Transfer Learning for Emotion Recognition

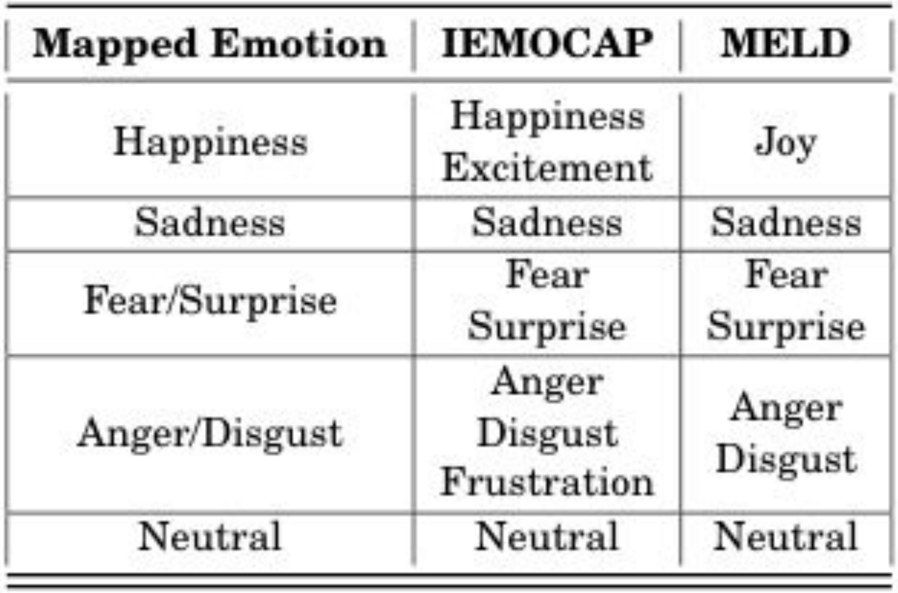
Jessica Huynh, Amith Ananthram, Kailash Karthik
jyh2127, aa4461, ks3740

Problem Description

- **problem:** given an utterance, classify its emotion into one of {happiness, sadness, fear/surprise, anger/disgust, neutral}

- **approach:** transfer learning from larger speaker identification corpus to generate MFCC + pitch based speech embeddings; combine with BERT-based text embeddings; neural classification, LDA/PLDA

- **training:** VoxCeleb1 and 2, MELD
- **test:** IEMOCAP

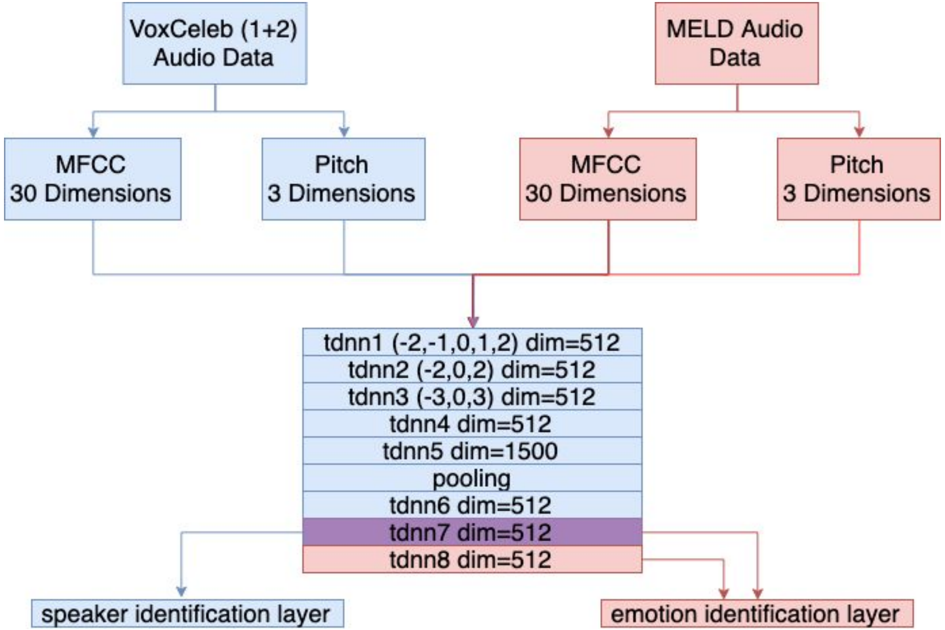


Mapped Emotion	IEMOCAP	MELD
Happiness	Happiness Excitement	Joy
Sadness	Sadness	Sadness
Fear/Surprise	Fear Surprise	Fear Surprise
Anger/Disgust	Anger Disgust Frustration	Anger Disgust
Neutral	Neutral	Neutral

