Speech emotion recognition from acoustic and text feature

Bagus Tris Atmaja
bagus@jaist.ac.jp

AIS-Lab
School of Information Science
JAIST

JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY 1990

Presented on Indonesian student association at Kanazawa University
September 25, 2018

About Me...

- Education:
  - B.Eng in Engineering Physics ITS (2009)
  - M.Eng in Engineering Physics ITS (2012)
  - Research Student at Kumamoto University (2011-2012)
- Experience:
  - Shimizu Seisakusyo, Kameyama-shi, Mie-ken (2012-2014)
  - VibrasticLab, Dept.of Engineering Physics ITS (2014 -)
  - PhD student at Acoustic Information Science-Lab, JAIST (2017-)
  - Instructor at the Carpentries (2017-)
- Research interest:
  - Speech processing
  - Noise control
  - Machine condition monitoring

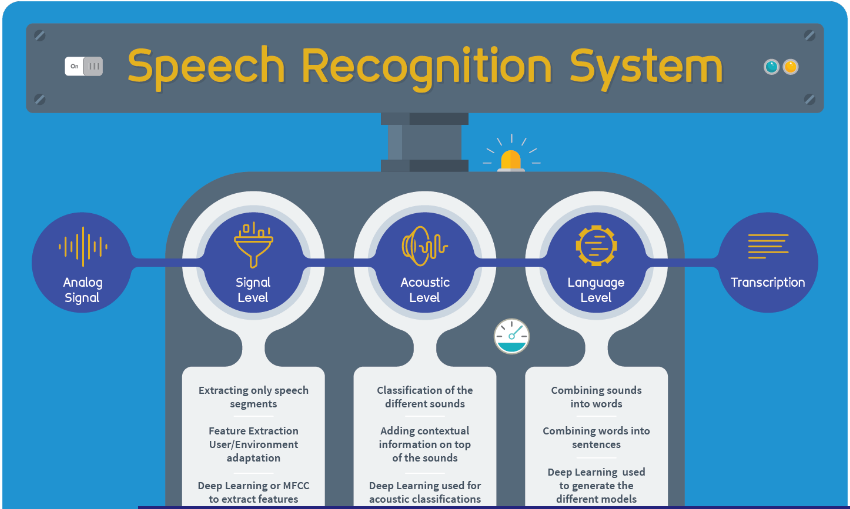
[0] This slide .tex source can be download here: github.com/bagustris/beamer-nomi

Research Motivation

- Speech contains a variety informations: linguistics, paralinguistics and nonlinguistics information.
- Ideally, speech should convey the correct message (intelligibility) while sounding like human speech (naturalness) with the right prosody (expressiveness).
- Most speech recognition system are focused on solving the first two issues above.
- We proposed to to recognize expressiveness in speech by using **linguistics (text)** and **paralinguistics (acoustic)** features to obtain nonlinguistics (emotion) information.
- Why? Because text features can be extracted from through Automatic Speech Recognition (ASR) or Speech to Text (SST) method (Google Assistant, Siri, Alexa, Cortana, DeepSpeech).

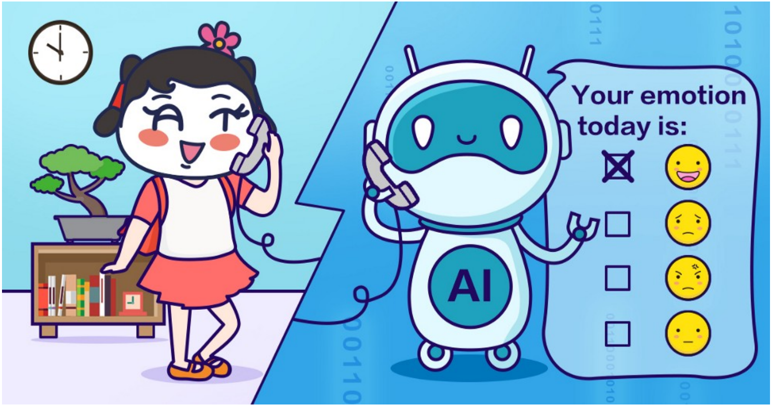
Speech Recognition:

The ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format



(Speech) Emotion Recognition

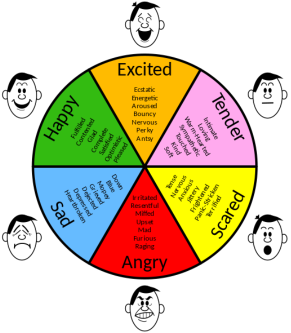
Speech emotion recognition is to study the formation and change of speakers emotional state from the speech signal perspective.



