Combining relative and absolute learning formulations to predict emotional attributes from speech

Abinay Reddy Naini, Shruthi Subramanium, Seong-Gyun Leem, Carlos Busso

UT Dallas MSP Multimodal Signal Processing Laboratory

THE UNIVERSITY OF TEXAS AT DALLAS

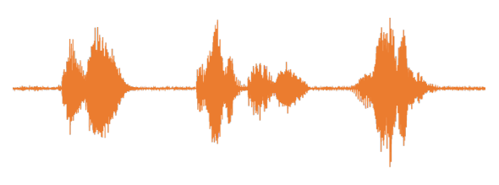
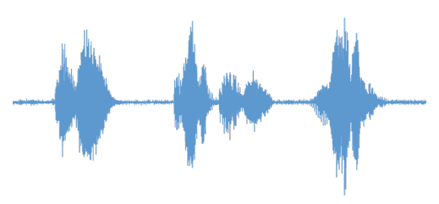
Erik Jonsson School of Engineering & Computer Science at the University of Texas at Dallas, Richardson, Texas 75080, USA

2023 ASRU Workshop on Automatic Speech Recognition and Understanding

Motivation

Background:

- Ordinal representations are more appropriate for emotional tasks (e.g., preference learning)

 <<  (e.g., more positive)

- Many applications require absolute emotional predictions
- Challenge:
  - Obtaining an absolute emotional label from a typical ordinal representation (preference learning in this case)

Our Work:

- A novel formulation that combines preference learning and regression formulations using multitask learning (MTL)

Emotional Corpus

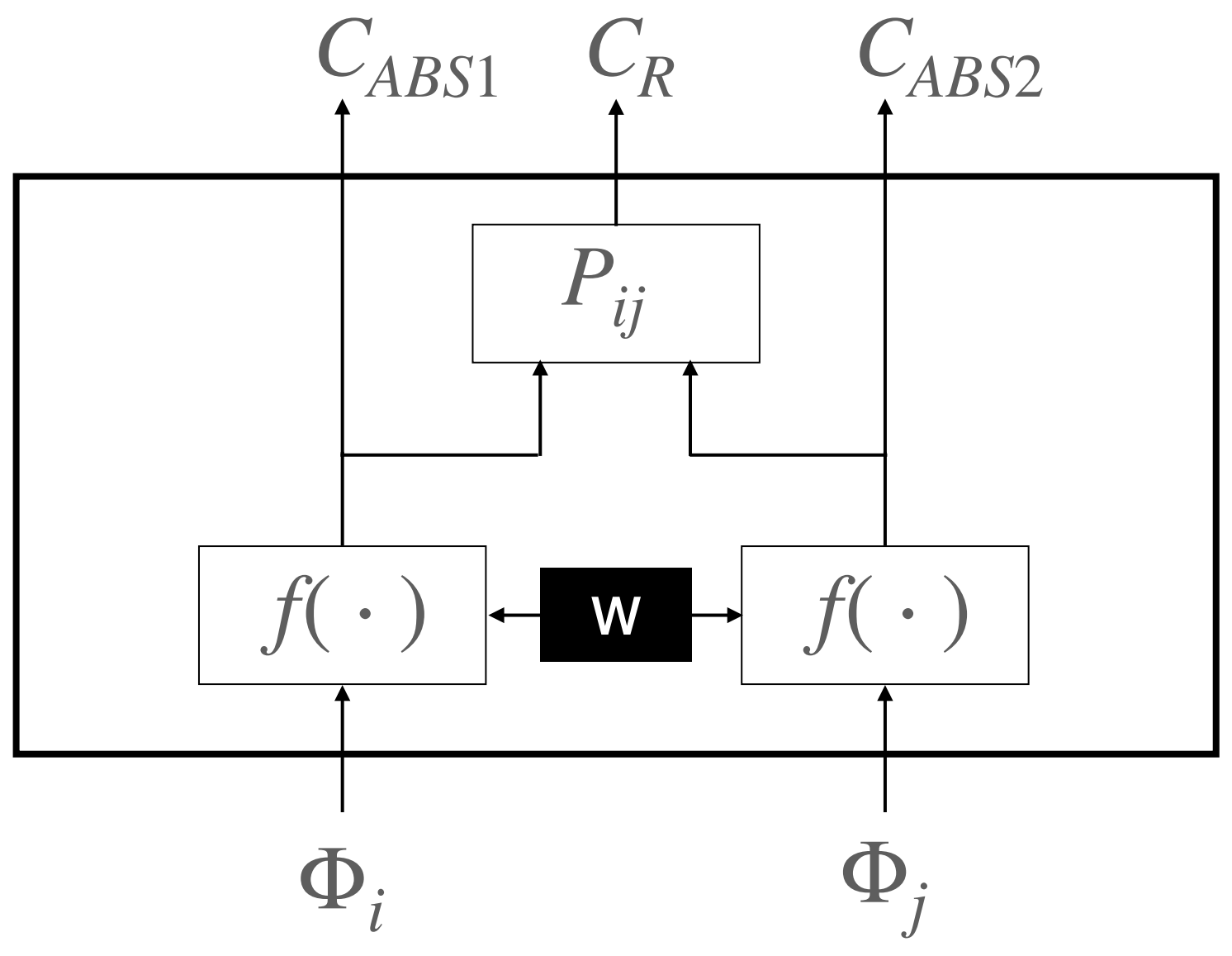
The MSP-Podcast corpus (v1.10)

