Recognizing Human Emotional State from Audiovisual Signals

Yongjin Wang, Yun Tie and Ling Guan

Department of Electrical and Computer Engineering
Ryerson University

Outline

- Background
- Audio Feature Representation
- Visual Feature Representation
- Recognition System
- Conclusions and Future Research
Background

What are Emotions?

– “cognitive” like thoughts?
– “physical” like the pounding of a heart?
– or some kind of combination?
Discoveries in Psychology

- Emotions are processed by a circuit of interconnected brain structures known as the limbic system.
- Basic emotions constitute the foundations of human emotion.
- Certain emotions are served by separate brain systems, and common to cultures throughout the world.
- Emotion space (Ekman's six): anger, disgust, fear, happiness, sadness, and surprise.

Limbic System Structure

- Cingulate gyrus
- Septum
- Olfactory bulb
- Hypothalamus
- Amygdala
- Mammillary body
- Hippocampus
- Fornix
**Indicators of Emotion**

- Speech
- Facial expression
- Body language: highly dependent on personality, gender, age, culture, etc
- Semantic meaning: two sentences could have the same lexical meaning but different emotional information

... ...

**Motivation**

- Some emotions are audio dominant, while others are visual dominant
- Complementary relationship of two modalities help to achieve higher recognition accuracy
- A generic recognition system should be independent of language and cultural background of the speaker
Objectives

- To develop an efficient language and cultural background independent system for recognition of human emotional state from audiovisual signals
- To enhance the Human-Computer Intelligent Interaction

Applications

- Human–Computer Interaction
- Learning Environments
- Consumer Relations
- Entertainment
- Educational Software
- Computer Animation
- Call Centers
- Smart Living
- Security/surveillance
- 3D Immersive Communication
RML Emotion Database

- **Eight subjects**: 2 Italian, 2 Chinese, 2 Pakistani, 1 Persian, and 1 Canadian
- **Six languages**: Italian, Mandarin, Urdu, Punjabi, Persian, and English
- **Six Emotions**: Anger, Disgust, Fear, Happiness, Sadness, and Surprise
- 500 video samples, with a sampling rate of 22050 Hz, using single channel 16-bit digitization

Reference Sentences

- More than ten reference sentences for each emotion
- Examples of the used sentences:
  - **Happiness**: I got A+ in that course.
  - **Anger**: Why do you always cheat on me.
  - **Sadness**: I messed up the mid-term.
  - **Surprise**: What? Am I the winner?
  - **Fear**: Please, please do not hurt me.
  - **Disgust**: That rotten fish has a disgusting smell.
Snapshots of Facial Expressions

ANGER  DISGUST  FEAR

HAPPINESS  SADNESS  SURPRISE

Examples of Audio and Video Samples

Sadness  Happiness  Disgust

Fear  Surprise  Anger

Disgust  Chinese  English  Italian  Persian  Punjabi  Urdu
Audio Feature Representation

Audio Feature Extraction System

Input Speech → Preprocessing → Windowing → MFCC Features → Audio Features

Prosodic Features

Formant Frequency
Preprocessing

• **Purpose:** reduce the effects of noise and silence which do not contribute to human emotion
  - **Noise reduction:** wavelet coefficients thresholding
  - **Leading and trailing edge elimination:** noise amplitude thresholding

Preprocessing Flow

Original Audio Signal

After Noise Reduction

After Preprocessing

Noise Reduction

Leading and Trailing Edge Elimination
Windowing

- Spectral analysis is only reliable when the signal is stationary
- **Purpose:** segment speech signal into short-time frames to approximate stationary signal
- **Hamming window:** low leakage effect

\[
w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad n = 0, \ldots, N-1
\]

where \( w(n) \) is the impulse response of the window, \( N \) is the size of the window

---

Hamming Window

![Hamming Window Diagram](image-url)
Prosodic Features

Prosody: rhythmic aspects of speech
- control of pitch
- segmental duration (rhythm)
- intensity