Emotion in Category

According to Ekman[1], there are six basic (facial) emotions:

- happy
- surprise
- fear
- disgust
- anger
- sad

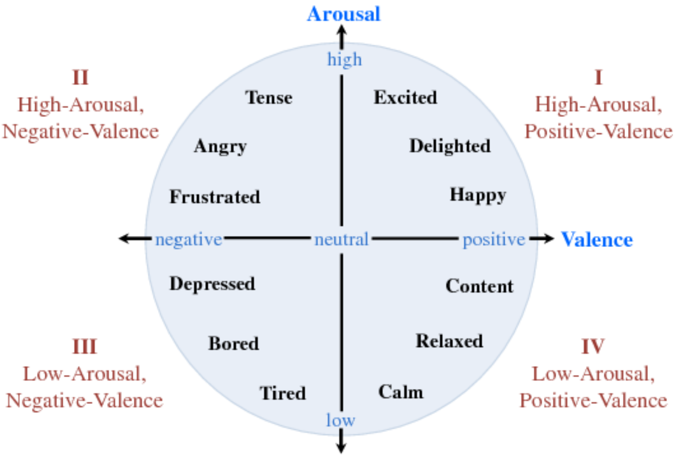


Recent research by Jack, R[2]. E et.al. (also supported by J.H Turner) revealed only 4 category of emotion: happy, sad, fear, anger.

[1] Ekman, P., Friesen, W. V., Ellsworth, P. "Emotion in the Human Face..". Pergamon (1972).

[2] Jack, Rachael E., et al. Four not six: Revealing culturally common facial expressions of emotion. Journal of Experimental Psychology: General 145.6 (2016): 708.

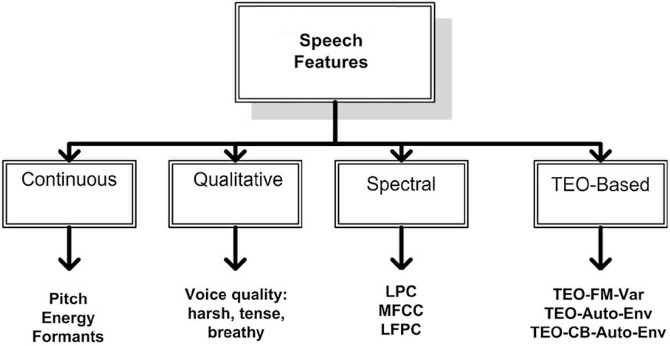
Emotion in Dimensional VA(D) Space



Valence: Happy/Unphappy, Arousal: Activity, Dominance: Potency

How to determine emotion from speech?

To determine emotion from speech we extract acoustic features and train them in (deep) learning method.



TEO: The Teager-energy-operator (proposed by Teager, 1990) based on theory that hearing is the process of detecting energy.

How to obtain text feature?

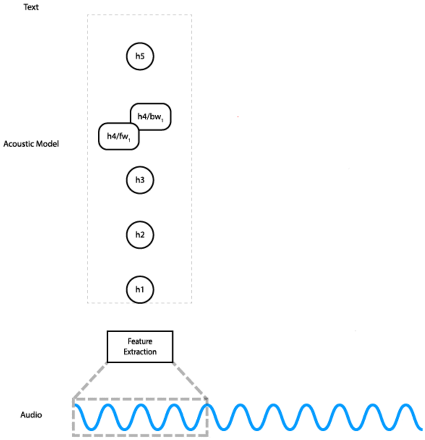


Figure 1: Data flow in Deep speech
https://hacks.mozilla.org/2017/11/a-journey-to-10-word-error-rate/

- Given speech spectrograms (time-frequency) as input, RNN train to produce output characters directly.
- The output of the network is a matrix of character probabilities over time
- For each time step, the network outputs one probability for each character in the alphabet (likelihood of that character corresponding to whats being said in the audio at that time).

How to use text feature?

- New Affective Norms of English Words (ANEW[3]) contain of 13,915 word with its valence arousal, and dominance value that can be used to predict emotion in text.

