ICPR 2020 Paper ID 2408

End-to-end Triplet Loss based Emotion Embedding for Speech Emotion Recognition

Puneet Kumar[†], Sidharth Jain[†], Balasubramanian Raman[†], Partha Pratim Roy[†] and Masakazu Iwamura[‡]

[†] Indian Institute of Technology, Roorkee, India.

[‡] Osaka Prefecture University, Sakai, Japan.

Outline

Introduction

- An end-to-end neural embedding system based on triplet-loss [1] and residual-learning [2] has been proposed for speech emotion recognition (SER).
- The proposed system learns the embeddings from the emotional information of the speech utterances. The embeddings are used for SER in the speech samples of various lengths.
- The proposed system is trained using softmax pre-training and triplet loss function. The weights between the fully connected and embedding layers of the trained network are used to calculate the embedding values.
- The embedding representations of various emotions are mapped onto a hyperplane, and the angles among them are computed using the cosine similarity. These angles are utilized to classify a new speech sample into its appropriate emotion class.

Introduction

Contributions

The contributions of the paper are as follows.

- A deep neural end-to-end SER system based on triplet loss and residual learning has been proposed. The proposed system is capable of learning emotion-related information from a labeled emotional speech dataset in the form of embeddings.
- The embeddings learned by the proposed system are used to classify the speech samples of various lengths into appropriate emotion classes. Using the embeddings, the proposed system can estimate the emotions in unseen speech utterances.

Literature Survey

- In recent years, several SER approaches have been developed. Feature-based speech recognition. For example, feature-based speech recognition systems [3], statistical SER methods [4], support vector machine based approaches. [5], neural SER models [6] and deep learning based methods [7].
- The existing SER approaches face challenges such as the need for manual crafting of acoustic features, bias towards the polarity of the emotional features in feature-based SER, and unreliability of statistical SER systems in estimating the parameters of the global speech features. The proposed system overcomes these issues.
- End-to-end SER using triplet loss and residual learning, along with deep neural networks, has not been explored to its full potential. With that as an inspiration, various state-of-the-art deep neural architectures have been implemented, and the best performing one is implemented in the proposed work.

Proposed System

Problem Formulation

Consider d-dimensional space $\mathbb{R}^d$ where elements in $\mathbb{R}^d$ are represented as $\{x_1, x_2, x_3, .., x_d\}$ where x_i is a d-tuple that denotes an embedding vector $f(x) \in \mathbb{R}^d$ mapped from a set of speech utterances $\sum_j y_j$. The projections of such embedding vectors are represented in a hyperplane where emotion similarity is measured using cosine similarity. The objectives of the proposed technique are to:

- (i) Learn the embeddings from input speech utterances,
- (ii) Visualize the embeddings projected in a hyperplane to analyze the learned emotion patterns,
- (iii) Use the learned embeddings to classify an unseen speech utterance into an appropriate emotion class.

Proposed System

Hypothesis

The core hypothesis of the proposed method is depicted in Fig. 1.