Spectral Analysis - Spectrogram

Anger
Disgust
Fear
Happiness
Sadness
Surprise

English  Italian  Persian  Urdu  Punjabi  Chinese
Vocal Cues of Emotion

<table>
<thead>
<tr>
<th>Emotional State</th>
<th>Cues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>many harmonics, fast tempo, high pitch level</td>
</tr>
<tr>
<td>Disgust</td>
<td>many harmonics, slow tempo, small pitch variation</td>
</tr>
<tr>
<td>Fear</td>
<td>pitch contour up, fast tempo, high pitch level</td>
</tr>
<tr>
<td>Happiness</td>
<td>few harmonics, fast tempo, large pitch variation</td>
</tr>
<tr>
<td>Sadness</td>
<td>pitch contour down, slow tempo, low pitch level</td>
</tr>
<tr>
<td>Surprise</td>
<td>few harmonics, fast tempo, high pitch level</td>
</tr>
</tbody>
</table>

List of Extracted Prosodic Features

<table>
<thead>
<tr>
<th>Index</th>
<th>Feature description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-5</td>
<td>Pitch mean, median, standard deviation, max, and range</td>
</tr>
<tr>
<td>6</td>
<td>Pitch variation rate</td>
</tr>
<tr>
<td>7</td>
<td>Rising/falling ratio</td>
</tr>
<tr>
<td>8</td>
<td>Rising pitch slope max</td>
</tr>
<tr>
<td>9</td>
<td>Falling pitch slope max</td>
</tr>
<tr>
<td>10</td>
<td>Rising pitch slope mean</td>
</tr>
<tr>
<td>11</td>
<td>Falling pitch slope mean</td>
</tr>
<tr>
<td>12</td>
<td>Pitch rising range max</td>
</tr>
<tr>
<td>13</td>
<td>Pitch falling range max</td>
</tr>
<tr>
<td>14</td>
<td>Pitch rising range mean</td>
</tr>
<tr>
<td>15</td>
<td>Pitch falling range mean</td>
</tr>
<tr>
<td>16-18</td>
<td>Overall pitch slope mean, standard deviation, median</td>
</tr>
<tr>
<td>19-23</td>
<td>Energy mean, median, standard deviation, max, and range (dB)</td>
</tr>
<tr>
<td>24</td>
<td>Average pause length</td>
</tr>
<tr>
<td>25</td>
<td>Speaking rate</td>
</tr>
</tbody>
</table>
Prosodic Feature Extraction

Pitch estimation: Top: waveform in time domain, Middle: magnitude spectrum, Bottom: DFT of log magnitude spectrum

- Pitch is estimated based on the Fourier analysis of the log amplitude spectrum of the signal
- Energy features are computed in time domain
- Speaking rate is approximated by:

$$ R_{spk} = \frac{1}{\text{mean} \_ \text{segment} \_ \text{length}} = \frac{N}{\sum_i T_i} $$

where $T_i$ is the length of voiced segment $i$ and $N$ is the number of voiced segments

Mel-frequency Cepstral Coefficients (MFCC)

- Popular feature in speech recognition, speaker verification, and identification
- Mimic the behavior of human ears by applying cepstral analysis
**MFCC Processing Flow**

- **Speech Signal** $x(n)$
- **Preprocessing** $x'(n)$
- **Window** $x_i(n)$
- **DFT** $X_i(k)$
- **Mel Filter Banks** $Y_t(m)$
- **IDCT** $y_t^m(k)$
- **MFCC** $y_t^m(k)$

**Ryerson Multimedia Research Laboratory (RML)**

---

**Mel-scaled Filter bank**

- Designed to capture and emphasize the information in low frequency band
- Constructed by triangular-shaped filters with frequency overlap
- The central frequency of each Mel filter is uniformly spaced before 1 kHz and it follows a logarithmic scale after 1 kHz
- The filter magnitude is set to 1 before 1 kHz, while decreasing logarithmically as the frequency increases

**Ryerson Multimedia Research Laboratory (RML)**

---
Mel-scaled Filter Bank Design

MFCC Computation

- MFCCs are the inverse discrete cosine transform of the logarithm of the magnitude of the filter bank output
- Magnitude and logarithm processing are also performed by the human ear.
MFCC Feature Mapping

- Common practice: use the 1st 9 to 13 coefficients
  --- Most of the signal energy is compacted in the first few coefficients due to the properties of the cosine transformation
- **Purpose:** make the features of each utterance uniform, and thus facilitate classification
- 65 MFCC features (use 13 coefficients): mean, median, standard deviation, max and min of the first thirteen coefficients

