Speech Emotion

Josh Miller, Ben Schmitz, Jill Donahue

Motivation

- Speech: primary form of communication, especially with a machine
- Goal: make interaction with machines more human-like
- Detecting speech emotion can be helpful for smart assistants





Overview of Speech Emotion Field (Schuller)

- Speech emotion detection has been a challenging problem to tackle
- An accurate representation of the emotions needs to be created
- Complex networks and large amounts of data are typically required for an accurate model
- Challenges
    - Cross-cultural differences in emotions
    - Complex inflections like irony and sarcasm

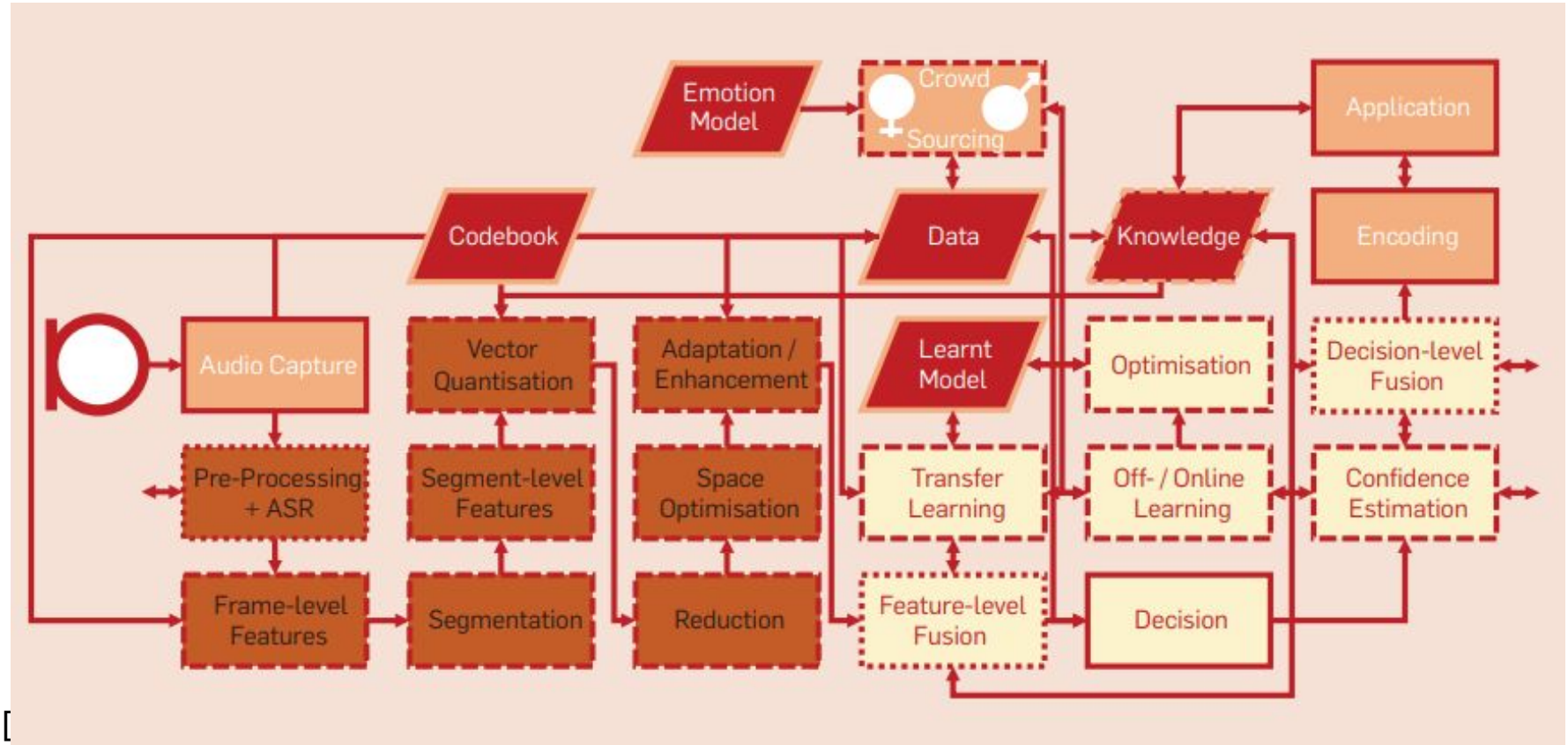
[8]