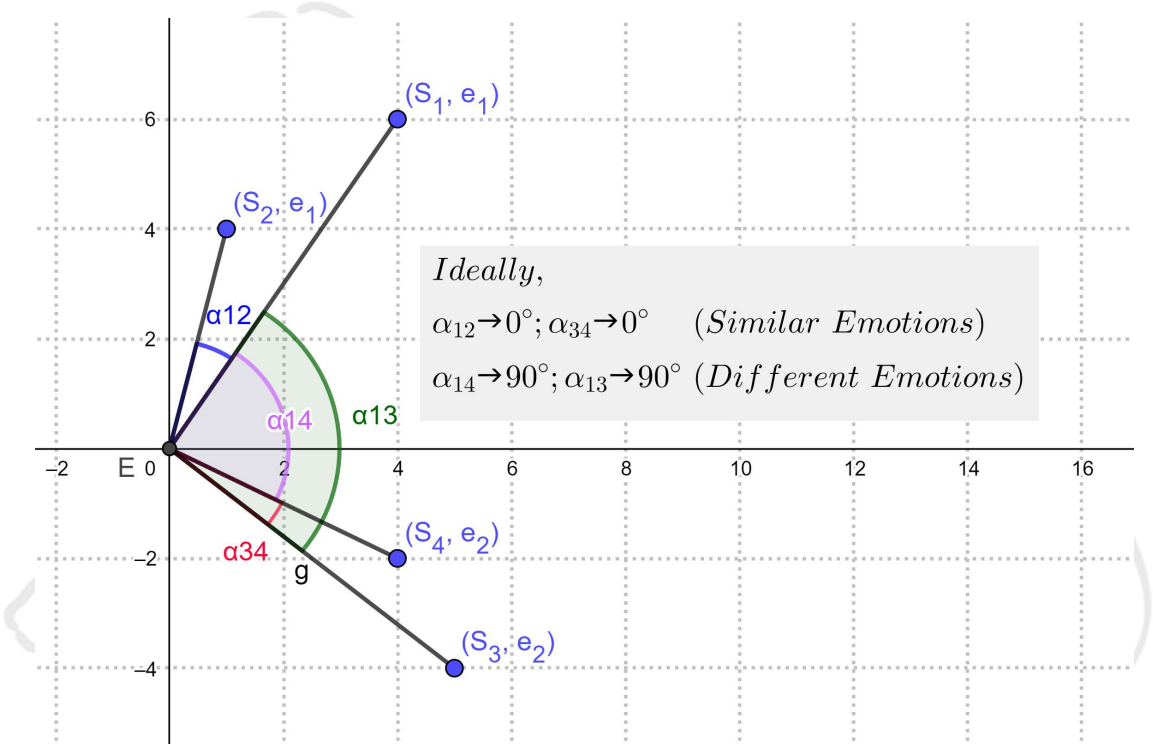

Figure 1: Hypothesis of the Speech Emotion Recognition

Proposed System

Methodolody

The proposed end-to-end system learns the embedding representations from emotional speech and uses them for speech emotion recognition. Various phases of the system and its architecture are described in Fig. 2 and Fig. 3 respectively.

Phase I: Initialization and Pre-processing - The embeddings are initialized and data is pre-processed.

Phase II: Embedding Training - A fully connected layer projects the utterance-level representations as embeddings. Emotion characteristics of the speech are learned by training embedding vectors for each emotion. The training process first carries out the softmax-pretraining and then performs embedding training with Triplet Loss.

