)
Happiness	6.1	13.0	871	144
Sadness	4.3	8.5	724	101
Anger	6.0	13.7	964	131
Fear	7	10.8	1025	105
Neutral	6.4	10.3	754	82

Feature	Walking speed (steps/min)	Acceleration (cm/s^2)	Variance (ms)	Step length (cm)
Happiness	96	362	64	65
Sadness	76	272	77	56
Anger	105	411	63	71
Fear	92	324	78	62
Neutral	90	323	58	61

Results

Means when forcing a 1-component GMM

2.What are the SIRE values defining emotions in speech and gait, and are they similar? (differences > 15% highlighted)

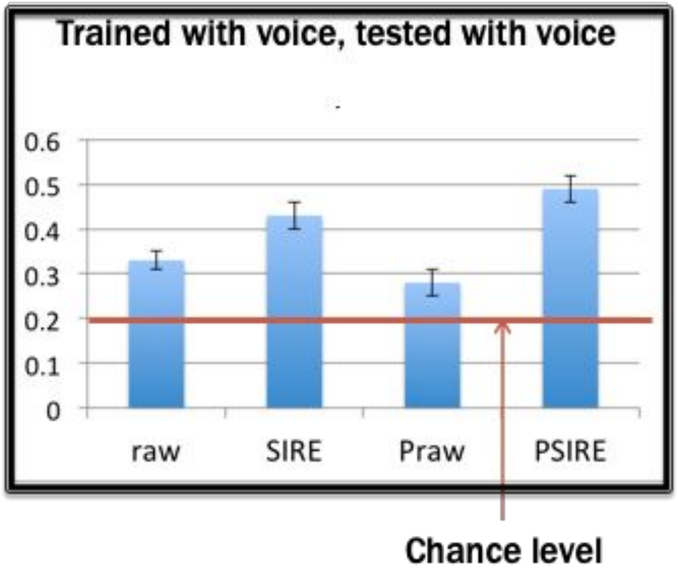
Voice	S	I	R	E
Happiness	0.59	0.63	0.49	0.74
Sadness	0.13	0.27	**0.29**	**0.40**
Anger	**0.56**	0.68	0.62	0.65
Fear	**0.81**	0.45	0.70	0.43
Neutral	0.66	0.41	0.34	0.25

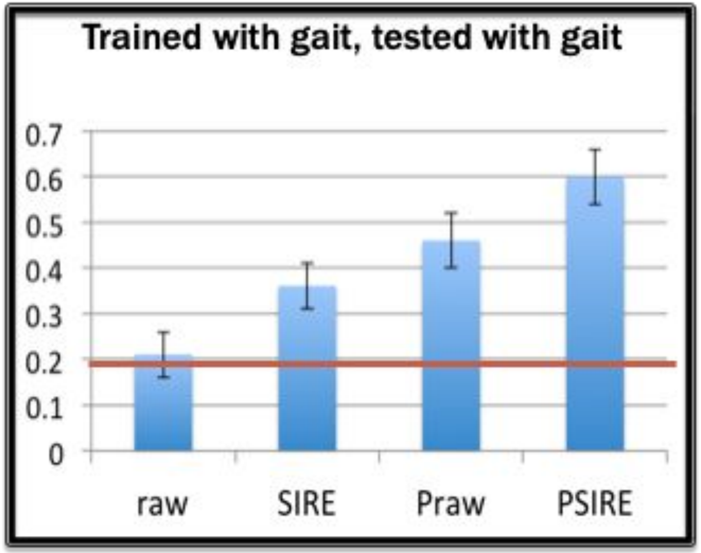
Gait	S	I	R	E
Happiness	0.60	0.61	0.49	0.64
Sadness	0.18	0.16	**0.58**	**0.19**
Anger	**0.78**	0.84	0.48	0.83
Fear	**0.51**	0.41	0.58	0.39
Neutral	0.46	0.41	0.44	0.39

Sadness: anguish vs. depressed **Anger**: hot vs. cold **Fear**: depends on source of fear

