RNN-based Dimensional Speech Emotion Recognition

Bagus Tris Atmaja*,
Reda Elbarougy,
Masato Akagi
*bagus@jaist.ac.jp

AIS-Lab
School of Information Science
JAIST

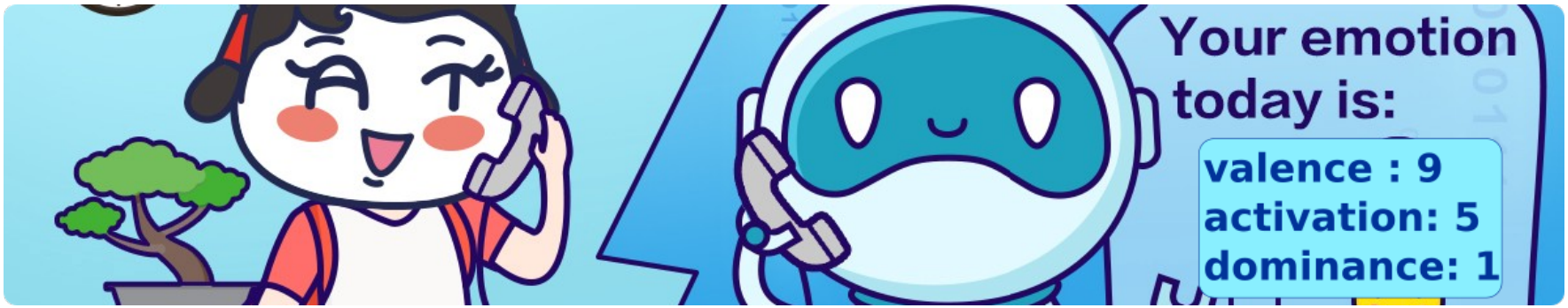
ASJ Autumn Meeting @Ritsumeikan Univ.
4 September 2019

Outline

Paper, slide, & experiment codes available at
`http://github.com/bagustris/asj_autumn_2019`

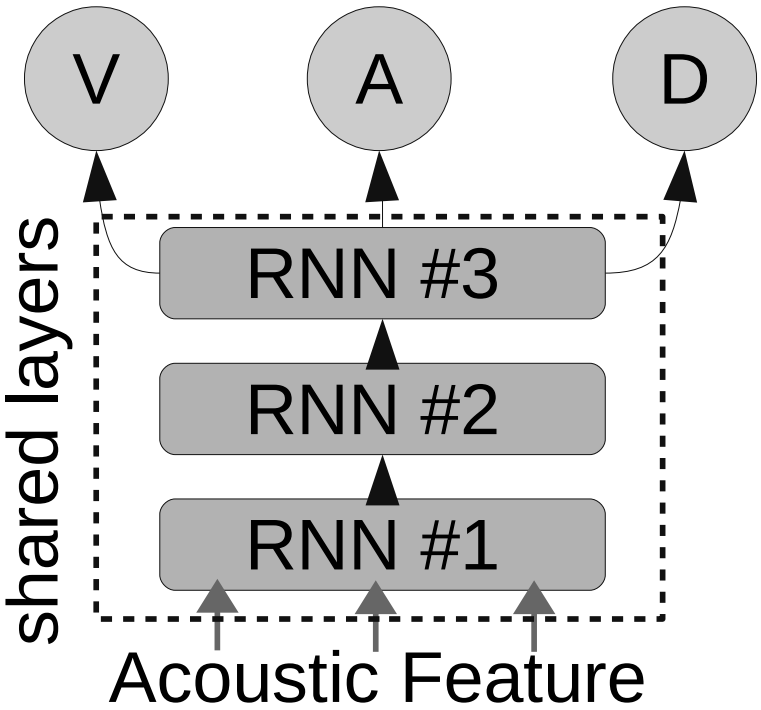
Motivation

- The need of recognizing human emotion automatically by machine increases on demand of such applications like call center or humanoid robotics.
- Most speech emotion recognition systems analyze human emotion in categorical views (angry, sad, fear, happy, etc)
- Recognizing "degree" of emotion (dimensional emotion) gives more benefits e.g. how fear, how angry (very angry, not so angry, etc)
- A recurrent neural network (RNN)-based is proposed to solve dimensional emotion recognition with multitask learning.