Summary of Audio Features

<table>
<thead>
<tr>
<th>Features</th>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosodic</td>
<td>25</td>
<td>[p(i), i=1,\ldots,25]</td>
</tr>
<tr>
<td>MFCC</td>
<td>65</td>
<td>[\text{mfcc}(j,\text{mean}),\text{mfcc}(j,\text{median}),\text{mfcc}(j,\text{std}),\text{mfcc}(j,\max),\text{mfcc}(j,\min), j=1,\ldots,13]</td>
</tr>
<tr>
<td>Formant</td>
<td>15</td>
<td>[\text{ff}(k,\text{mean}),\text{ff}(k,\text{median}),\text{ff}(k,\text{std}),\text{ff}(k,\max),\text{ff}(k,\min), j=1,2,3]</td>
</tr>
<tr>
<td>Overall Audio</td>
<td>105</td>
<td>[p(1),\ldots,p(25),\text{mfcc}(1,\text{mean}),\ldots,\text{mfcc}(13,\min),\text{ff}(1,\text{mean}),\ldots,\text{ff}(3,\min)]</td>
</tr>
</tbody>
</table>
Visual Feature Representation

Visual Feature Extraction System

- Input Image Sequence
- Key Frame Extraction
- Maximum Speech Amplitude
- Face Detection
- Visual Features
- Feature Mapping
- Gabor Filter Bank
**HSV Color Model**

- Planar envelop approximation method in HSV color space:
  \[ S \geq T_h, \quad V = T_v, \quad S \leq -H - 0.1V + 110; H \leq -0.4V + 75; \]
  \[ \text{If } H \geq 0, \quad S \leq 0.08(100-V) H + 0.5V \quad \text{Else} \quad S \leq 0.5H + 35 \]

- **Clearing procedure**: morphological operation
  - Closing: connect narrow gaps
  - Opening: remove small isolated bulbs
  - Filling: remove black isolated holes

- **Normalization**: gray-level image of size 128×128
**Face Detection Flow**

Key Frame  →  Binary  →  Opening

Normalized Face Image  ←  Filling  ←  Closing

---

**Gabor Wavelet Representation**

**Multi-resolution analysis:**
Allows description of spatial frequency structure while preserving information about spatial relations

**Space domain:**

\[ g(x, y) = s(x, y) \times w(x, y) \]

- \( g(x, y) \): Gabor function
- \( s(x, y) \): Complex sinusoidal
- \( w(x, y) \): Gaussian kernel

**Gabor filter dictionary design:**
4 scales, and 6 orientations

**Feature mapping:**
Mean and standard deviation of the output of each filter (48 visual features)
Gabor Filter Dictionary Design

Half-amplitude contour of the designed Gabor filter dictionary

Gabor Filter Bank

Space domain

Frequency domain
Gabor Filter Bank

Example of Gabor wavelet transformed image

Recognition System
Overview of Proposed System

Input video
- Key frame extraction
- Face detection
- Gabor wavelet

Audio feature extraction
- Pre-processing
- Windowing
- MFCC
- Formant

Visual feature extraction

Classification scheme

Experimental Data Setup

- Experiments were performed on 500 video samples (RML Emotion Database)
- Eight subjects, speaking six languages
- Six emotion labels: Anger, Disgust, Fear, Happiness, Sadness, and Surprise
- 360 samples (from six subjects) were used for training, and the rest 140 (from the remaining two subjects) for testing, there is no overlap between training and testing subjects
Classification Algorithms

- Which classifier can best model our problem?
- Comparison of different classification algorithms:

  - Gaussian Mixture Model – Parametric method
  - K-nearest Neighbors – Non-parametric method
  - Neural Networks – Nonlinear method
  - Fisher’s Linear Discriminant Analysis – Linear method

Gaussian Mixture Model (GMM)

- GMM models the probability density function (pdf) of the data $P_{GMM}(y)$ as weighted sum of several different Gaussian density functions $p_k(y)$:

$$P_{GMM}(y) = \sum_{k=1}^{K} c_k p_k(y)$$

where $c_k$ are the component probabilities, and $K$ is the number of components
**Expectation Maximization (EM)**

- Parameter estimation: EM algorithm

\[
c_k^{(i+1)} = \frac{1}{N} \sum_{n=1}^{N} \psi_k(n) \\
\mu_k^{(i+1)} = \frac{\sum_{n=1}^{N} \psi_k(n) y_n}{\sum_{n=1}^{N} \psi_k(n)} \\
\sigma_k^{(i+1)} = \frac{\sum_{n=1}^{N} \psi_k(n)(y_n - \mu_k^{(i)})^2}{\sum_{n=1}^{N} \psi_k(n)}
\]

where \( \psi_k(n) \) is the posteriori probability, \( N \) is the size of the feature vector, and \( i \) is the iteration number.