- Naturalist data sourced from various audio-sharing websites with Creative Commons licenses
  - Train set: 63,076 speech segments
  - Development set: 10,999 speech segments
  - Test set: 16,903 speech segments
- We use emotional attributes
  - Arousal, valence, and dominance

Features

- We use the pre-trained Wav2vec2-large-robust2 model [1] from the HuggingFace library
  - We pruned the top 12 transformer blocks and fine-tuned the rest of the blocks with the train set
  - Sentence-level representation is obtained with the average pooled vector across all frames

Proposed multi-task framework (MTL)



Training

- Model trained with a pair of sentences at a time, with three set of labels:
  - Preference label
  - Two absolute scores for the samples
- C_{ABS} losses force the $f(\cdot)$ output to be a normalized score between 0,1 indicating the absolute emotional label

$$\mathcal{C} = \alpha_{PL}\mathcal{C}_{\mathcal{R}} + (\alpha_{ABS})(\frac{1}{2}\mathcal{C}_{ABS1} + \frac{1}{2}\mathcal{C}_{ABS2})$$

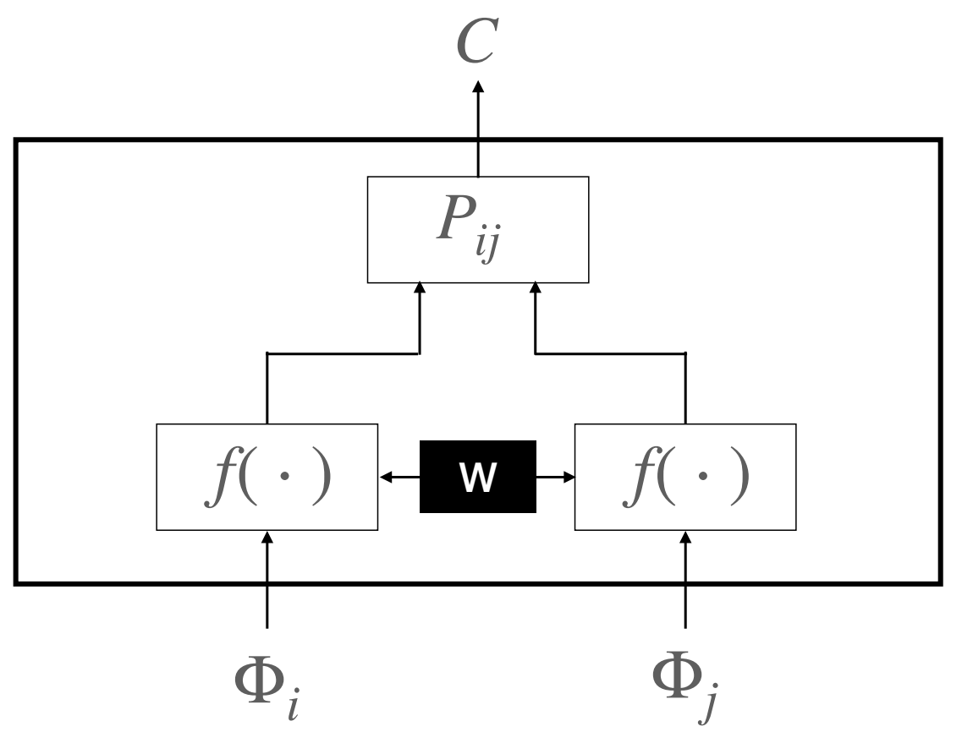
Inference

- $f(\cdot)$ produces an emotional score that can predict preference rank and absolute attribute labels

C_{ABS} = CCC loss
C_R= Rank-Net cost

Single-task Frameworks

RankNet Framework for Preference Learning (PL) [2]

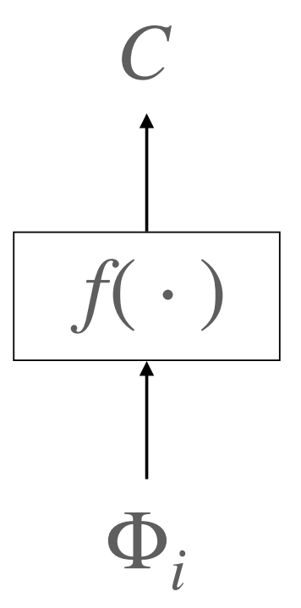


- This study relies on the RankNet-based implementation for preference learning
- $f(\cdot)$ has two fully connected layers

$$\boxed{P_{ij} = \frac{1}{1+e^{-\sigma(e_i - e_j)}}}$$

$$\boxed{\mathcal{C}_{\mathcal{R}} = -\bar{P}_{ij}\log P_{ij} - (1-\bar{P}_{ij})\log(1-P_{ij})}$$

Framework for absolute score (ABS)



- A similar function $f(\cdot)$ is trained to predict the absolute attribute score

Ordinal labels:

- QA (Qualitative Agreement): Preference labels obtaining using the QA method
  - Trained using randomly selected 200K pairs

	Sentence 1	Sentence 2
Rater 1	3.0	2.0
Rater 2	4.0	2.0
Rater 3	3.0	3.0
Rater 4	5.0	3.0
Rater 5	-	4.0

		Sentence 2				
		R1	R2	R3	R4	R5
Sentence 1	R1	↑	↑	=	=	↓
	R2	↑	↑	↑	↑	=
	R3	↑	↑	=	=	↓
	R4	↑	↑	↑	↑	↑