Standard Approach (Schuller)

- Modeling
  - Representing emotion so it can well be handled by a machine
  - Discrete classes
    - Happy, sad, etc.
  - Continuous dimension
    - Activation/Arousal
    - Valence
- Annotation
  - Acquisition of labelled data
- Features
  - Audio
    - MFCC, spectral flux, etc.
  - Textual
    - Identifying meaning
    - Can be combined with audio feature analysis

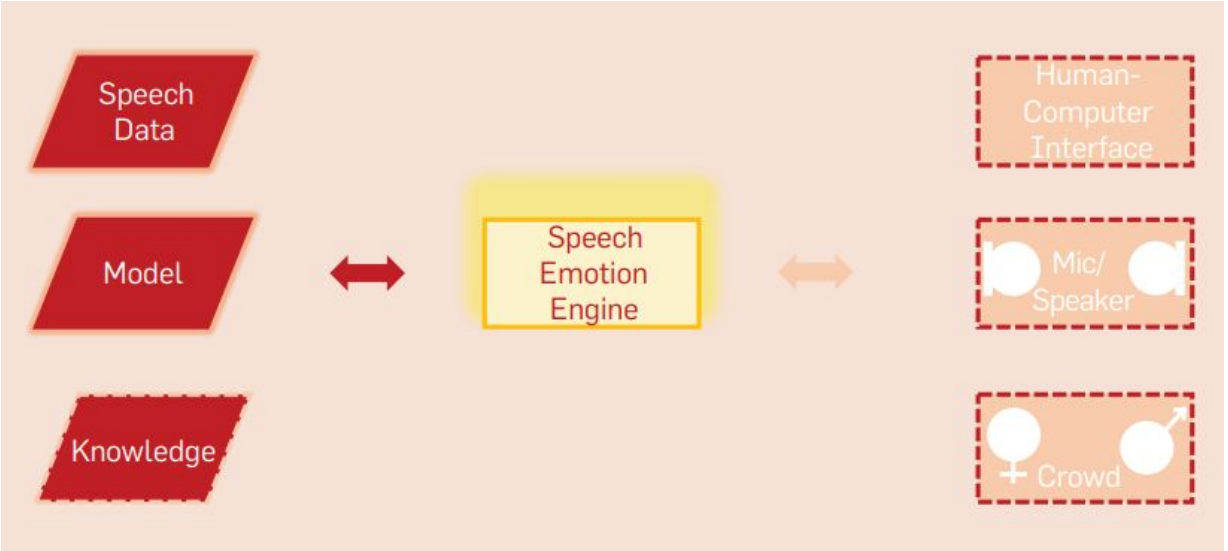
[8]

Standard Approach (Schuller)



Current State-of-the-Art (Schuller)

- Neural nets are popular because they have the ability to learn complex patterns like those that indicate emotion in speech
- Depends on large amounts of data



Busso et al.: Analysis of Emotion Recognition

Methods:

- Acted speech for happy, sad, angry and neutral
- Acoustical features:
  - Mean, standard deviation, range, maximum/minimum, and median of pitch and energy
  - Ratio of voiced and unvoiced speech
- SVM classifier

Results:

	Anger	Sadness	Happiness	Neutral
Anger	0.68	0.05	0.21	0.05
Sadness	0.07	0.64	0.06	0.22
Happiness	0.19	0.04	0.70	0.08
Neutral	0.04	0.14	0.01	0.81

[1]

Busso et al.: Analysis of Emotion Recognition

Criticism

- The authors generally see the downfall of their experiments
  - Not enough features for input
- Limitation of global pitch and energy features
  - Recognize MFCC - don't use it
- Database was only one voice actress

