THE UNIVERSITY OF WARWICK

Compact Graph Architecture for Speech Emotion Recognition

Amir Shirian   Tanaya Guha

University of Warwick, UK

ICASSP 2021 TORONTO Canada

June 6-11, 2021

Metro Toronto Convention Centre

Motivation

- Pervasive use of speech emotion recognition in many human-centric systems, such as behavioral health monitoring and empathetic conversational systems.
- Why modeling speech with graph?
  - Graph is a compact, efficient, and scalable way to represent data.
  - The temporal and spatial information can be coded into a graph.
- Modeling all samples with the same graph structure leads to a lot fewer number of trainable parameters in comparison with the recurrent models.

Contribution

- First work that takes a graph classification approach to SER.
- Leveraging accurate graph convolution, we obtain the state-of-the-art results on **IEMOCAP** and **MSP-IMPROV** databases.
- Our model has significantly fewer trainable parameters ($\sim$30K only) with better performance.

Problem Formulation

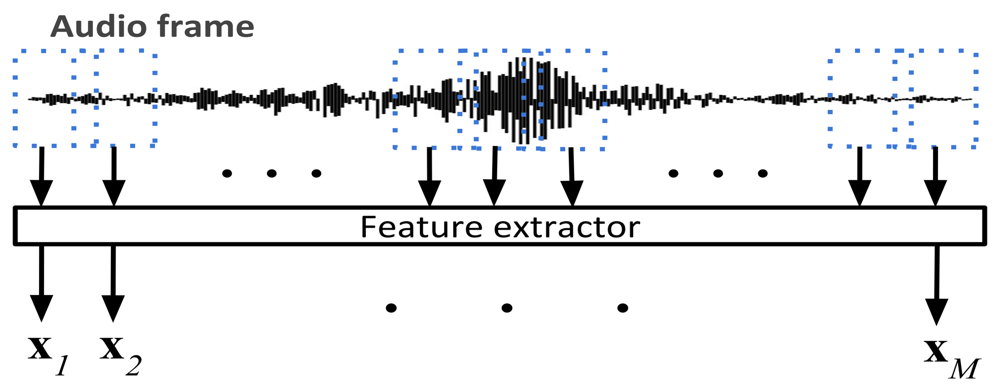
- Given:
  - Speech graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
  - Data graph specified by the adjacency matrix $\mathbf{W} \in \{\mathbf{A}_c, \mathbf{A}_l\}$

$$\mathbf{A}_c = \begin{bmatrix} 0 & 1 & 0 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 1 & 0 \end{bmatrix} \quad \mathbf{A}_l = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

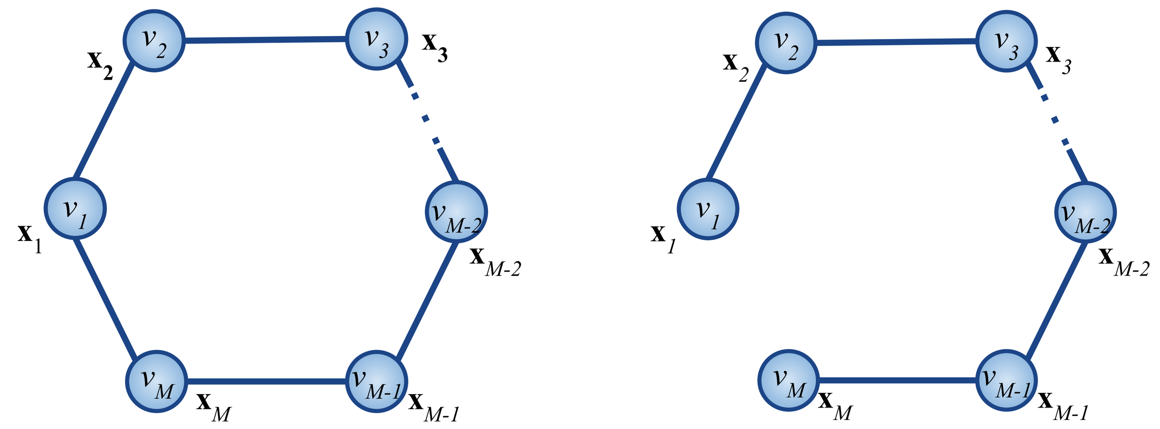
  - Each graph is associated with label $\mathbf{y}_i$
- Goal:
  - We want to predict the emotion related to the speech graph

Graph Construction

- It's a frame to node transformation.