Table 6 Correlations between emotional dimensions and semantic variables reported in prior studies [degrees of freedom are based on the numbers of data points reported as *N* (Overlap)]

Source	Measure	*N* (Source)	*N* (Overlap)	Valence	Arousal	Dominance
a	Imageability	5,988	5,125	.161	–.012	.031
b	Imageability	326	318	–.037	.099	–.160
	Concreteness	326	318	.109	–.244	–.019
	Context Avail.	326	318	.196	–.147	.044
c	Concreteness	1,944	1,567	.105	–.258	.009
d	Imageability	3,394	2,906	.152	–.045	.006
	Familiarity	3,394	2,906	.206	–.028	.215
e	AoA[1]	30,121	13,709	–.233	–.062	–.187
	% Known[2]	30,121	13,709	.094	.078	.103
f	Sensory Exp.	5,857	5,007	.067	.228	–.044
g	Body–Object	1,618	1,398	.203	–.143	.172
h	Familiarity	559	503	.272	–.193	.329
	Pain	559	503	–.456	.579	–.343
	Smell	559	503	.139	.052	–.043
	Color	559	503	.401	.052	.081

[3] A. B. Warriner, V. Kuperman, and M. Brysbaert, Norms of valence, arousal, and dominance for 13,915 English lemmas, Behav. Res. Methods, vol. 45, no. 4, pp. 11911207, 2013

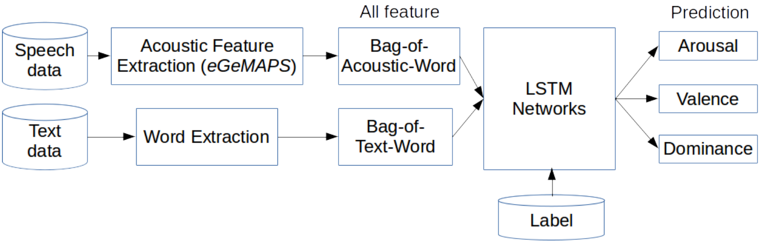
How to use text feature? [2]

- Perform ANEW analysis to calculate VAD score from IEMOCAP[4] dataset.
- To compute total VAD score in each utterances, we currently use mean and median method for each words in utterances that has VAD score in ANEW list.
- Compare result from ANEW analysis with IEMOCAP evaluation.
- Expected result:
  Score of CCC, CC and RMSE.

[4] C. Busso et al., IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval., vol. 42, no. 4, pp. 335359, 2008.

How to train speech-text feature?

- This network consist of two part, feature extraction part and LSTM-RNN part.
- In this scenario, we extracted all feature from acoustic and text as described in the previous page.
- The extracted features are concatenated to feed 2 layer LSTM and model the contextual information in the label.



How to evaluate the system?

We use the following three different objective function to measure the performance. x is each VAD (valence, arousal, dominance) score from dataset, and y is predicted each VAD score from our algorithm.

- Concordance Correlation Coefficient (CCC):

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{1}$$

- Pearson Correlation Coefficient (CC):

$$\rho_{xy} = \frac{cov(x, y)}{\sigma_x\sigma_y} \tag{2}$$

- Root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(x-y)^2}{N}} \tag{3}$$

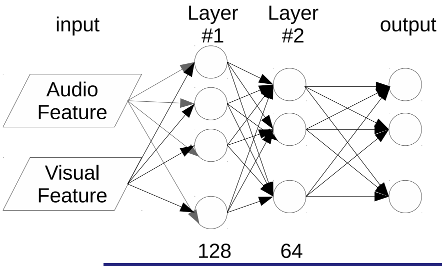
What's the contribution of this research?

- The use of deep learning technique (LSTM-RNN/CNN) for dimensional speech emotional recognition.
- The number of dominant feature extracted from bag-of-acoustic-words (BoAW) and bag-of-text-words (BoTW) that contributes significantly to speech emotion recognition performance by feature selection algorithm.
- A VAD-based text emotion recognition method by (1) ANEW analysis, and (2) machine learning algorithm.
- A method to integrate acoustic and text feature for speech emotion recognition.

Terima Kasih

Current works:
Cross-cultural Video/Audio Emotion Recognition

- Task: Given German utterances with its label (Valence, arousal, and liking score) to predict Hungarian utterances emotion score.
- Proposed solution:Using LSTM algorithm to train valence, arousal and liking from German language to predict its dimension in Hungarian from different number of acoustic and visual features.
- Network architecture:



Current works:
Cross-cultural Video/Audio Emotion Recognition

Table 1: Parameters in LSTM network

Parameter	Value
batch size	34
learning rate	0.001
num iter	50
num units 1	128
num units 2	64
bidirectional	False
dropout	0.2