# Speech-based Emotion Recognition

陳嘉平

Multimedia Information Technology Lab
Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan

October 2012

*"If you talk to a man in a language he understands, that goes to his head. If you talk to him in his language, that goes to his heart."*

– Nelson Mandela

*"If you talk to a man in a language he understands, that goes to his head. If you talk to him in his language, that goes to his heart."*

– Nelson Mandela

talk = speech

*"If you talk to a man in a language he understands, that goes to his head. If you talk to him in his language, that goes to his heart."*

– Nelson Mandela

$$\text{talk} = \text{speech}$$

$$\text{heart} = \text{emotion}$$

## Background

- speech
- emotion
- motivation

# Speech

## multitude of information in speech

1. what is spoken?
   $\rightarrow$ speech recognition
2. who is speaking?
   $\rightarrow$ speaker verification
3. which language is it?
   $\rightarrow$ language identification
4. how is it spoken?
   $\rightarrow$ emotion recognition

# Emotions

## happy

- mission accomplished
- when Jeremy Lin was running the Knicks

## angry

- when partner blames you for his blunder
- plane delayed, flight missed, and nobody's sorry about it

## fear

- in a car accident
- losing passport in Europe

## sad

- watch movie ``no country for old men''
- when the cafeteria you often eat at moves out of campus

## disgust

- rotten tomatoes
- fake stuff

## surprise

- winning a lottery
- secret guests

# Motivation

## who needs to recognize emotions?

- always hear and observe
- in a long-term relationship
  - boss and subordinate
  - parent and child
  - teacher and student
  - husband and wife
  - friends
- during a short-term relationship (brief encounter)
  - customers and waiters
  - strangers

## who wants to be emotional?

- strong emotional impacts lead to strong intellectual impacts
  $\rightarrow$ emotion for better learning
- emotional ups and downs happen at the critical times in life
  $\rightarrow$ emotion for better life
- memories, good or bad, remain for emotional experiences
  $\rightarrow$ they get sweeter as time goes by
- affects are the essence of human being (one who shows no emotions is difficult to be around with)
  $\rightarrow$ emotion for better social networking
- showing emotions releases pressures
  $\rightarrow$ emotion for better health, longer life
- being emotional is not the same as being irrational
  $\rightarrow$ it means touched, moved, engaged, etc.

## who can automatic emotion recognition help?

- those who want to but cannot recognize emotions
  - expressive agnosia: inability to perceive emotional expressions, e.g., Anton Chigurh in ``no country for old men''
  - machines
    - robot
    - servers
- those who are prone to be emotional
  - athletes
  - in-pregnancy
  - kids
  - silver-age
  - hospitalized

## Status Quo

- naïve definition of emotion states
- machine learning methodology
- data
- features
- models

# Emotional State

## continuous space

- valence (attitude)
- arousal (intensity)

## discrete states

- happy
- sad
- angry
- fear
- disgust
- surprise
- ... other ... mixed ...

# Recognition and Machine Learning

## background

Machine learning (a.k.a. data-driven) methodology is now familiar to the research community.

## emotion recognition via machine learning

- data collection
  $\rightarrow$ labeled or not labeled emotional speech
- feature design for data representation
  $\rightarrow$ informative, robust, etc.
- recognition model design
  $\rightarrow$ easy to learn, deploy, test, adapt, etc.
- performance evaluation and feedback

# Data, Features, and Methods

- emotional speech databases
  - number of emotional states
  - language
  - number of speakers
  - kind: natural/simulated/elicited
- acoustic features
  - pitch
  - formants
  - vocal-tract cross-section area
  - MFCC
  - TEO-based features
  - intensity
  - speaking rate
- classification methods
  - `HMM ANN LDA kNN SVM`

## Example (Dellaert et. al. 1996)

- 4 emotion categories
  - `happy sad anger fear`
- 1,000+ utterances with one emotion per utterance
- basic prosodic features
  - {mean, std, max, min, range} of pitch signal
  - global slope (of pitch) of linear regression
  - speaking rate
- basic classification methods
  - maximum-likelihood Bayes classifier
  - kernel regression
  - kNN

## Example (Schuller et. al. 2004)

- 6+1 emotion categories
  - `joy sad anger fear disgust surprise natural`
- 3,000+ utterances
- acoustic + linguistic features
  - phrase spotting
- classification methods
  - kMeans
  - kNN
  - GMM
  - MLP
  - SVM
  - belief network (l.f.)
  - fusion (a.f. + l.f.)

# Data Collection

## label issue

- **straightforward** to transcribe speech, which is local and objective
- **challenging** to label the emotion, which is highly contextual and somewhat subjective
  - ground truth
  - unlabelled data

## authenticity issue

- **easy** to collect speech data
- **difficult** to collect emotional speech data
  - acted data?

# Features

## detectable and indicative of emotion

- common features for ER

$$\{\text{rate, energy, pitch}\} \times \{\text{average, range, variation }\}$$

- features are inconclusive

$$\text{tears} = (\text{fears} \mid \text{sorrow} \mid \text{angry} \mid \text{happiness} \mid \text{dry eye})$$

## features for ASR vs. features for ER

- ASR: spectral, short-time analysis
- ER: prosodic, long-time analysis

# Recognition Models: ASR

## generative models of speech

- acoustic model, e.g. hidden Markov models (HMMs)
- language model, e.g. n-grams

## parameter estimation

expectation-maximization, count smoothing, parameter-tying ...

## search

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \; p(\mathbf{W}|\mathbf{O}) = \arg\max_{\mathbf{W}} \; p(\mathbf{W}) \; p(\mathbf{O}|\mathbf{W}).$$

- $A^*$ decoding
- dynamic programming + beam pruning

# Recognition Models: ER

## criteria

- plausibility
- feasibility
- scalability
- performance

## framework

- deep vs. shallow
- cognitive vs. responsive
- general vs. limited domain
- multi-model and fusion

## Future Works

- impacts
- application
- discussion
- conclusion

# Impacts

## impacts of emotion recognition

- on ASR
  $\rightarrow$ E.S.R.
- on spoken dialogue systems
  $\rightarrow$ interaction styles
- on voice search
  $\rightarrow$ negative/positive links
- on "orange technology"
  $\rightarrow$ barometer of personal emotion, slow-down of aging
- on education
  $\rightarrow$ affective computing for effective learning

# Killer Applications

## just some thoughts

- kids' talk
  → the conjecture is that kids are emotionally "pure" or "primitive"
- motherese
  → the conjecture is that motherese is consistent, at least from the baby's perspective
- e-Barometer
  → for people who are emotional
- entertainment industry (movie, TV, music . . . )
  → enormous and tremendous data
- You name it!

# Discussion

- Different people express emotions differently, depending on age, culture, gender, and personality.
- Emotion is not yet an accurate science. It is more of an engineering problem (application-oriented, as long as it works), rather than a discovery in science.
- There is not a single definition of emotion that works well for every application.

# Conclusion

- $e$ for emotion.
- Emotions are important. We rely on emotions.
- Using machines for emotions are uncharted seas.
- With lots of data, we can apply machine learning approaches to emotion recognition.
- The front of spoken language technology may be changing tack and moving towards emotion recognition.
- Emotion recognition impacts other areas of spoken language technology.
- Emotionally speaking, I hope you find emotions interesting. That's the most important thing of all.