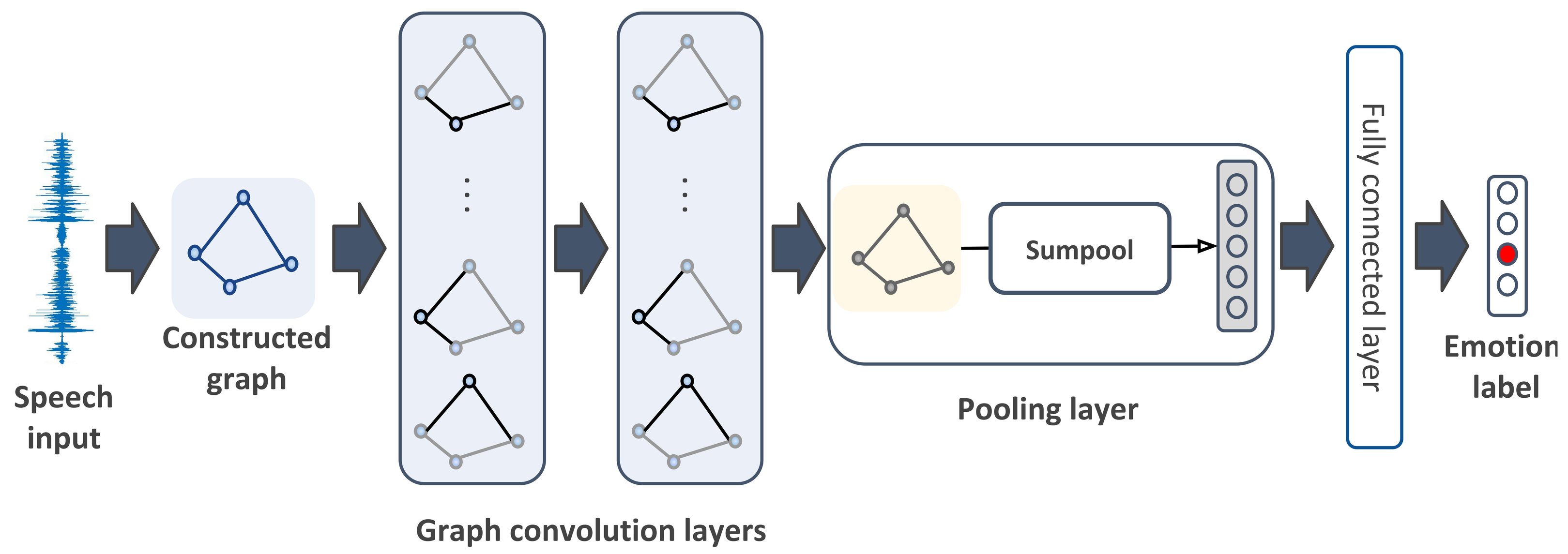
- LLD features were extracted from M frames (short, overlapping segments).



- Each of these M frames are associated with a node in a graph.
- Either a *cycle* ($\mathbf{A}_c$) or *line* ($\mathbf{A}_l$) structure is selected manually for the graph.

Model

The overview of our proposed graph-based architecture for SER



- Gives constructed graph as input.
- Produces node embedding with two graph convolution layers
- Produces graph embedding with a pooling function.

Results

Compare with SOTA and graph baselines

Results on MSP-IMPROV

Model	WA (%)	UA (%)
Graph baselines		
GCN	54.71	51.42
PATCHY-SAN	55.47	52.33
PATCHY-Diff	56.18	53.12
SER models		
ProgNet 2017	**58.40**	-
CNN 2019	50.84	-
LSTM 2019	51.21	-
CNN-LSTM 2019	52.36	-
Ours (cycle)	**57.82**	**55.42**
Ours (line)	57.08	54.75
Ours (cycle w/o MLP)	56.82	53.22

Results on IEMOCAP

Model	WA (%)	UA (%)
Graph baselines		
GCN	56.14	52.36
PATCHY-SAN	60.34	56.27
PATCHY-Diff	63.23	58.71
SER models		
Attn-BLSTM 2016	59.33	49.96
BLR 2017	62.54	57.85
RNN 2017	63.50	58.80
CRNN 2018	63.98	60.35
SegCNN 2019	64.53	**62.34**
LSTM 2019	58.72	-
CNN-LSTM 2019	59.23	-
Ours (cycle)	**65.29**	**62.27**
Ours (line)	64.69	61.14
Ours (cycle w/o MLP)	64.19	60.31

Model size comparison

GCN	PTCHY-SAN	PTCHY-Diff	BLSTM	Ours
$\sim$76K	$\sim$60K	$\sim$68K	$\sim$0.8M	$\sim$**30K**

Conclusions

- First graph-based approach to SER.
- We transformed speech utterances to graphs with simple structures that largely simplify the convolution operation on graphs.
- Defining the same structure for samples leads to a light-weight GCN architecture which outperforms LSTMs, standard GCNs and several other recent graph models in SER.

GitHub link: github.com/AmirSh15/Compact_SER

Contact: amir.shirian@warwick.ac.uk

SIPLab