**Architecture of applied GMM scheme**

INPUT

- ANGER
- DISGUST
- FEAR
- HAPPINESS
- SADNESS
- SURPRISE

GMM

Y = argmax(Y_i)

DECISION

Recognized Emotion
K-nearest Neighbors (KNN)

- Non-parametric method for classification
- Three ingredients:
  - a positive integer, \( k \)
  - a metric (or pattern similarity function) \( D(x, x') \)
  - a reference sample of \( n \) correctly labeled feature vectors:
    \[ S_n = \{(x^1, l^1), (x^2, l^2), \ldots, (x^n, l^n)\} \]

Algorithm: Given an input feature vector \( x \), identify the \( k \) nearest neighbors: the subset of \( k \) feature vectors from \( S_n \) that lie closest to \( x \) with respect to the metric equals the class that occurs with greatest frequency within the subset of \( k \) nearest neighbors.

- Leave-one-out cross validation can be used to determine the proper \( k \) value

Neural Networks (NN)

- Three-layer feed-forward architecture
- Back-propagation for training
- Number of input neurons is equal to the dimension of the feature vector
- Six output neurons correspond to the six emotional classes
- Number of neurons in the hidden hyperbolic sigmoid layer is equal to the square root of the product of input and output neurons
Fisher’s Linear Discriminant Analysis (FLDA)

- LDA assumes the discriminant function $g(x)$ to be a linear function of data $x$
  $$g_i(x) = w_i^T x + w_{i0}$$
- In FLDA, the weights $w$ are estimated by maximizing Fisher’s criterion function $J_F(w)$
  $$J_F(w) = \frac{w^T S_b w}{w^T S_w w} = w^T S_b S_w^{-1} w$$
  where $S_b$ and $S_w$ are the between-class and within-class scatter matrix respectively
### Comparison of Classification Algorithms

#### Audiovisual Recognition

<table>
<thead>
<tr>
<th></th>
<th>GMM</th>
<th>NN</th>
<th>K-NN</th>
<th>FLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosodic</td>
<td>60.60</td>
<td>49.29</td>
<td>55.00</td>
<td>65.71</td>
</tr>
<tr>
<td>Audio</td>
<td>64.72</td>
<td>51.43</td>
<td>62.14</td>
<td>66.43</td>
</tr>
<tr>
<td>Visual</td>
<td>28.28</td>
<td>35.00</td>
<td>35.71</td>
<td>49.29</td>
</tr>
<tr>
<td>Audiovisual</td>
<td>65.38</td>
<td>56.43</td>
<td>62.86</td>
<td>70.00</td>
</tr>
</tbody>
</table>

#### Audiovisual Recognition of Emotions

<table>
<thead>
<tr>
<th></th>
<th>Anger</th>
<th>Disgust</th>
<th>Fear</th>
<th>Happiness</th>
<th>Sadness</th>
<th>Surprise</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>88.46</td>
<td>61.90</td>
<td>59.09</td>
<td>48.00</td>
<td>57.14</td>
<td>80.00</td>
<td>66.43</td>
</tr>
<tr>
<td>Visual</td>
<td>34.62</td>
<td>47.62</td>
<td>88.18</td>
<td>52.00</td>
<td>47.62</td>
<td>48.00</td>
<td>49.29</td>
</tr>
<tr>
<td>Audiovisual</td>
<td>76.92</td>
<td>76.19</td>
<td>77.27</td>
<td>56.00</td>
<td>61.90</td>
<td>72.00</td>
<td>70.00</td>
</tr>
</tbody>
</table>
Discussion

- FLDA outperforms GMM, KNN, and NN
- Audiovisual performs better than audio or visual only
- Prosodic features are very powerful indicator of emotion in speech

Discussion (cont.)

