RNN-based Dimensional Speech Emotion Recognition

Bagus Tris Atmaja* , Reda Elbarougy, Masato Akagi (JAIST)

1 Introduction

Automatic speech emotion recognition is ordinarily performed in categorical views by predicting class of speech emotion whether it belongs to to happy, sad, angry, or other emotion categories. However, recent research suggested that emotion attributes, such as arousal, valence, and dominance, is more challenging as it enables analysis of emotion in 2D or 3D space where emotion category also can be obtained.

This paper presents dimensional speech emotion recognition using recurrent neural networks (RNN). Two sets of acoustic feature are evaluated, 31 features built from [1] and eGeMaps feature set [2]. As the baseline, dense neural network is used with the same number of layer and nodes. Two variants of RNN i.e. gated rated unit (GRU) and long short-term memory (LSTM) are evaluated for shared layers. The obtained result is measured in term of mean squared error (MSE), mean absolute percentage error (MAPE), mean absolute error (MAE) and concordance coefficient correlation (CCC) [3]. The first three metrics calculate the error (lower is better) for predicted and true emotion dimension, while the last one calculate agreement between true value and predicted emotion (higher is better). Among those metrics, the main focus is CCC as it is proposed in [4] for the emotion recognition evaluation.

2 Dataset and Acoustic Feature Set

We used "interactive emotional dyadic motion capture database" (IEMOCAP), collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). This dataset consists of multimodal measurement of speech and gesture including markers of the face, head, and hands, which provide detailed information about facial expressions and hand movements during dyadic conversation. Among those modalities only speech utterance is used. The total utterances is 10039 turns with three emotion attributes: arousal, valence and dominance. To extract acoustic features from the speech, two feature sets are used. The first set includes 31 acoustic features, the second is eGeMaps feature set. The descriptions of those two feature sets are explained below.

31 Features: Thirty-one of acoustic features were extracted from IEMOCAP speech dataset. Those feature are: 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. These 31 features are extracted for each frame with 20ms window and 10 ms stride. The total of frames for each utterance is limited to 100 frames. This feature set is derived from [1].

eGeMaps Feature: This is a feature set proposed by paper [2] for voice research and affective computing. There are 23 features in this set: 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. The total frames used for this feature set is the longest one i.e. 3409 frames.

3 RNN-based Emotion Recognition

The proposed system for dimensional emotion recognition consists of three shared recurrent neural network (RNN) layers. Figure 1 shows the architecture of the proposed system. The RNN layer is implemented either using gated rated unit (GRU) or long short-term memory (LSTM) networks. For the baseline, we choose traditional deep neural networks with three dense networks and flattened before connected to three output layers. For the RNN, we choose 64 nodes of each layer to close the total number of trainable parameters to dense layer. A batch normalization layer is added at input layer, and a dropout layer is added before the output layer with factor 0.4. The implementation of this architecture

*Corresponding author, Email: bagus@jaist.ac.jp

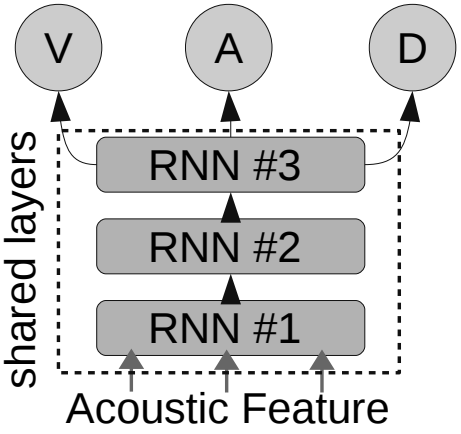

Fig. 1 RNN system for dimensional emotion recognition. RNN #1, #2, and #3 are either GRU or LSTM layer; V: valence, A: arousal, D: dominance.

is available in public repository[1] for more detail.

For each method, we run an experiment on 100 epochs. As the main target is CCC, then the used loss function is CCC loss (CCCL) which is formulated as follows,

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

$$CCCL = 1 - CCC \quad (2)$$

where ρ is Pearson correlation between predicted emotion degree x and true emotion degree y, σ^2 is a variance and μ is a mean. As the learning process minimizes three variables, we used the following multi-task learning approach to optimize CCC score.

$$CCCL_{tot} = \alpha CCCL_V + \beta CCCL_A + \gamma CCCL_D \quad (3)$$

where α, β, and γ are parameters for CCCL valence (V), arousal (A), and dominance (D). We set all parameters to 1 at initial experiment and find the optimum with $\alpha : 0.7$, $\beta : 0.3$, and $\gamma : 0.6$ by trials.

4 Results and Discussion

Table 1 shows result of speech emotion recognition performance among different methods and metrics for 100 epochs. Clearly, both GRU and LSTM performs better than dense deep neural network (DNN) for both acoustic feature sets. The highest CCC score by LSTM is **[0.11, 0.43, 0.36]** for [valence, arousal, dominance] obtained using earlystopping method. eGeMaps feature set obtained significantly higher performance than 31 acoustic features for both error and agreement measurement. The length of used frames for each feature set, where 31 features set used smaller frames, may affect this results. For future works, an extension to find the optimum multi-task learning parameter should be performed by using a such automatic method, e.g. grid search.

[1] https://github.com/bagustris/asj_autumn_2019/

Table 1 Results of dimensional emotion recognition among different method and metric; Each score is average scores from V, A, and D (e.g. Fig. 2).

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

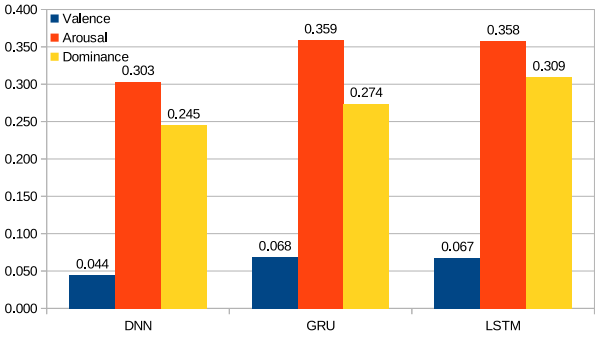

Fig. 2 CCC score for each emotion dimension.

References

[1] Theodoros Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis." PloS one 10, no. 12, 2015.
[2] F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set for Voice Research and Affective Computing." IEEE Trans. Affect. Comput., vol. 7, no. 2, pp. 190202, 2016.
[3] L. I-Kuei Lin, "A concordance correlation coefficient to evaluate reproducibility." Biometrics, 255-268, 1989.
[4] Fabien Ringeval et al., "AVEC 2019 Workshop and Challenge: State-of-Mind, Depression with AI, and Cross-Cultural Affect Recognition." AVEC '19, ACM Int'l. Conf., 2019.