Beamer Nomi:
A beamer template for JAIST

Bagus Tris Atmaja
bagus@jaist.ac.jp

AIS-Lab
School of Information Science
JAIST

June 30, 2019

Motivation

- First motivation, `within \itemize`
- Second motivation, `with \item`

Problem

- Given a set of (cross-cultural) dataset (train and devel) with its label, how to predict emotion dimension on test data?
- How to build multimodal emotion recogniton with LSTM network?
- How feature selection can improve AER?

Purpose:

- Build multimodal emotion recognition using feature selection and LSTM, compare with baseline sytem (without feature selection)

Example of Table

- SEWA dataset is used, consisting audiovisual spontaneous behaviors of participant recorded *in-the-wild*.
- Annotation (labels) is available for the emotional dimensions arousal and valence, and a third dimension describing liking (or sentiment), by 6 (German) or 5 (Hungarian) native speakers.

Table 1: Number of instance for each partition of recordings

Partitions	German	Hungarian	Labels	Total
Training	34	-	✓	34
Development	14	-	✓	14
Testing	16	66	-	82
Total	64	66	130	130

Example of footnote as reference

- Use `\footnotetext` for Reference.
- Use `\footnotemark` to mark.
- Available features:

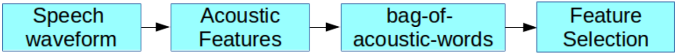
Table 2: Number of audio and video features

Model	LLDs	Bag of words
Audio	23 eGeMAPS	100
	39 MFCCs	100
Visual	17 FAUs	100

[0] eGeMAPS: Geneva Minimalistic Acoustic Parameter Set, Eyben et. al., 2013
MFCCs: 1 to 13, inculding deltas and deltas-deltas, totally 39
FAU: Facial Action Units

Example of footnote

- Use `\centering` for centering footnote, sometimes looks ugly.
- Use `\tiny` to make footnote text tiny[1].
- Line above footnote is automatic.
- figure within `\item` doesn't need `\centering`



[1] F. Ringeval et al., AVEC 2017-Real-life Depression, and Affect Recognition Workshop and Challenge, 2017.

Example of Table: plain latex

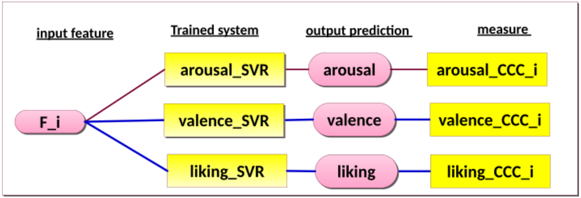


Table 3: Result of Optimal Set algorithm using visual features (BoVW FAUs)

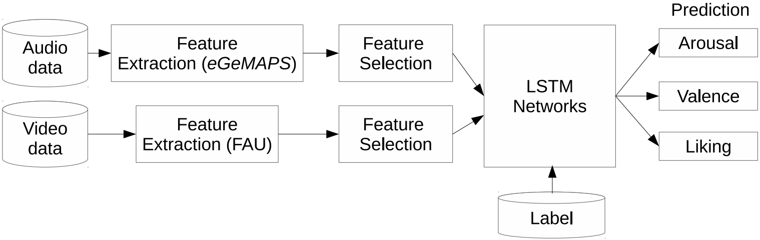
Valence FID	CCC	Arousal FID	CCC	Liking FID	CCC
VI31	0.192	VI31	0.274	VI99	0.098
VI17	0.170	VI17	0.239	VI74	0.055
VI60	0.153	VI62	0.187	VI41	0.041
VI20	0.147	VI20	0.177	VI40	0.027
VI35	0.125	VI60	0.164	VI12	0.019

Example algorithm, need algorithmic package

1: Input gold standard values for emotion dimension ED_i for development partition.
2: Input features sorted by abs(CCC): $f_1, f_2, f_3, ..., f_n$
3: Input impact CCC_1 of f_1
4: $Optimal\_Set = \{f_1\}, CCC\_Optimal = CCC_1$
5: **for** f_j in $f_1, f_2, f_3, ..., f_n$ **do**
6: $\quad CCC_j$=CCC(Predict ([$Optimal\_Set, f_j$]))
7: $\quad$ **if** $CCC_j > CCC\_Optimal$ **then**
8: $\quad\quad CCC\_Optimal = CCC_j$
$\quad\quad Optimal\_Set = [Optimal\_Set, f_j]$
9: $\quad$ **end if**
10: **end for**
11: **return** $Optimal\_Set, CCC\_Optimal$