Dimensional Speech Emotion Recognition

- Input: Acoustic Features
  - 31 Features
  - eGeMaps Feature set
- Classifer: RNN
  - GRU
  - LSTM
- Evaluation Metrics:
  - MSE (mean squared error)
  - MAE (mean absolute error)
  - MAPE (mean absolute percentage error)
  - **CCC (concordance coefficient correlation)**



The dataset

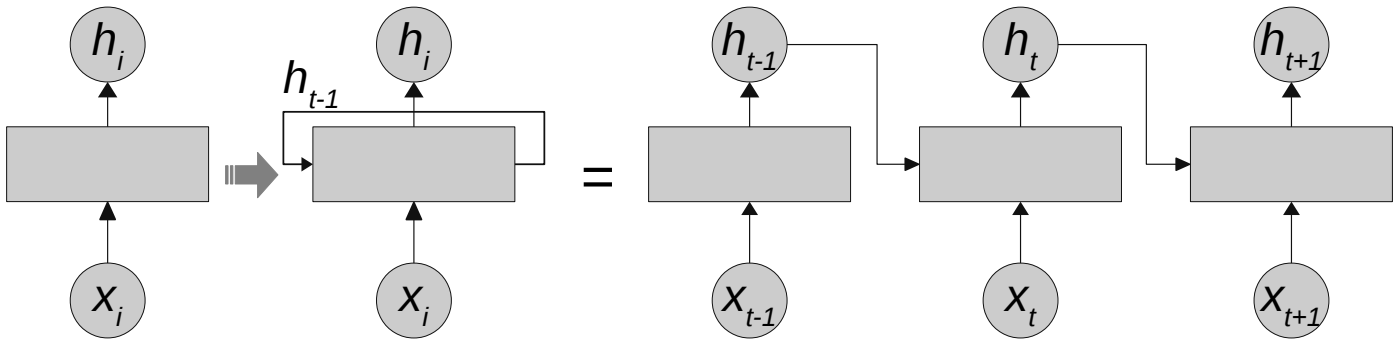
- IEMOCAP (interactive emotional dyadic motion capture database) was used, among many modalities (speech, face, movement, text) only speech is used.
- Total utterances is 10039 sentences from 10 speakers in dyadic conversation (12h).
- Training/test split: 8000:2039, 20% of training data is used for validation/development.

Text	v	a	d
'Excuse me.'	2.5	2.5	2.5
'That's out of control.'	2.5	3.5	3.5
'Did you get the mail? So you saw my letter?'	2.5	2.0	1.5
'Did you get the letter?'	3.5	3.0	2.0