Results

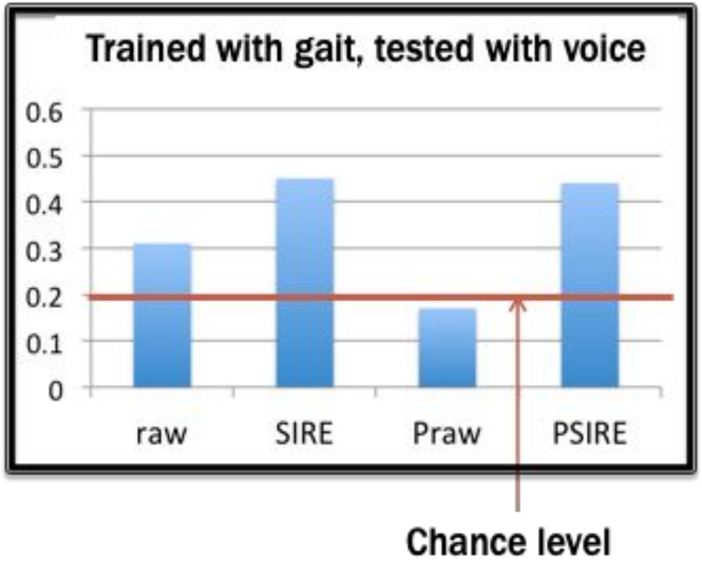
3.What is the effect of using SIRE mapping and personalization on emotion training and recognition?