Qualitative Agreement

Performance Analysis for Speech Emotion Recognition

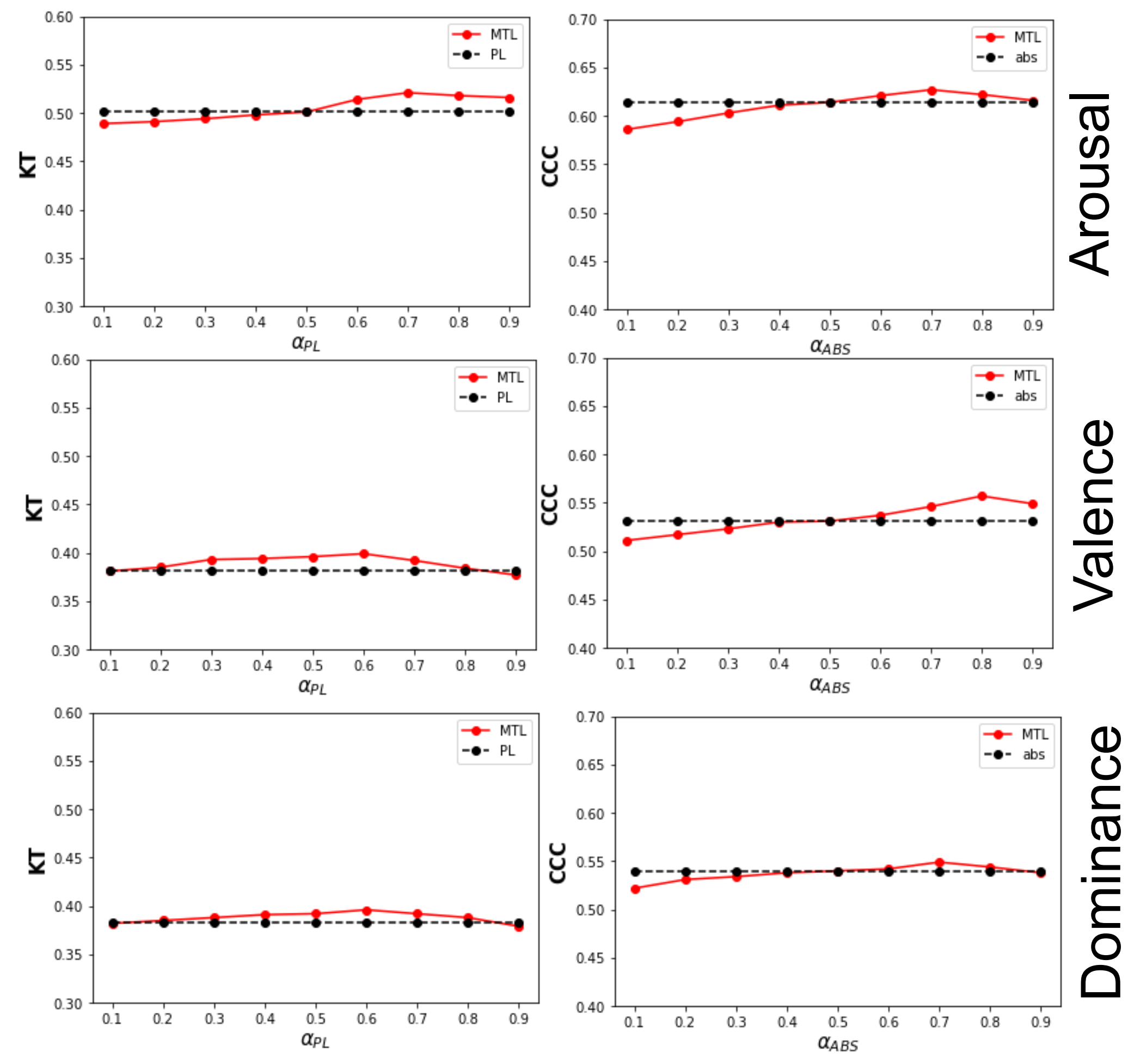
Experimental Results

- We consistently observed the best performance for each task when the corresponding weight is above 0.5 but less than 1
- The proposed MTL formulation performs either significantly better or similarly compared to the PL and ABS models
- The model obtains significantly better performance in each case by setting weights for ABS, PL in Case-2,3

KT: Kendall's Tau coefficient

Case-1	Arousal	Valence	Dominance
(α_{PL}, α_{ABS})	(0.45,0.55)	(0.3,0.7)	(0.4,0.6)
MTL (KT)	**0.507**	**0.393***	**0.391***
PL (KT)	0.502	0.381	0.383
MTL (CCC)	**0.619**	**0.546***	**0.542**
ABS (CCC)	0.614	0.531	0.540

	Case-2,3	Arousal	Valence	Dominance
Weights optimized for PL task	(α_{PL}, α_{ABS})	**(0.7,0.3)**	**(0.6,0.4)**	**(0.6,0.4)**
	MTL (KT)	**0.521***	**0.399***	**0.396***
	PL (KT)	0.502	0.381	0.383
Weights optimized for ABS task	(α_{PL}, α_{ABS})	**(0.3,0.7)**	**(0.2,0.8)**	**(0.3,0.7)**
	MTL (CCC)	**0.627***	**0.557***	**0.549***
	ABS (CCC)	0.614	0.531	0.540



Conclusions

- We proposed a novel Multi-task framework
  - Preserves the relative preference while predicting the absolute emotional score
- Explored the tradeoff between both the absolute and ordinal predictions in the proposed MTL framework

Future Work

- Explore alternative objective functions that will improve the performance of both tasks
- Explore strategies to estimate relative labels from absolute labels

References:

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representa- tions," (NeurIPS 2020)

[2] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," (ICML 2005)

This work was supported by Laboratory for Analytic Sciences