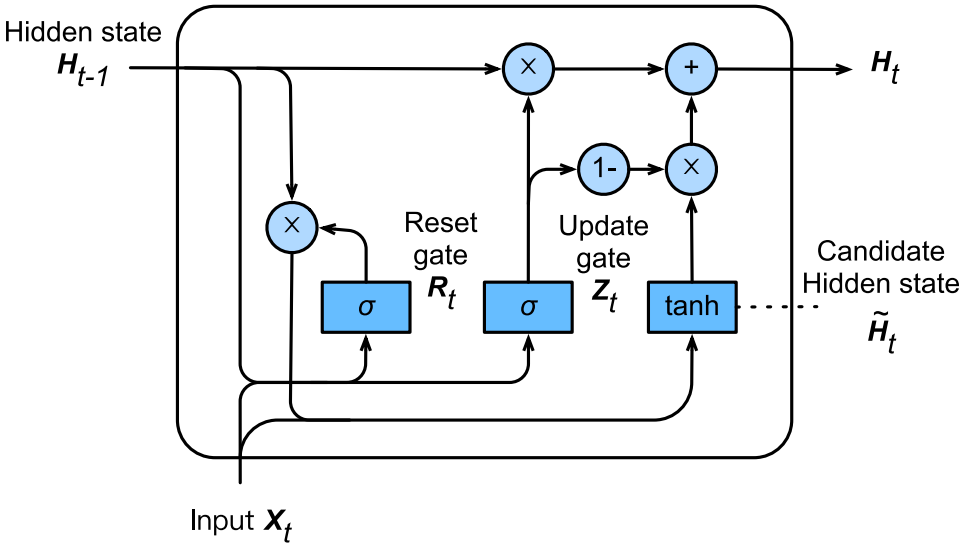
Acoustic Features Extraction

- 31 Features
  - 3 time domain features (Zero Crossing Rate, Energy, Entropy of Energy); 5 frequency domain features (Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral Flux, Spectral Rollof); 13 Mel-frequency cepstral coefficients (MFCCs), 5 Fundamental frequencies, 5 Harmonics.
  - Extracted on each frame with 20 ms window size and 10 ms stride.
- eGeMaps (Geneva Minimalistic Acoustic Feature Set, 23 Features)
  - loudness, alpha ratio, hammarberg index, spectral slope 0-500 Hz, spectral slope 500-1500 Hz, spectral flux, 4 MFCCs, F0, jitter, shimmer, Harmonics- to-Noise Ratio (HNR), Harmonic difference H1-H2, Harmonic difference H1-A3, F1, F1 bandwidth, F1 amplitude, F2, F2 amplitude, F3, and F3 amplitude.
  - Extracted on each frame with 25 ms and 10 ms stride.

