Text Classification and Information Retrieval

Automatic Recognition of Emotions from Recorded Speech Data

• Final Project • Hanna M. Dettki • April 28, 2022

香 港 大 學
THE UNIVERSITY OF HONG KONG

Ruth Campbell. The processing of audio-visual speech. 2008

Automatic Speech Recognition (ASR) and NLP

- ASR incorporates many disciplines (e.g., linguistics, cognitive and computer science, engineering, human physiology, etc.)
- Speech perception and thus information retrieval on the receiver's side is **multimodal** (i.e.)

Emotions:

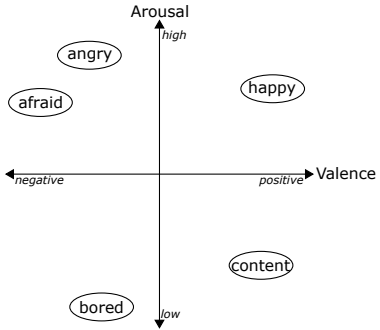

Figure: A two-dimensional emotion space with an Arousal and a Valence axis. Basic emotions are marked as ellipses within the quadrant.

Acoustic Properties in Speech:

Table: Some variations of acoustic variables observed in relation to emotions, from **?**.

Emotion	Pitch	Intensity	Speaking Rate	Voice Qualtiy
Anger	high mean, wide range	increased	increased	breathy; blaring timbre
Joy	increased mean and range	increased	increased	sometimes breathy; moderately blaring timbre
Sadness	normal or lower than normal mean, narrow range	decreased	slow	resonant timbre

Emotions:

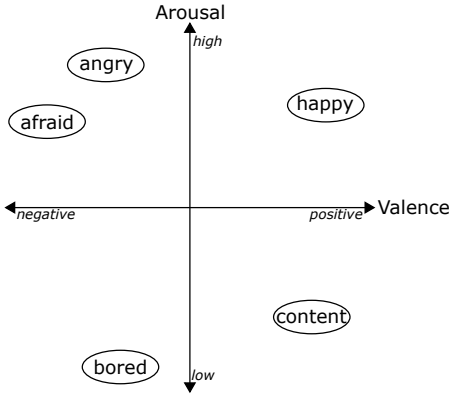

Figure: A two-dimensional emotion space with an Arousal and a Valence axis. Basic emotions are marked as ellipses within the quadrant.

Different Databases to classify emotions on:

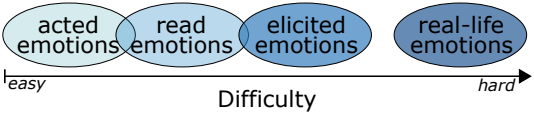

Figure: Types of databases used for emotion recognition and their difficulty.

Automatic Speech Emotion Recognition System:

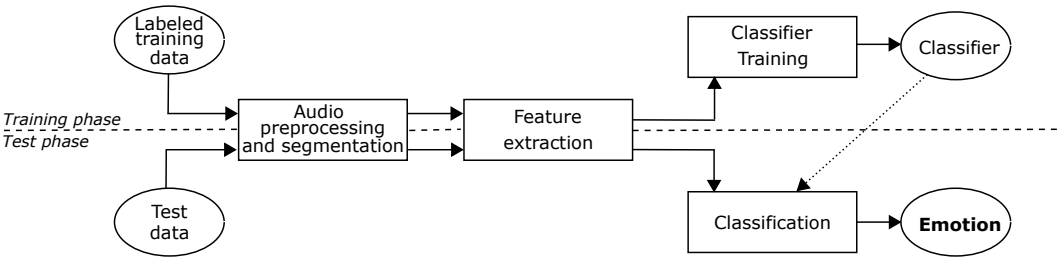

Figure: Overview of speech emotion recognition system.

Data Processing

How did you feel?

香 港 大 學
THE UNIVERSITY OF HONG KONG

Dataset: https://zenodo.org/record/1188976

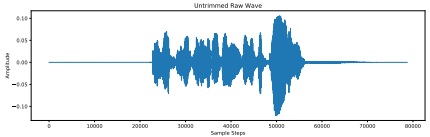

Figure: Un-trimmed sound wave of woman saying 'Dogs are sitting by the door.' in an fearful manner. There is still a lot of meaningless silence before and after the woman speaking.

Dataset

- *Ryerson Audio-Visual Database of Emotional Speech and Song* (RAVDESS)
- The RAVDESS database 'is a validated multimodal database of speech and song
- 24 professional actors and actresses, equally balanced in gender, were recorded phrasing main clauses such as 'Dogs are sitting by the door.'
- North American English
- in one of the following eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, surprised.
- 1440 RAVDESS files used (audio-only)

Data Preprocssing

Turning speech into numbers ...

香 港 大 學
THE UNIVERSITY OF HONG KONG

Audio Preprocessing ...

Speech Processing:

Classification Algorithms

(1) Convolutional Neural Network

- CNNs are a special kind of neural network
- inspired by the concept of the mammalian retina
- have proven to perform well on high-dimensional input
- use convolution as a specialized kind of linear operation in place of a general matrix multiplication
- Advantageous to CNNs is their high efficiency in terms of computational complexity while using a sparse set of parameters
- The kernels which the CNN learns during training process, are reused over the entire input which markedly exceeds regular feed forward networks in terms of memory and computational efficiency.

Support Vector machine:

- **Goal:** find a hyperplane of the form $\mathcal{H} = \{x \in \mathbb{R}^d | \langle w, x \rangle + b = 0\}$ such that it separates the data while maximizing the distance between the hyperplane and the closest data point.
- This distance is referred to as the margin.
- By maximizing the margin, the classifier is most robust to noise in new data

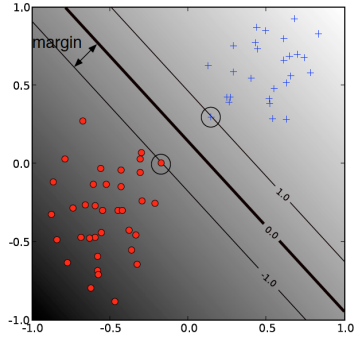

Figure: A linear SVM. The circled data points are the *support vectors* – i.e., the examples that are closest to the decision boundary. The support vectors determine the the margin with which the two classes are separated. In this work, eight classes need to be separated.

香 港 大 學
THE UNIVERSITY OF HONG KONG

SVM-Model Parameters:

$l = 745$ (number of principal components of PCA)
$\gamma_{\text{PCA}} = 1$ (inversed kernel width of PCA)
$C = 10$ (soft-margin constant of SVM)
$\gamma_{\text{SVM}} = 10$ (inversed kernel width of SVM)

Comparison

SVM	Training	Test
Accuracy	97.05%	63.39%

Table: Results of evaluating the SVM-model on the training and test data.

→

Summary

香 港 大 學
THE UNIVERSITY OF HONG KONG

SVM:

SVM	Training	Test
Accuracy	97.05%	63.39%

CNN:

- did not learn the problem at all
- probably training data too small
- $\rightarrow$ was not further pursued

Improvements:

- incorporate transcribed text data
- include videos to have facial expressions as additional feature
- using the validation set to find better hyperparameters
- $\rightarrow$ would make a plethora of established methods used in text processing available.
  - feature extracting with BoW- or tf-idf method
  - n-grams to capture context of a word

Thank You!