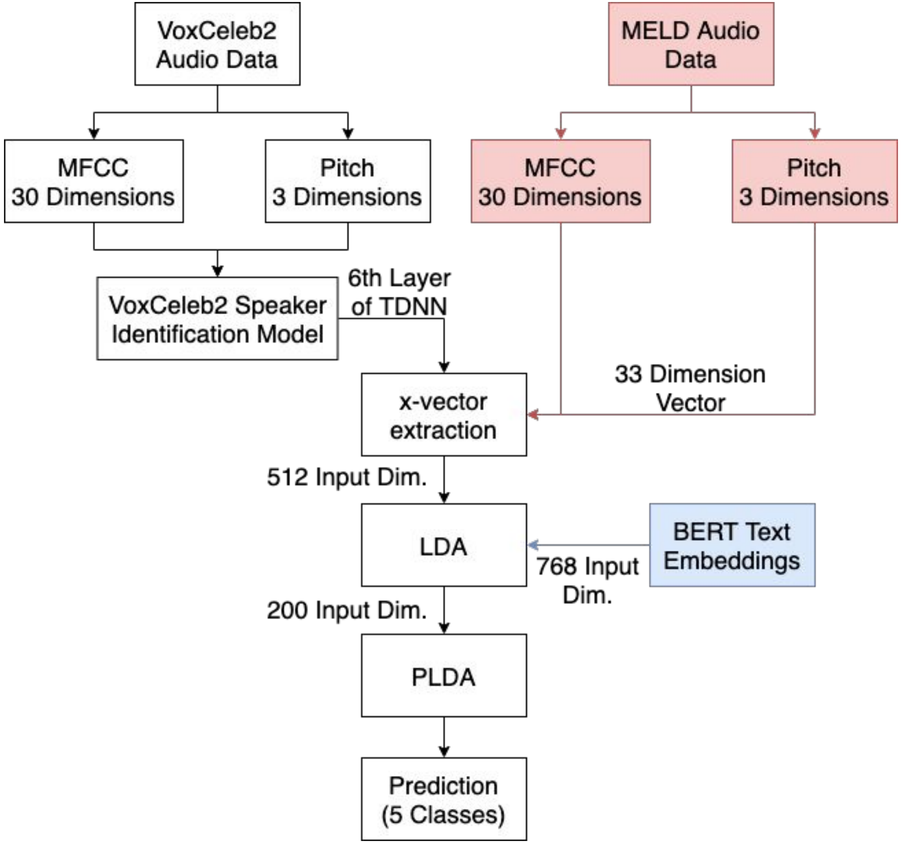
Table 1. Emotion Label Mapping

Approach

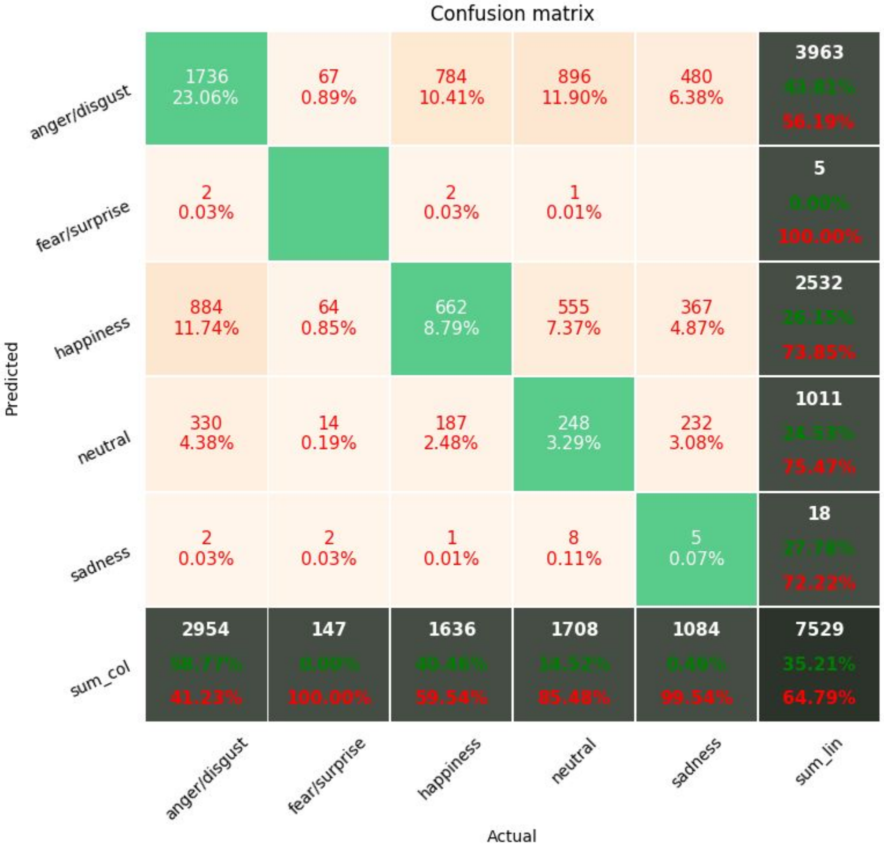
Network Training



LDA/PLDA Training



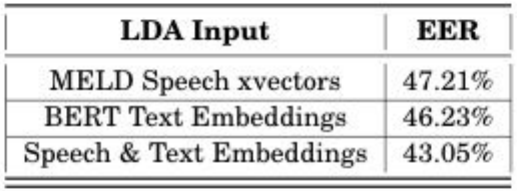
Neural Network Results



Confusion matrix

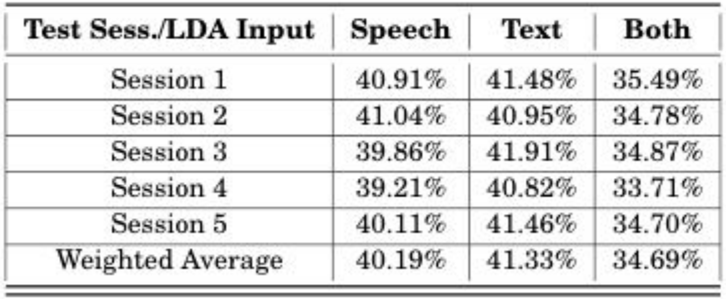
Predicted \ Actual	anger/disgust	fear/surprise	happiness	neutral	sadness	sum_lin
anger/disgust	1736 23.06%	67 0.89%	784 10.41%	896 11.90%	480 6.38%	3963 43.81% 56.19%
fear/surprise	2 0.03%		2 0.03%	1 0.01%		5 0.00% 100.00%
happiness	884 11.74%	64 0.85%	662 8.79%	555 7.37%	367 4.87%	2532 26.15% 73.85%
neutral	330 4.38%	14 0.19%	187 2.48%	248 3.29%	232 3.08%	1011 24.53% 75.47%
sadness	2 0.03%	2 0.03%	1 0.01%	8 0.11%	5 0.07%	18 27.78% 72.22%
sum_col	2954 58.77% 41.23%	147 0.00% 100.00%	1636 40.46% 59.54%	1708 14.52% 85.48%	1084 0.46% 99.54%	7529 35.21% 64.79%

LDA/PLDA Results



LDA Input	EER
MELD Speech xvectors	47.21%
BERT Text Embeddings	46.23%
Speech & Text Embeddings	43.05%

Table 3. EER On All IEMOCAP



Test Sess./LDA Input	Speech	Text	Both
Session 1	40.91%	41.48%	35.49%
Session 2	41.04%	40.95%	34.78%
Session 3	39.86%	41.91%	34.87%
Session 4	39.21%	40.82%	33.71%
Session 5	40.11%	41.46%	34.70%
Weighted Average	40.19%	41.33%	34.69%

Table 4. EER On Fifths Of IEMOCAP

Challenges

1. cross-domain evaluation (train on VOX/MELD, evaluate on IEMOCAP)
2. partial signal: human annotators used visual cues + audio cues
3. imbalanced corpora: not all emotions are represented equally
4. context independent combination of text + speech vectors
    a. we generate context dependent dependings of text and speech separately and combine them

Future Work

1. neural network
   a. reference model
      i. train speaker identification model with additional layers
      ii. train speaker identification model with silence frames
   b. domain adaptation
      i. include portions of IEMOCAP in re-training (in-progress)
   c. emotion conversion
      i. using the emotion detector, train an auto-encoder (in-progress)
2. target emotions (experiment with other clusterings)
3. model entire conversations for latent emotional state
4. LDA/PLDA
   a. context dependent combinations of speech + text embeddings
   b. Use embeddings from 8th layer of neural network for LDA/PLDA (in-progress)