RNN-based Dimensional Speech Emotion Recognition

JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY 1990

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

Background

- The need of recognizing human emotion by machine automatically increases on demand of such applications like call center or humanoid robotics.
- Most speech emotion recognition analyze human emotion in categorical views (angry, sad, fear, happy, etc).
- Recognizing "degree" of emotion is important because it enables deeper analysis on how weak/strong emotion.
- Dimensional emotion represents emotions in a two- or three-dimensional space e.g. VAD (Valence – positive/negative, Arousal – excited/calm, and dominance – degree of control).

Problem & Purpose

- **Problem**: How to recognize emotion in 3 dimensional VAD space, i.e. predicting score V, A, and D from speech utterances?

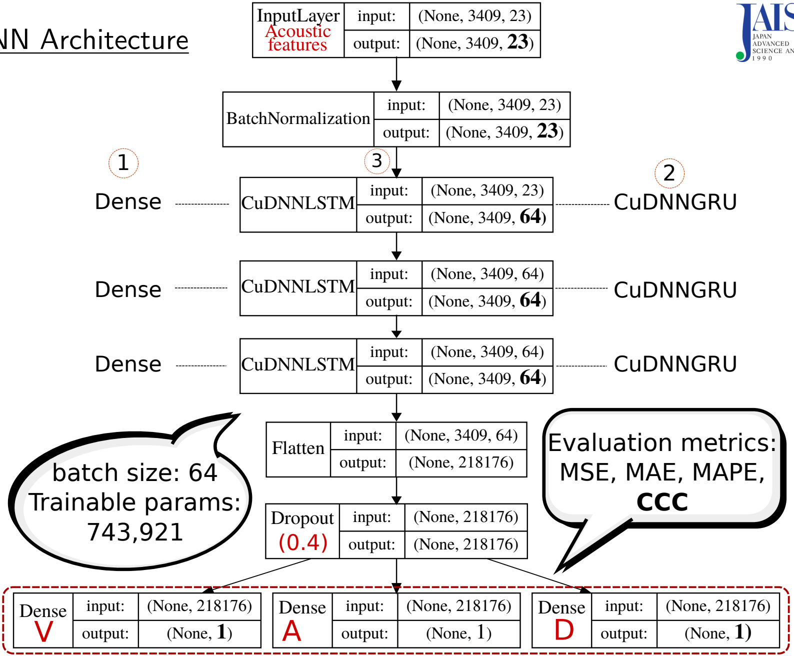
- **Purpose**: Evaluate a recurrent neural network (RNN)-based system for predicting dimensional emotion degree with multitask learning which learns together to predict score of V, A, and D.

Dataset and Acoustic Feature

- Dataset:
  - Name: IEMOCAP (interactive emotional dyadic motion capture database)
  - Modality: Speech
  - Number of utterances: 10039 (Training/Test : 80/20)
  - Duration: 12h
- Acoustic Features:
  - 31 Features: 3 time domain features, 5 frequency domain features, 13 MFCCs, 5 F0, 5 Harmonics.
  - eGeMaps Feature set (Geneva Minimalistic Acoustic Parameter Set)[1] : 23 features, e.g. Loudness, alpha ratio, hammarberg index, spectral slope, spectral flux, 4 MFCCs, F0, jitter, shimmer, Harmonics-to-Noise Ratio (HNR), etc.

[1] F. Eyben et al., The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing, IEEE Trans. Affect. Comput., vol. 7, no. 2, pp. 190202, 2016.

RNN Architecture



Multi-task Learning

- Instead of minimizing error (e.g. MSE), we minimize concordance correlation coefficient (CCC) loss (CCCL).
- CCC measures the association between variables and penalizes the score even if the model predicts the emotion well, but it shifts the value.

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

$$CCCL = 1 - CCC$$

- Where CCCL is summation of CCCL from valence, arousal, and dominance. We define our multitask 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).

Result among different methods

Table 1: Results of dimensional emotion recognition among different methods and metrics; Each score is averaged score from V, A, and D. For error, the smaller the better. For CCC, the higher the better (-1 $\sim$ 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 on each V, A, and D dimension

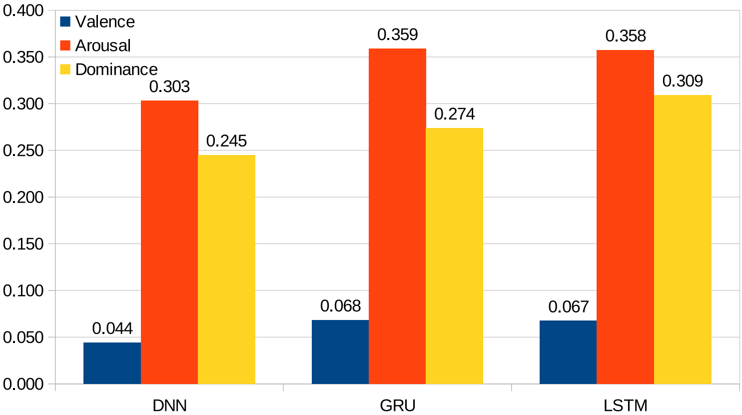


Figure 1: CCC score for each emotion dimension after 50 epochs using eGeMaps feature set.

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

Conclusion

- An RNN-based dimensional speech emotion recognition is presented by utilizing three network layers using either LSTM or GRU layers and split the last RNN layer into three dense layer with 1 unit to represent/predict score of valence, arousal, and dominance.
- A multitask learning is employed by introducing new loss function namely CCC loss which minimizes concordance correlation between true value and predicted score for V, A, and D degree simultaneously.
- By tuning the parameters in that multitask learning (i.e. $\alpha = 0.7, \gamma = 0.3,$ *and* $\beta = 0.6$.), the highest CCC score among 2 acoustic feature sets and 3 network architectures is [0.11, 0.43, 0.36] for valence, arousal, and dominance, which is obtained using eGeMaps feature set and LSTM networks.