- However, in some of the emotions, such as anger and surprise, the combination of audio and visual information performs worse than audio only
- High dimensionality → high computational cost

Solution: Feature selection?
Stepwise Method

- Combination of forward and backward procedures,
- Criterion: Mahalanobis distance
- For each step, one feature is added to or removed from the selected feature subset to maximize the between-class Mahalanobis distance

Results with Feature Selection

<table>
<thead>
<tr>
<th></th>
<th>Dimension</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original features</td>
<td>153</td>
<td>70.00%</td>
</tr>
<tr>
<td>Stepwise method</td>
<td>45</td>
<td>75.71%</td>
</tr>
</tbody>
</table>

- Stepwise method achieves both low dimensionality and high accuracy
**Comparison of Results**

- **Audio**: 88.46, 61.90, 59.09, 48.00, 57.14, 80.00, 66.43
- **Visual**: 34.62, 47.62, 68.18, 52.00, 47.62, 48.00, 49.29
- **Audiovisual**: 76.92, 76.19, 77.27, 56.00, 61.90, 72.00, 70.00
- **Stepwise**: 80.77, 61.90, 77.27, 80.00, 76.19, 76.00, 75.71

**Discussion**

- In Disgust, the performance is actually lower than before feature selection.
- In anger and surprise, the performance is improved, but still lower than audio only.
  
  **Why**
  - Stepwise method is a suboptimal method.
  - Feature selection is performed on a six-class basis.

  **Solution**: Individual class based analysis.
Multi-classifier Scheme

- Six one-against-all (OAA) classifiers, the output of which is the probability of belonging to the corresponding emotion
- Corresponding classifiers: 15 two-class, 20 three-class, 15 four-class, 6 five-class, and 1 six-class (global) classifiers
- Feature selection is performed on all the classifiers by using stepwise method

Multi-classifier Scheme (cont.)

Rule 1:
If N >= 2 or N=0,
→ global classifier
Accuracy: 79.29% overall

Rule 2:
If N >= 2,
→ corresponding classifiers
Else if N=0,
→ global classifier
Accuracy: 82.14% overall

Where N is the number of output that exceeds 50%
Overall Experimental Results

<table>
<thead>
<tr>
<th></th>
<th>Anger</th>
<th>Disgust</th>
<th>Fear</th>
<th>Happiness</th>
<th>Sadness</th>
<th>Surprise</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>88.46</td>
<td>61.90</td>
<td>59.09</td>
<td>48.00</td>
<td>57.14</td>
<td>80.00</td>
<td>66.43</td>
</tr>
<tr>
<td>Visual</td>
<td>34.62</td>
<td>47.62</td>
<td>68.18</td>
<td>52.00</td>
<td>47.62</td>
<td>48.00</td>
<td>49.29</td>
</tr>
<tr>
<td>Audiovisual</td>
<td>76.92</td>
<td>76.19</td>
<td>77.27</td>
<td>56.00</td>
<td>61.90</td>
<td>72.00</td>
<td>70.00</td>
</tr>
<tr>
<td>Stepwise</td>
<td>80.77</td>
<td>61.90</td>
<td>77.27</td>
<td>80.00</td>
<td>76.19</td>
<td>76.00</td>
<td>76.71</td>
</tr>
<tr>
<td>Multi-classifier</td>
<td>88.46</td>
<td>80.95</td>
<td>77.27</td>
<td>80.00</td>
<td>80.95</td>
<td>84.00</td>
<td>82.14</td>
</tr>
</tbody>
</table>

Discussion

- Dynamic visual information is missing.
- Prosodic features are very powerful indicator of emotion in speech
Acknowledgement

• This work is supported in part by the Canada Research Chair program and Canada Foundation for Innovation (CFI)
• Special thanks are due to the member of Ryerson Multimedia Research Laboratory (RML) for their help in collecting the emotional video database

References

Roadmap 2

- No class on Tue, March 29.
- Final Test: Tue, April 5, 1 hour, open/close notes?
  - Coverage of material: everything studied in the course.
  - Office hour: Tue, Apr 5, 10:00-11:45am.
- Project presentation:
  - Tue, Apr 12, 12:00pm in the classroom, VIC110 (tentative)
  - 20 minutes for individual and 25 minutes for team of two for presentation and a short demo.
- Report:
  - Up to 15 pages single space or up to 25 pages double space.
  - Due Tue, Apr 26, 5pm.

Course Instructor:
Prof. Ling Guan
Department of Electrical & Computer Engineering
Room 315, ENG Building
Tel: (416) 979-5000 ext 6072
Email: lguan@ee.ryerson.ca

Participating Instructor:
Prof. Yifeng He
Multimedia Communications
(Tuesday, March 22)

Instructor: Prof. Yifeng He
Department of Electrical & Computer Engineering
Room 326, ENG Building
Tel: (416)979-5000 ext 4904
Email: yhe@ee.ryerson.ca