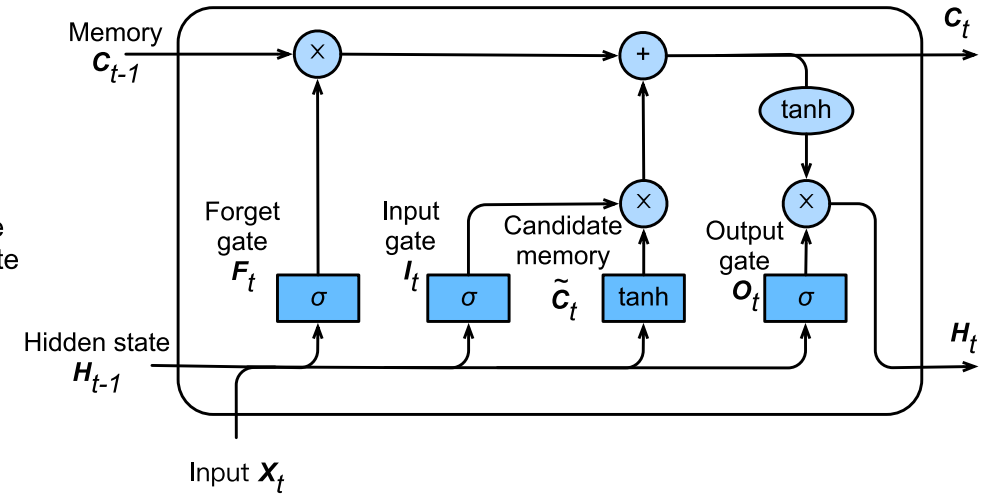
Recurrent Neural Network



NN RNN Unrolled RNN

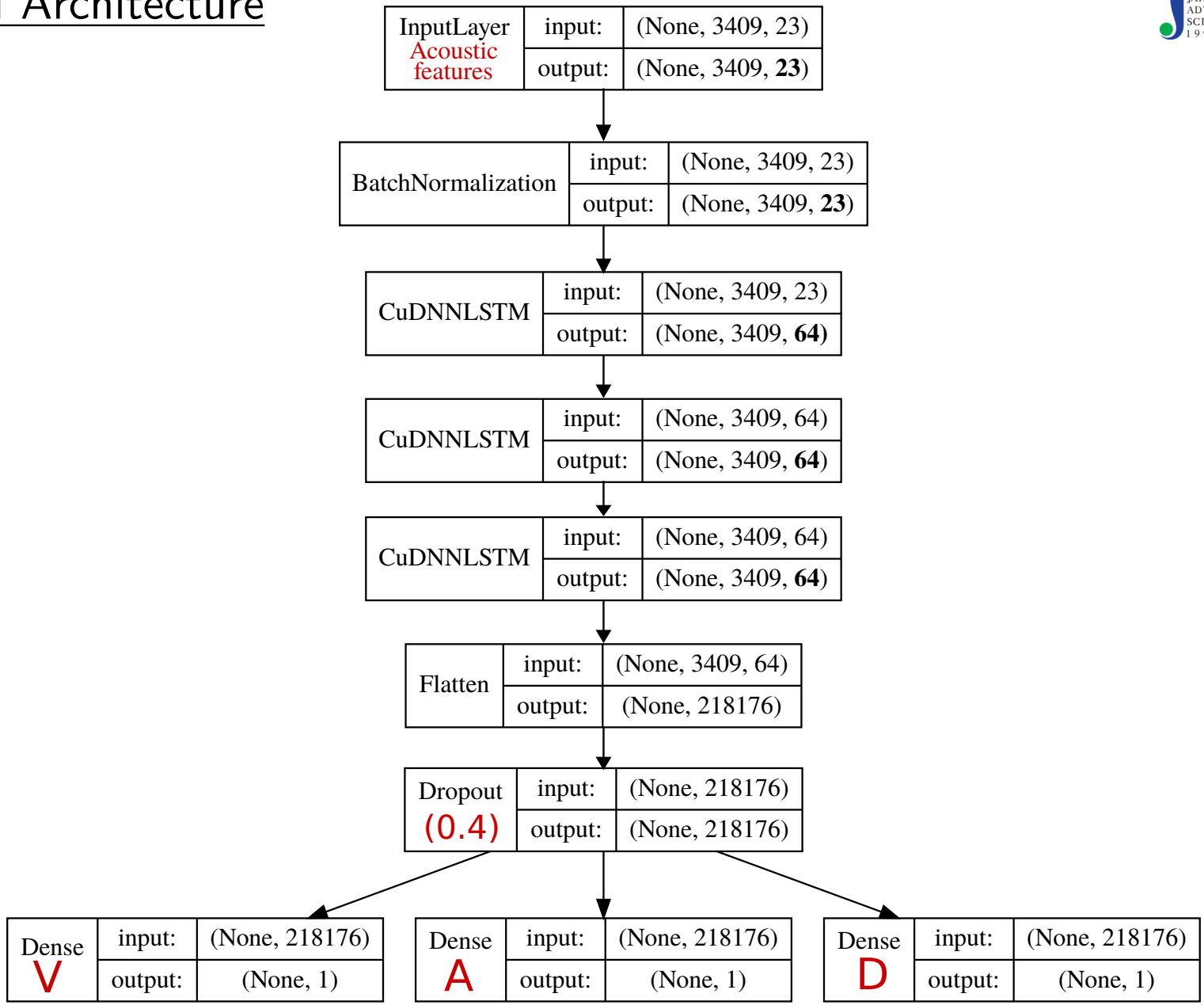


GRU



LSTM

RNN Architecture



Multi-task Learning

- Instead of minimizing error (e.g. MSE), we minimize concordance correlation coefficient (CCC) loss,

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

$$CCCL = 1 - CCC$$

- Where CCCL is summation of CCCL from valence, arousal, and dominance. We define out mulitask learning as follows,

$$CCCL_{tot} = \alpha CCCL_V + \beta CCCL_A + \gamma CCCL_D$$

- Where α, β, γ are obtained by experiments (0.7, 0.3, and 0.6).
- The goal is to improve the concordance coefficient correlation between true emotion degree with predicted emotion degree.

Result

Table 1: Results of dimensional emotion recognition among different method and metric; Each score is averaged scores from V, A, and D. For error, lower is better. For CCC, higher is better (-1 to 1).

Method	MSE	MAPE	MAE	CCC
	31 Features			
DNN	1.441	32.372	0.965	0.050
GRU	1.332	30.802	0.925	0.076
LSTM	1.068	28.278	0.823	**0.088**
	eGeMaps			
DNN	0.955	25.855	0.7	0.198
GRU	0.663	23.488	0.644	0.234
LSTM	0.683	23.814	0.655	**0.245**