Method: LSTM network

- Using LSTM algorithm to train valence, arousal and liking from German language to predict its dimension from different number of acoustic and visual features.
- LSTM network can model the context (VAD value) while insensitive to outliers[2]
- Network architecture:

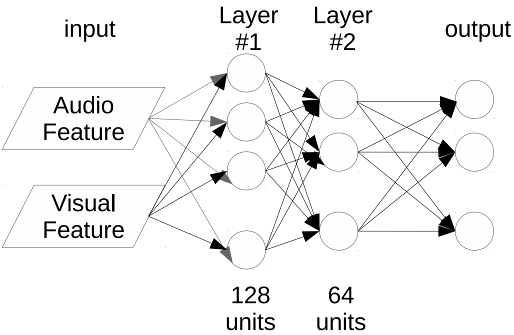


[2] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, End-to-End Multimodal Emotion Recognition using Deep Neural Networks, vol. 14, no. 8, pp. 19, 2017.

Use column two divide into 2 sides

Table 4: 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



Example of Equations

We use the following 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):

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

The average of CCC for the three-emotion dimension is used as a measure for the performance of the whole system; this measure is defined by the following equation.

$$CCC_{avg} = \frac{\sum CCC_i}{n} \tag{2}$$

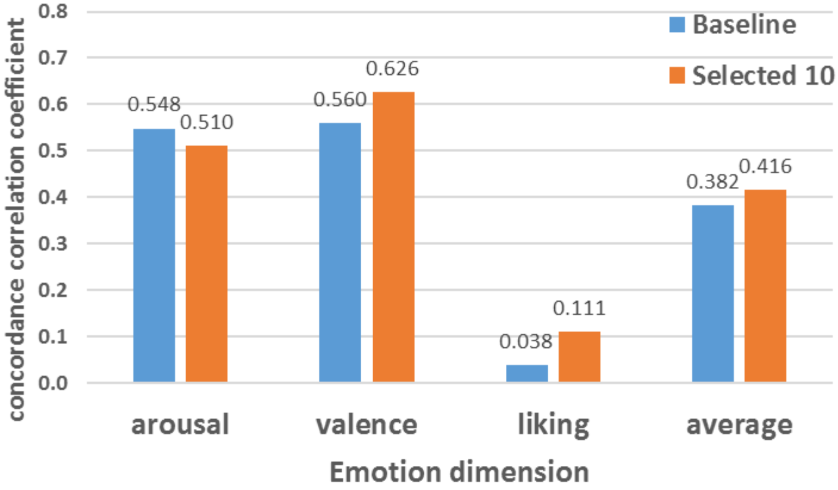
where n is number of i dimension, i.e. 3 (valence, arousal, liking).

Result: Development

Table 5: Evaluation results using mono-language case by using the development partition form German language using selected features from audio-video modalities

Features Set	Arousal	Valence	Liking	Average
Baseline (100)	0.552	0.563	0.238	0.451
Selected (6)	0.641	0.636	0.278	0.518
Selected (10)	0.660	0.620	0.298	0.526
Selected (15)	0.622	0.623	**0.314**	0.520
Selected (20)	0.616	0.596	0.299	0.504
Optimal Set	**0.678**	**0.654**	0.304	**0.545**

Result: Testing Data



Conclusion

- The use of deep learning technique (LSTM-RNN/CNN) for dimensional speech emotional recognition from multimodal feauture has been presented.
- 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.
- The result shows promising result on development data, slightly improvement on testing data, but poor performance on cross-cultural test data, due to limitation of dataset.

Remaining Problem/Future Works

- Use total CCC instead of averaged CCC as overall accurate prediction is desired.
- Compare Acoustic feature only to Audio/Video-based.
- Implement the similar scenario for IEMOCAP dataset.