Proposed System

Architecture

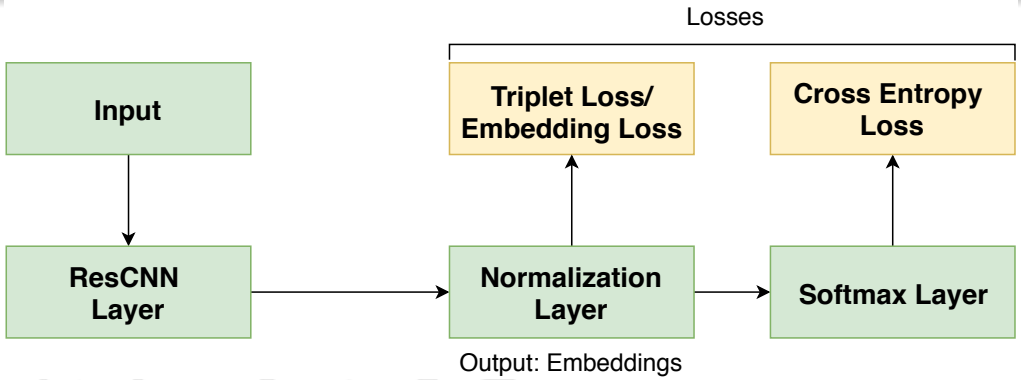

Figure 2: Proposed methodology for Speech Emotion Recognition

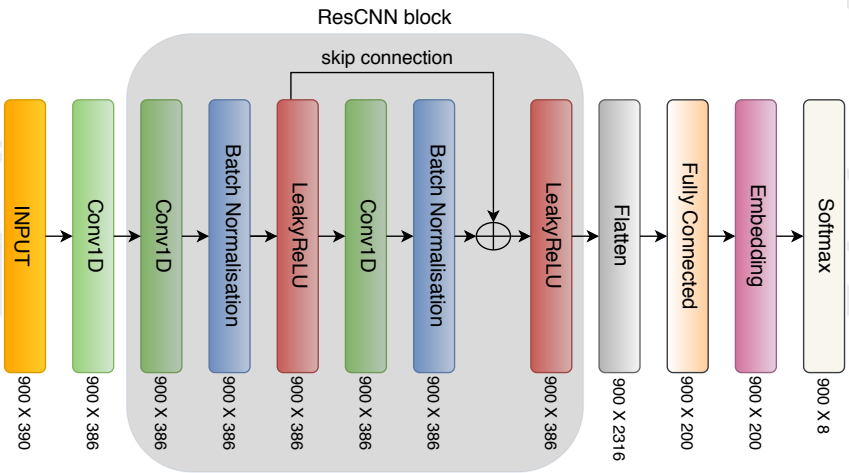

Figure 3: Proposed architecture

Experimental Setup

Experimental Platform & Dataset

- **Experimental Setup**: Model training has been performed on Nvidia RTX 2070 GPU with 2304 CUDA cores, 288 Tensor cores, and 8 GB Virtual RAM. Model testing has been carried out on Intel(R) Core(TM) i7-7700, 3.70 GHz, 16GB RAM CPU system with Ubuntu 18.04.
- **Dataset and Training Strategy**: SER experiments have been performed on *RAVDESS* and *IEMOCAP* datasets. Final implementation has been carried out using 70%-30% training-testing split and 10-fold cross-validation.

Experimental Setup

Ablation Study

The ablation study to choose the appropriate network architecture has been performed. Fully Connected (FC) network, CNN, Residual Neural Network (ResNet), RNN, Long Short Term Memory (LSTM) based RNN, and Gated Recurrent Unit (GRU) based RNN has been evaluated. Their details have been presented in Table 1. Here 'x' represents the total number of layers. The analysis is performed for x = 6 to 15.

Table 1: Summary of the ablation study for Speech Emotion Recognition

Architecture	**Details**	**x**	**Accuracy**
FC-x	Fully Connected network with x layers	7	47.22%
CNN-x	Convolutional Neural Network with x layers	8	58.33%
RNN-x	Simple x-layered Recurrent Neural Network	8	55.56%
LSTM-x	x-layered RNN with LSTM units	7	56.94%
GRU-x	x-layered RNN with GRU units	8	54.16%
ResNet-x	Residual Neural Network with x layers	11	**61.11%**

Results

Emotion Embeddings

The proposed approach showed an accuracy of 91.67% for RAVDESS dataset and 64.44% for IEMOCAP dataset. The emotion embeddings have been visualized in Fig. 4 and Emotion Classification results are presented in the upcoming slides.