Result

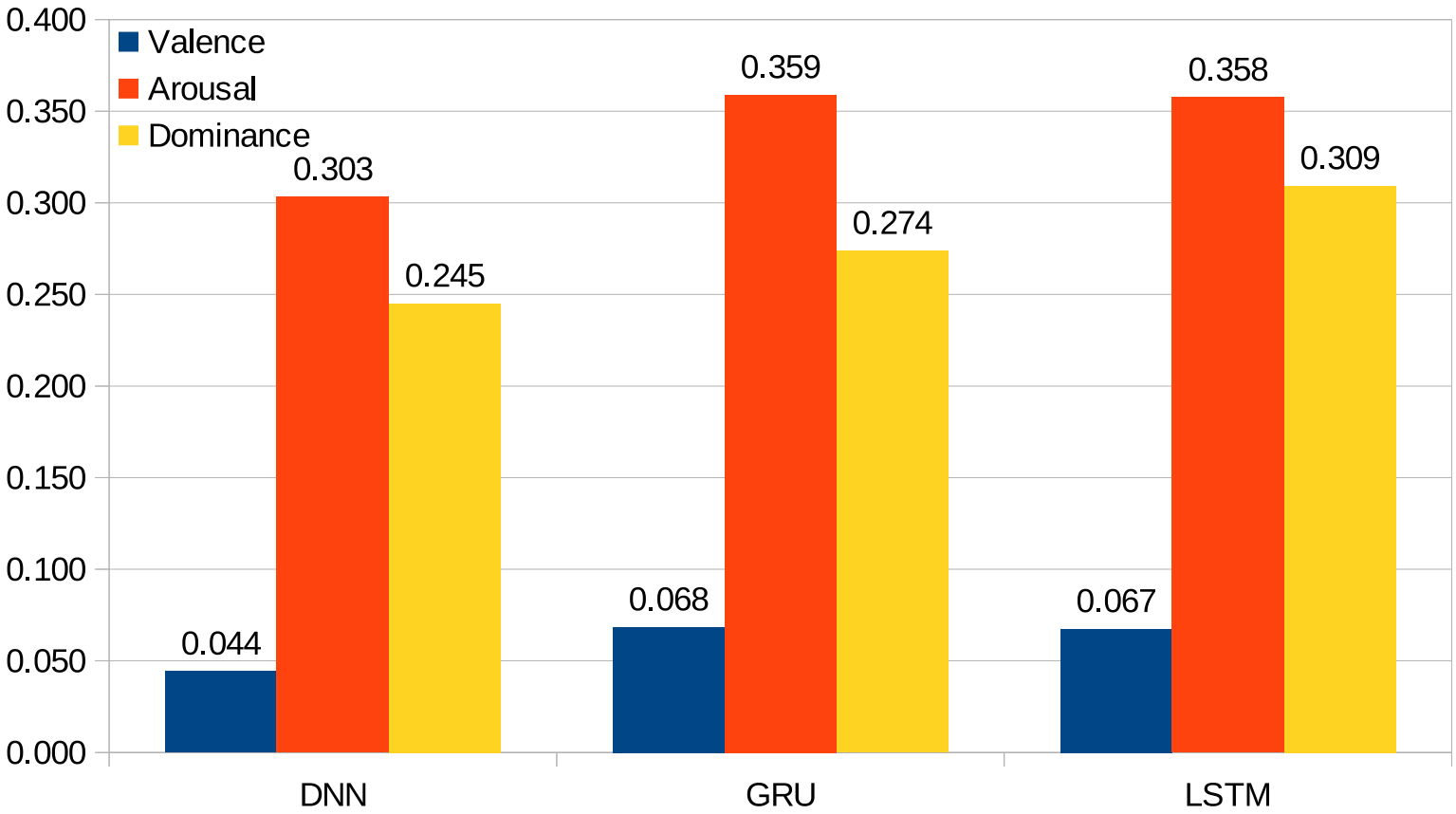


Figure 1: CCC score for each emotion dimension from 50 epochs.

- CCC score of LSTM networks from 100 epochs: [0.11, 0.43, 0.36]

Conclusion

- A RNN-based dimensional speech emotion recognition is presented. For feature set, we evaluate two sets of acoustic features: 31 Features and eGeMaps feature set which shows eGeMaps obtained the better result with LSTM network.
- A multitask learning by employing CCC loss for valence, arousal, and dominance is proposed with three parameters. By experiment, the highest CCC score is achieved using $\alpha = 0.7, \gamma = 0.3,$ *and* $\beta = 0.6$.
- The highest CCC score by those parameters within LSTM network is [0.11, 0.43, 0.36] for valence, arousal, and dominance.