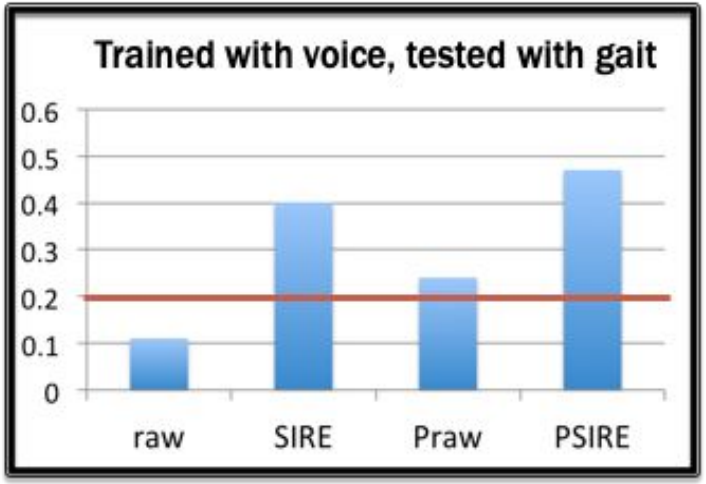


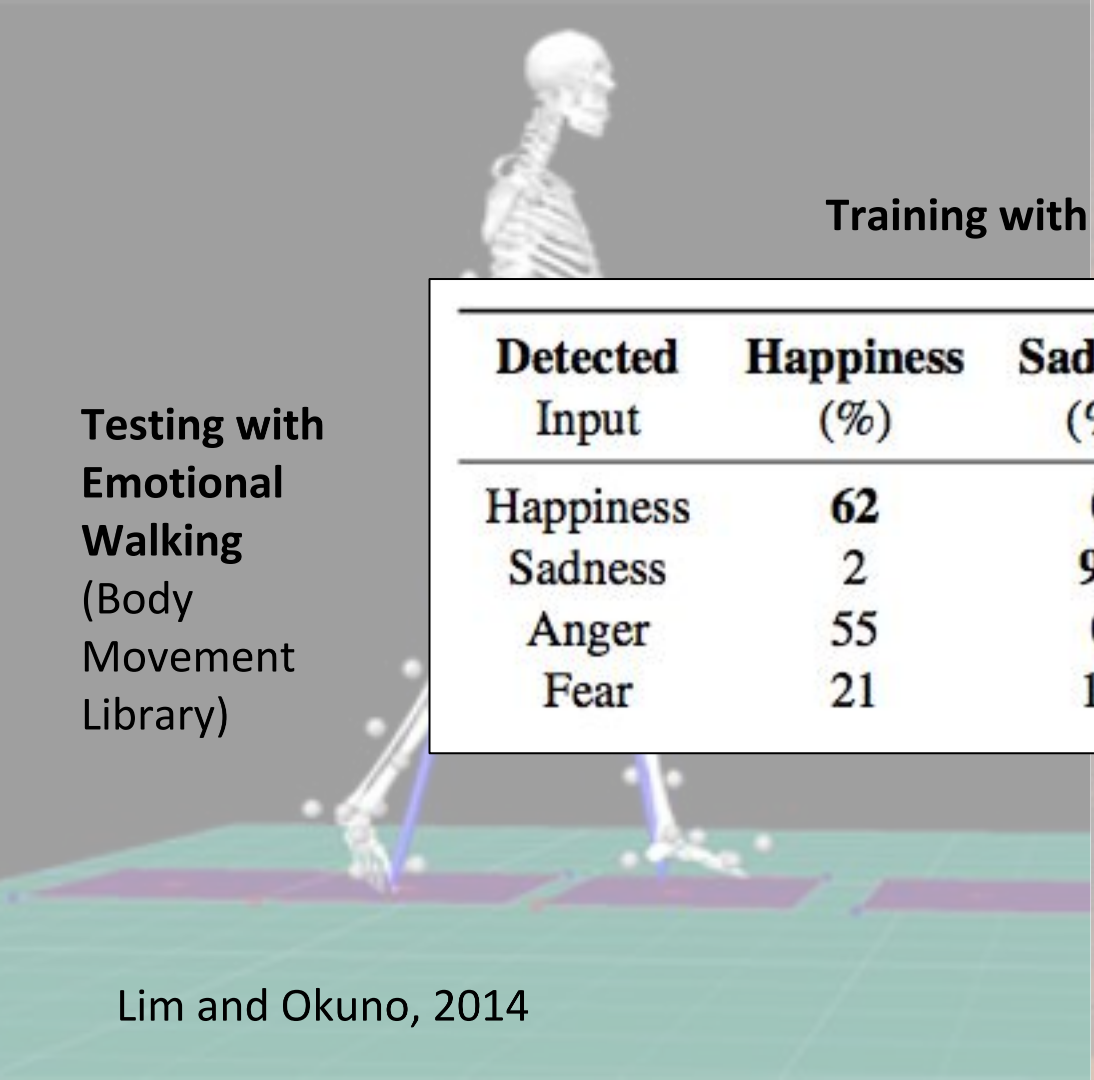


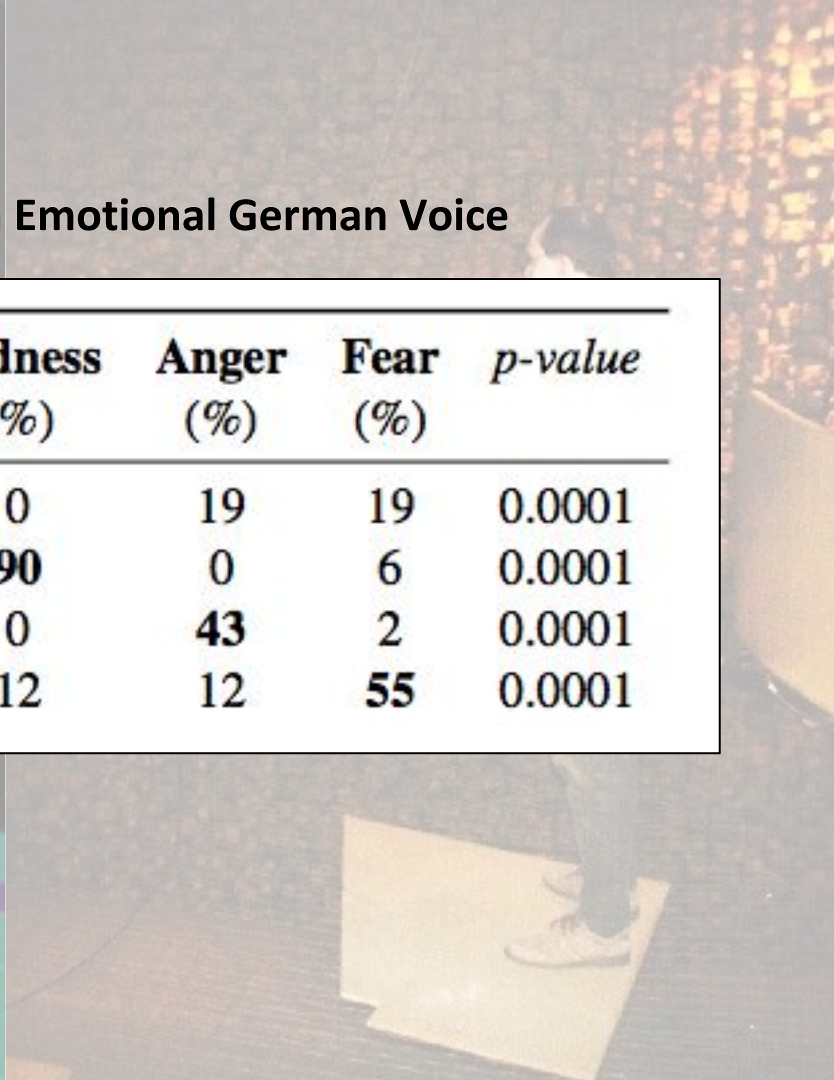
Results

4.Can an emotion classifier be trained with one modality and tested with another?









Training with Emotional German Voice

Testing with Emotional Walking
(Body Movement Library)

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

Lim and Okuno, 2014