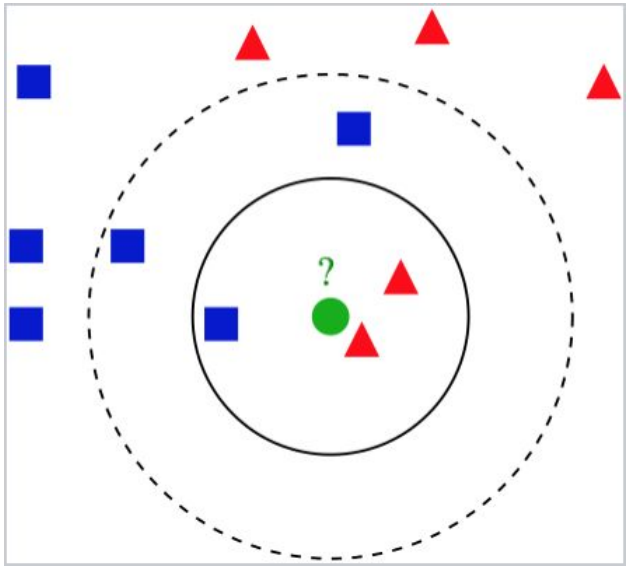
SVM

- Effective for more complex (non-linear) data
- Used for high-dimensional spaces
- Long training time

Aun and Rajoo: Influence of Languages

Methods:

- Malay, English and Mandarin
- Acted speech for happy, sad, angry and neutral used to compare languages
- Acoustical features:
  - MFCC
  - Pitch
  - Energy
  - Zero-Crossing Rate
- kNN classifier
  - Euclidean Distance



$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

[2]

Aun and Rajoo: Influence of Languages

Results:

- First experiment:

Language	Classification Rate				
	Happiness	*Sadness*	*Anger*	*Neutral*	*Average*
English	75	86	83	70	78.5
Malay	70	75	74	65	71.0
Mandarin	75	71	74	70	72.5
Average	73.3	77.3	77	68.3	74.0

- Second experiment:

Language	Classification Rate				
	Happiness	*Sadness*	*Anger*	*Neutral*	*Average*
Malay	73	78	77	71	74.75
Mandarin	78	80	77	69	76
Average	75.5	79	77	70	75.4

Aun and Rajoo: Influences of Languages

Criticism

- Use the same sentence
- Small set of data

kNN

- Effective with large training sets
- Robust to noisy data
- Computation time is large due to distance calculations
- Determine k and hand-pick features

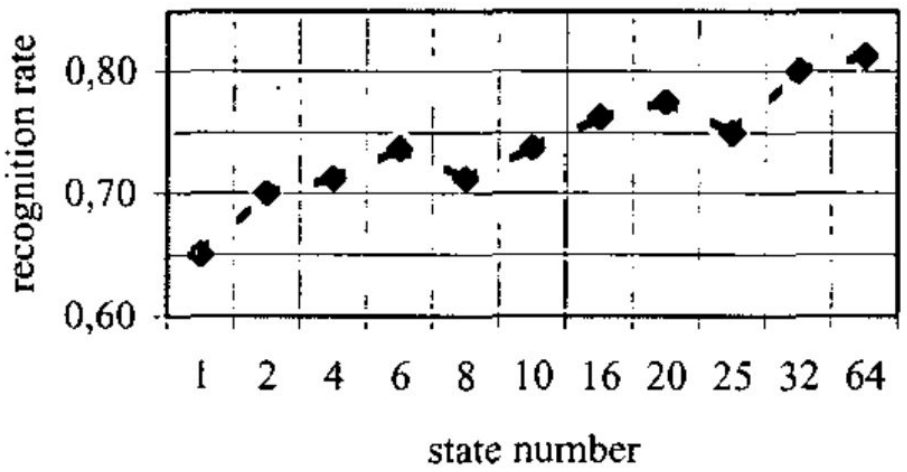
[3]

Feature Based Approach: Hidden Markov Models (Schuller et al.)

- Energy related features
  - Relative maximum and position of maximum of derivation of energy
  - Average and standard deviation of derivation of energy
- Pitch related features
  - Mean duration of voiced sounds
  - Average pitch, relative pitch minimum/maximum
- Hand pick optimal features and vectorize after modeling probabilities with Gaussian Mixture Models (GMM's) [4]

Schuller et al. Results:

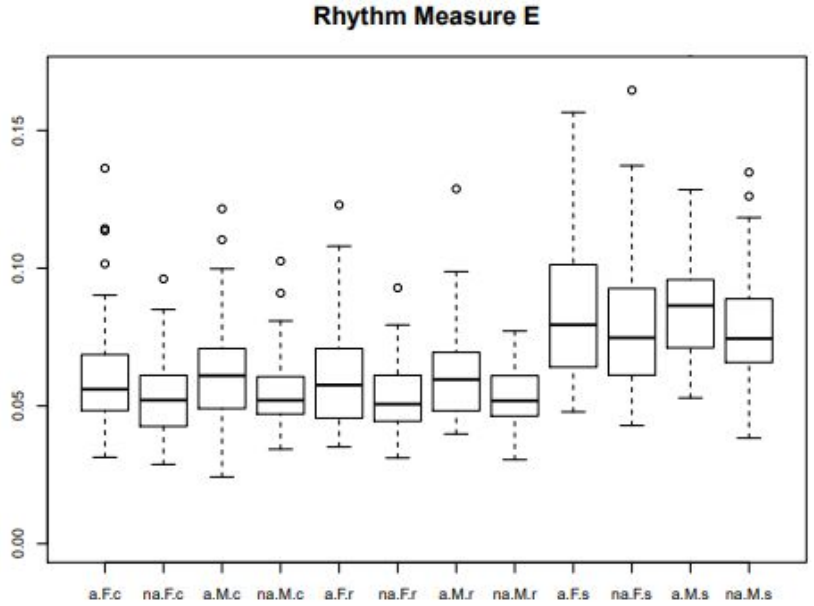
- Each emotion represented by one GMM
- Results dependent on number of states used
- Generally derived features give better results for less states
- **Results:**



Labeled emotion	Recognized emotion						
	ang	dis	fea	sur	joy	ntl	sad
ang	**68.5**	12.7	2.6	**1.8**	2.7	8.4	3.3
dis	12.8	**84.7**	2.1	0.3	0.0	0.1	0.0
fea	1.8	0.1	**95.4**	0.2	2.0	0.4	0.1
sur	6.3	6.7	6.3	**73.5**	6.1	0.9	0.2
joy	10.1	11.8	7.9	**1.2**	**68.0**	0.5	0.5
ntl	10.4	0.9	1.0	0.1	1.9	**79.6**	6.1
sad	5.9	10.1	2.8	2.1	2.2	1.8	**75.1**

Rhythm and Formant Features for Automatic Alcohol Detection (Schiel et al.)

- RMS rhythmicity and formant frequencies F1-F4 analyzed
- Alcohol Language Corpus dataset
- RMS measure primarily based on the median of the absolute value of the differences of the RMS values between maxima



[7]

Formant Analysis (Schiel et al.)

- Track vowels without any manual correction to the speech recordings
- Formant tracker configured to use a nominal F1 frequency
    - 531 Hz for male, 595 Hz for female
- Median formant frequencies F1 – F4 derived from the middle 30% of each vowel segment

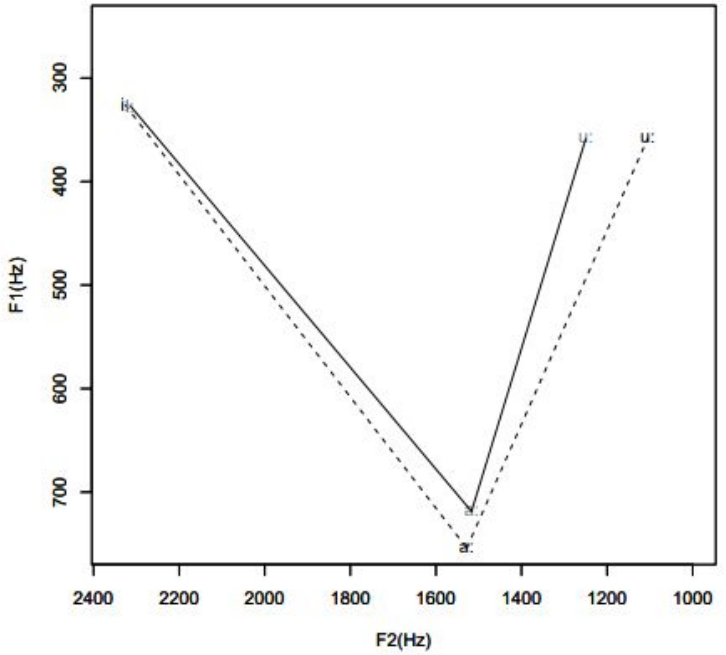


Figure 2: *Vowel triangle in the F1/F2 plane. The solid line combines vowel centroids of sober speech, the dashed line those of alcoholized speech.*

Schiel et al. Results and Opinions

- Significant differences between drunk and sober speech were found with formant based methods
- Even still, usefulness is limited for our desired applications
    - No prognosis rate is given by paper
- Differences detected likely too subtle for accurate detection