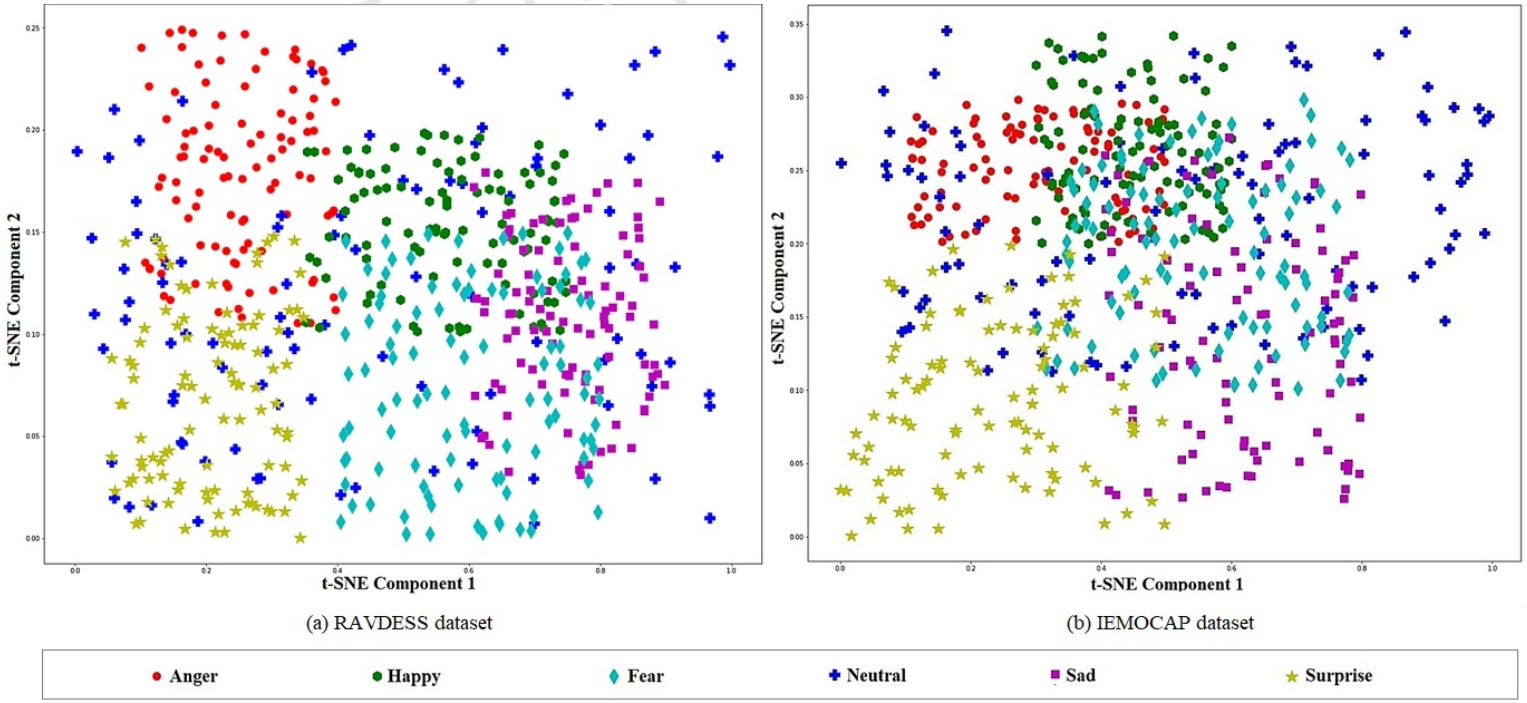

Figure 4: Visualization of Emotion Embeddings

Emotion Classification

Results for RAVDESS dataset

Table 2: RAVDESS dataset: Angles among emotional embedding vectors

	neutral	**calm**	**happy**	**sad**	**angry**	**fearful**	**disgust**	**surprise**
neutral	0.37°	30.75°	60.34°	89.78°	69.73°	62.63°	80.28°	77.27°
calm		0.51°	78.91°	65.01°	64.61°	66.37°	89.93°	71.71°
happy			0.03°	25.75°	83.22°	66.30°	62.76°	64.39°
sad				0.35°	60.95°	61.52°	63.69°	84.88°
angry					0.63°	84.97°	61.23°	83.40°
fearful						0.03°	67.89°	67.06°
disgust							1.64°	83.63°
surprise								0.50°

Table 3: Result comparison with state-of-the art methods

Method	**Author**	**Accuracy**
Proposed Method		**91.67%**
Convolutional Neural Network	M. G. Pinto [7]	91.53%
Artificial Neural Network	K. Tomba et al. [8]	89.16%
Multi Task Hierarichel SVM	B. Zhang et al. [9]	83.15%
Bagged Ensemble of SVMs	A. Bhavan et al. [10]	75.69%
Convolutional Neural Network	D. Issa et al. [11]	71.61%

Emotion Classification

Results for IEMOCAP dataset

Table 4: IEMOCAP dataset: Angles among emotional embedding vectors

	anger	sadness	happiness	neutral	excitement	surprise	fear	disgust	frustration
anger	4.51°	51.97°	78.08°	53.39°	56.82°	79.20°	52.68°	62.32°	84.85°
sadness		0.96°	87.23°	70.54°	29.96°	52.28°	67.51°	85.14°	73.09°
happiness			0.99°	67.23°	53.74°	60.81°	77.82°	50.09°	77.06°
neutral				1.12°	56.43°	72.48°	61.06°	45.56°	82.68°
excitement					0.87°	50.74°	68.97°	88.14°	39.29°
surprise						2.66°	83.84°	78.18°	77.65°
fear							0.77°	32.51°	83.56°
disgust								0.73°	73.84°
frustration									2.44°

