Faculty Mentor: Dr. Carlos Busso
Student Mentor: Wei-Cheng Lin
Program: Electrical Engineering

Creating an Attention Based Network for Speech Emotion Recognition in Conversations

Luz Martinez-Lucas
Luz.Martinez-Lucas@utdallas.edu

UT Dallas MSP Multimodal Signal Processing Laboratory
THE UNIVERSITY OF TEXAS AT DALLAS
Erik Jonsson School of Engineering & Computer Science at the University of Texas at Dallas, Richardson, Texas 75080, USA

2020 Jonsson School UG Research Award Recipient

FEARLESS engineering™

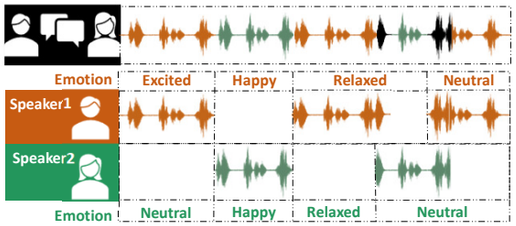
Motivation

Background:

- *Speech Emotion Recognition* (SER) models often use only the speech information of the current time step
  - Emotions often depend on previous speech information as well as the emotional context of the scene

Speakers in conversations influence the emotional state of other speakers

Emotion Mimicking in a Conversation



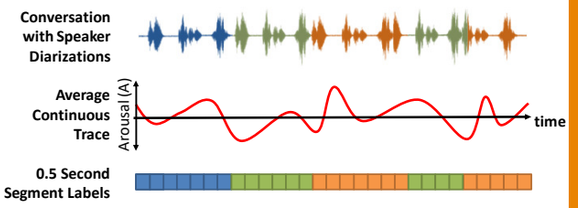
Our Work:

- Utilizes current and previous speech as well as surrounding speaker context to train an SER Model for conversations

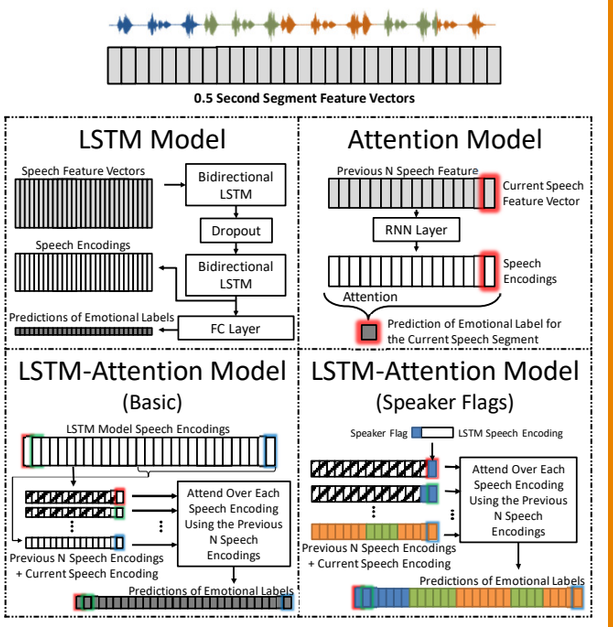
MSP-Conversation

Emotional corpus collected at UT-Dallas

- Conversations: 74 audio clips (10-20 min) taken from podcasts
- Continuous annotations of three emotional attributes: arousal, valence, and dominance
- The mean traces are averaged over 0.5 second segments to obtain segment labels
- The segment is labeled with the speaker that is active the longest during that speech turn
  - If two or more speakers are active during the entire segment, priority is given to the speaker "speaking" last



SER Models



Current Results

Training Parameters:

- 6,373 acoustic features extracted with openSMILE
- All models trained on arousal, valence, and dominance labels
- The CCC loss is used for the training loss
- The CCC's of the test conversations are averaged for the evaluation metric
- Employ early stopping over the validation set during training
- For the attention mechanism: use 30 previous turns (N=30)
  - Zero padding for speech segments that do not have N previous segments
- The speaker flags are the speaker numbers (e.g., 0, 1 ,2)
  - If the segment is more than 50% silence or other noises, the flag is set to -1

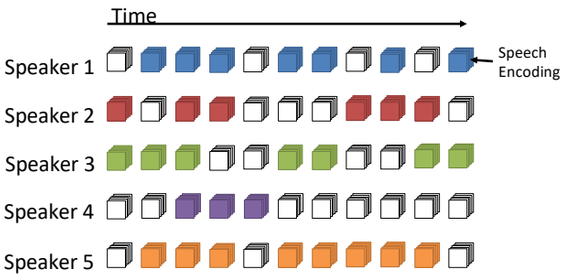
Mean CCC of the Test Conversations for the Best Model According to the Validation Set

Model (N = 30)	Emotional Attribute		
	Arousal	Valence	Dominance
LSTM	0.710	0.281	0.716
Attention	0.334	0.136	0.247
LSTM-Attention (B)	**0.744**	**0.307**	0.725
LSTM-Attention (SF)	0.740	0.265	**0.742**

- The best models are the LSTM-Attention models
- The addition of the speaker flags seems to confuse the models trained on arousal and valence but aid the dominance model

Future Work

- Speaker flags do not add enough speaker context
- Create 3D map of the speech segment encodings belonging to each speaker at each time step:



- We can use a Convolutional Neural Network and Attention to introduce the speaker context to the emotional inferences