	$F1_m$	$F1_{qq}$	$F4_{qq}$	ED
alc	⇑	⇑	⇑	⇑
score	p<0.0001	p=0.046	p=0.004	p<0.0001
sex	F,M	F,M	F,M	F
vow	/a:/	/u:/	/a: i: u:/	/u:/
sty	r,s,c	r,s,c	r,s,c	r,s,c

	A	B	C	D, E
alc	⇑	⇑	⇑	⇑
score	p<0.0001	p<0.0001	p<0.001	p<0.0001
sty	r,s	r,s	c	r,s,c

	F	G	H	I
alc	⇑	⇑	⇑	⇑
score	p<0.0001	p<0.0001	p<0.0001	p<0.0001
sty	r,s,c	r,s	r,s	r

Neural Network Approaches

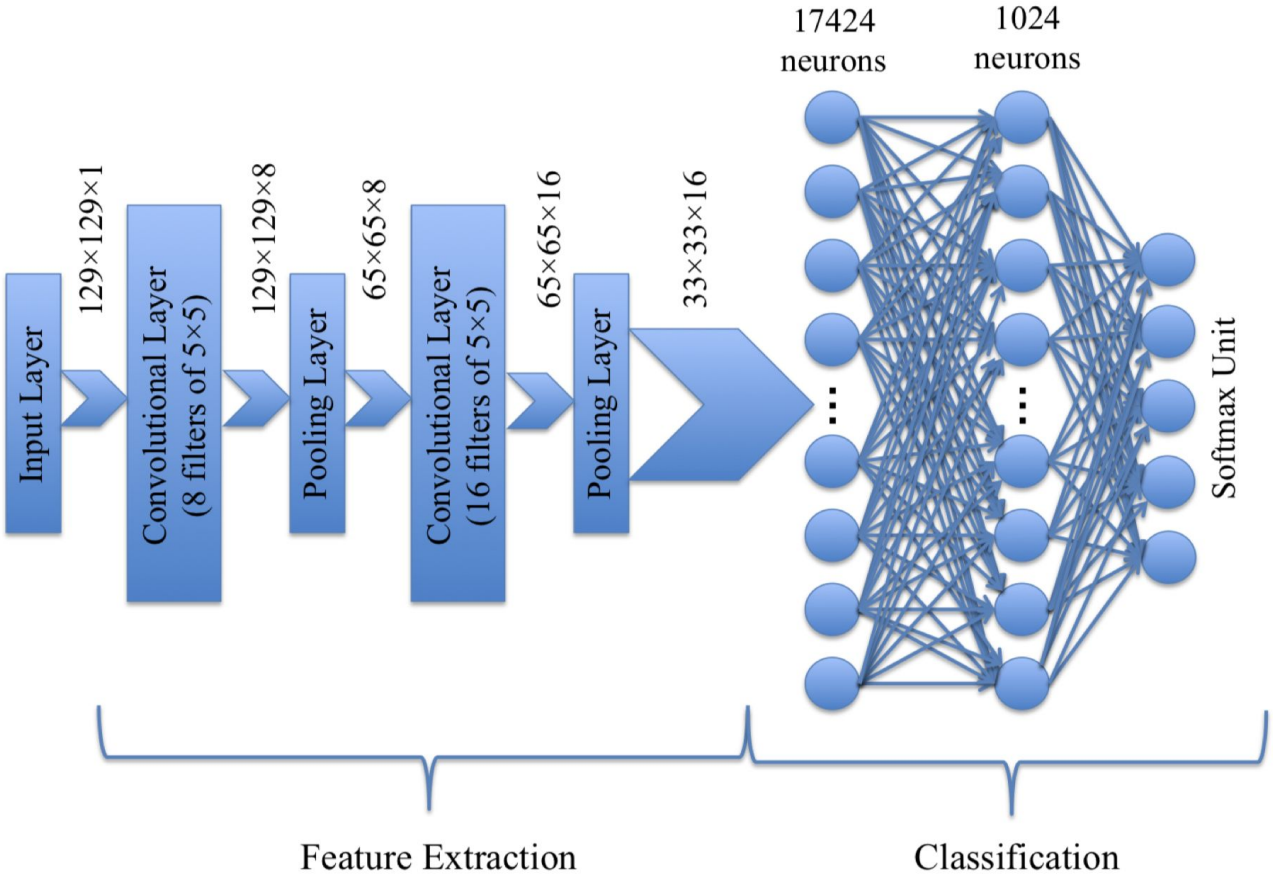
- Valence and arousal vs. direct labels
- Spectrogram as input data (no need to select features)
- May require more training data than other approaches
- Curriculum Learning may improve accuracy of model (Lotfian & Busso) [5]
    - Emulates human learning
    - Use disagreement on emotion from human evaluators as measure of difficulty
    - Minimax Conditional Entropy to account for noisy annotations