Table 5: Result comparison with state-of-the art methods

Method	Author	Accuracy
RNN + Attention	N. Majunder [12]	64.50%
Proposed Method		**64.44%**
Memory Network	D. Hazarika et al. [13]	63.50%
CNN + Mel Filterbanks	Z. Aldeneh and E. Provost [14]	61.80%
Memory Network	S. Poria et al. [15]	56.13%
CNN + LSTM	J. Zhao [16]	52.14%

Conclusion

- An end-to-end emotion embedding system has been proposed to learn the emotional patterns from speech in the form of an embedding matrix.
- The emotion embedding matrix thus prepared has been used for speech emotion recognition, and it demonstrated comparable recognition results to the state-of-the-art methods.
- It is required to check the angles for each speech utterance with each emotion class. This process can be optimized to reduce computational requirements.
- It is also aimed to use the learned embeddings for other speech processing tasks such as emotional speech synthesis.

Key References I

[1] Florian Schroff, Dmitry Kalenichenko, and James Philbin.
"FaceNet:Unified Embedding for Face Recognition and Clustering".
In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
"Deep Residual Learning for Image Recognition".
In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[3] Chul Min Lee, Shrikanth S Narayanan, et al.
Toward detecting emotions in spoken dialogs.
IEEE Transactions on Speech and Audio Processing, 13(2):293–303, 2005.

[4] Jaime Lorenzo-Trueba, Roberto Barra-Chicote, Ruben San-Segundo, Javier Ferreiros, Junichi Yamagishi, and Juan M Montero.
Emotion transplantation through adaptation in hmm-based speech synthesis.
Computer Speech & Language, 34(1):292–307, 2015.

[5] Udit Jain, Karan Nathani, Nersisson Ruban, Alex Noel Joseph Raj, Zhemin Zhuang, and Vijayalakshmi GV Mahesh.
Cubic svm classifier based feature extraction and emotion detection from speech signals.
In *2018 International Conference on Sensor Networks and Signal Processing (SNSP)*, pages 386–391. IEEE, 2018.

[6] André Stuhlsatz, Christine Meyer, Florian Eyben, Thomas Zielke, Günter Meier, and Björn Schuller.
"Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks".
In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5688–5691, 2011.

[7] Marco Giuseppe de Pinto.
Emotion classification RAVDESS, 2019.

[8] Kevin Tomba, Joel Dumoulin, Elena Mugellini, Omar Abou Khaled, and Salah Hawila.
Stress detection through speech analysis.
In *International Joint Conference on e-Business and Telecommunications (ICETE)*, pages 560–564, 2018.

Key References II

[9] Biqiao Zhang, Georg Essl, and Emily Mower Provost.
Recognizing emotion from singing and speaking using shared models.
In *IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 139–145, 2015.

[10] Anjali Bhavan, Pankaj Chauhan, Rajiv Ratn Shah, et al.
Bagged support vector machines for emotion recognition from speech.
Knowledge-Based Systems, 184:104886, 2019.

[11] Dias Issa, M Fatih Demirci, and Adnan Yazici.
Speech emotion recognition with deep convolutional neural networks.
Biomedical Signal Processing and Control, 59:101894, 2020.

[12] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria.
Dialoguernn: An attentive rnn for emotion detection in conversations.
In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, volume 33, pages 6818–6825, 2019.

[13] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann.
Icon: interactive conversational memory network for multimodal emotion detection.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2594–2604, 2018.

[14] Zakaria Aldeneh and Emily Mower Provost.
Using regional saliency for speech emotion recognition.
In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2741–2745, 2017.

[15] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann.
Conversational memory network for emotion recognition in dyadic dialogue videos.
In *Conference of the North American Chapter of the Association for Computational Linguistics (ACL)*, pages 2122–2132, 2018.

[16] Jianfeng Zhao, Xia Mao, and Lijiang Chen.
Speech emotion recognition using deep 1D & 2D CNN & LSTM networks.
Biomedical Signal Processing and Control, 47:312–323, 2019.

Thank You.