	Aro. [F-score]	Val. [F-score]
w/o curriculum	0.779	0.592
With random curriculum	0.771	0.590
Criterion 1-Error of predicted label	0.783	0.607*°
Criterion 2-Disagreement between annotators	0.789*°	0.615*°
Criterion 3-Minmax entropy	0.791*°	0.616*°

Speech Emotion Detection Using Convolutional Neural Networks (Shahsavarani)

- Large databases with varying emotional speech data
- Audio preprocessing
  - Wide band spectrogram data below 4 kHz, normalized and resized to 129 x 129
- Used data augmentation and dropout to reduce overfitting
- K-fold cross validation for training/testing network
  - Randomly divide data into K folds
  - K-1 folds used for training, remaining fold for testing
  - Repeat this process K times so that each fold is tested on once [6]

Convolutional Neural Network Architecture



Shahsavarani Results:

- Results for an American English Database dependent on convolution filter dimensions, pooling type, dropout probability, data augmentation, and epochs
- High accuracy potentially due to speech data

f_1	f_2	pooling	$p_{dropout}$	augmentation	epoch	CV accuracy (%)
5×5	5×5	Max	0	10x	100	66.36
5×5	5×5	Max	0.5	10x	100	73.81
10×10	5×5	Max	0.5	10x	100	73.38
10×10	5×5	Average	0.5	10x	100	75.55
5×5	5×5	Max	1	20x	100	63.64
5×5	5×5	Max	0.5	20x	100	72.91
10×10	5×5	Max	0.5	20x	100	68.9
10×10	5×5	Max	0.2	20x	100	78.17
10×10	**5×5**	**Average**	**0.2**	**20x**	**100**	**88.91**
10×10	**5×5**	**Average**	**0.2**	**20x**	**800**	**92.63**
10×10	**5×5**	**Average**	**0.2**	**20x**	**4000**	**94.51**

Summary of Findings

- Neural network methods are the cutting edge for speech emotion detection
- Drunkenness detection has a small pool of research mostly based on bag-of-features models
- Combining a neural net drunkenness detector with a general emotion detector could prove very useful for applications like AI assistants

References

[1] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, & Shrikanth Narayanan. 2004. “Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information.” In Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI '04). ACM, New York, NY, USA, pp 205-211.

[2] R. Rajoo & C.C. Aun. "Influences of languages in speech emotion recognition: A comparative study using Malay English and Mandarin languages" *Computer Applications & Industrial Electronics (ISCAIE) 2016 IEEE Symposium on. IEEE* , pp. 35-39, 2016.

[3] G. Xiao, C. Zha, X. Zhang & L. Zhao. 2015. “A Speech Emotion Recognition Method in Cross-Languages Corpus Based on Feature Adaptation.” In Proceedings of the International Conference on Information Technology Systems and Innovation (ICITSI ‘15). pp. 1-4. DOI: 10.1109/ICITSI.2015.7437680

[4] B. Schuller, G. Rigoll and M. Lang, "Hidden Markov model-based speech emotion recognition," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, Hong Kong, 2003, pp. II-1.

[5] R. Lotfian and C. Busso, “Curriculum learning for speech emotion recognition from crowdsourced labels,” arXiv preprint arXiv:1805.10339, May 2018.

[6] Shahsavarani, B. S. (2018). Speech Emotion Recognition using Convolutional Neural Networks (Master thesis, The University of Nebraska-Lincoln).

[7] Florian Schiel, Christian Heinrich, Veronika Neumeyer (2010). Rhythm and Formant Features for Automatic Alcohol Detection (Bavarian Archive for Speech Signals, Institute for Phonetics and Speech Processing, Ludwig-Maximilians-Universitat, Munchen, Germany).

[8] Bjorn W. Schuller (2018). Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks,and Ongoing Trends (Communications of the ACM).