Preference Learning Labels by Anchoring on Consecutive Annotations

Abinay Reddy Naini, Ali N. Salman, Carlos Busso

UT Dallas MSP Multimodal Signal Processing Laboratory

UTD THE UNIVERSITY OF TEXAS AT DALLAS

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

INTERSPEECH 2023

20th-24th August 2023 - Dublin, Ireland

MOTIVATION

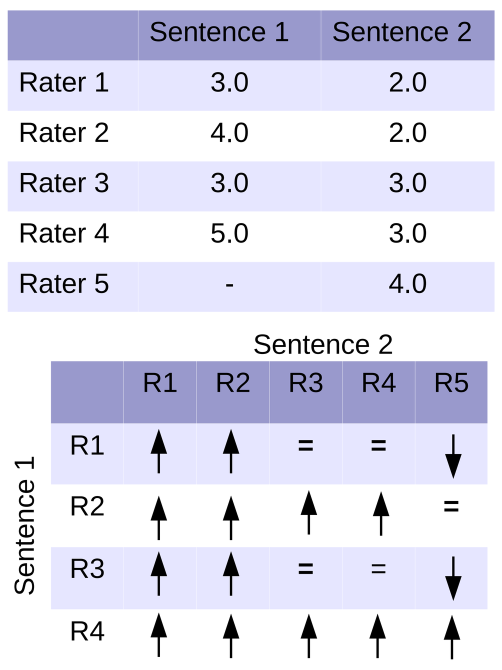
Background:

- An important task in human-computer interaction is to rank speech samples according to their expressive content.
- Preference learning framework is appropriate for obtaining emotional rank order for a set of speech samples.
- Challenge:
  - Obtaining reliable preference labels indicating preference between pair of speech samples

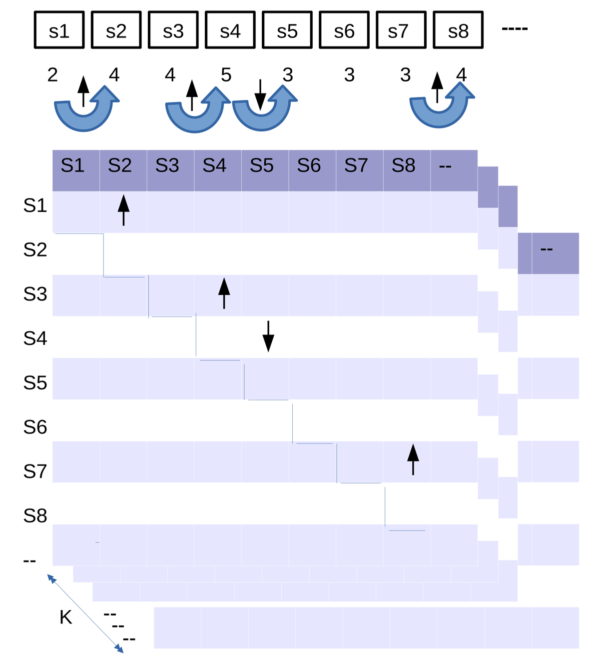
Our Work:

- We propose a method to obtain preference labels
  - That leverage the anchoring process by considering preference across consecutive annotations
  - The proposed consecutive labels (CL) are result of annotation trends assigned by a rater to consecutive samples
  - Achieved better performance with sparse set of relative labels

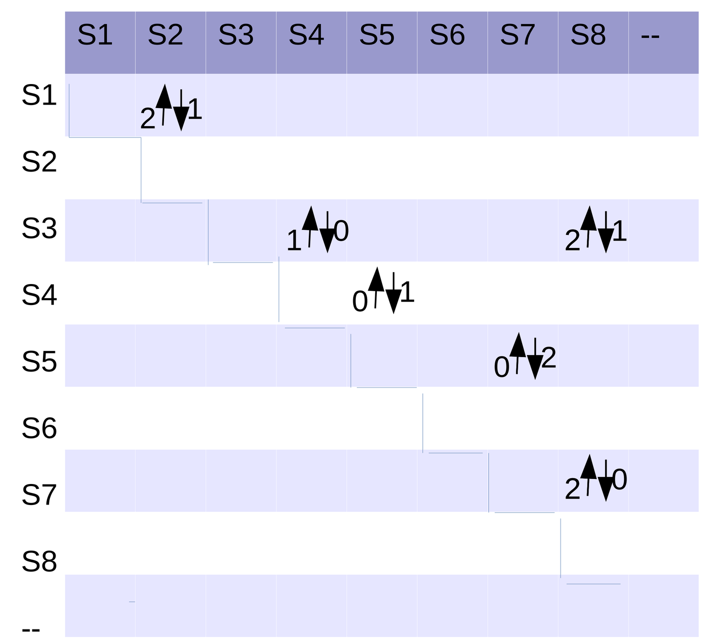
Proposed Anchor-based Ordinal Labels



Qualitative Agreement (Baseline)



individual CL matrix of annotators



Cumulative CL matrix

Step1:

- Obtained individual CL matrices for all k annotators indicating their consecutive preferences

Step2:

- Combined all individual CL matrices to obtain cumulative CL matrix.

CL1: All pairs

CL2: Pairs preferred by at least two annotators

CL3: Pairs preferred by at least three annotators

	CL1			CL2			CL3		
	A	V	D	A	V	D	A	V	D
Num. of Pairs in[K]	204	193	174	43	47	39	21	21	19
Coverage (%)	96.5	95.6	96.8	85.2	88.5	82.7	61.3	64.6	59.8

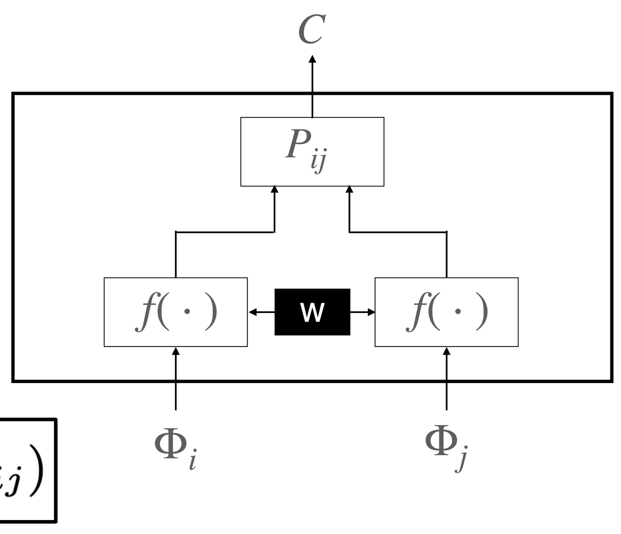
A: Arousal, V: Valence, D: Dominance, out of ~63K training samples

Preference Learning Framework and Performance Analysis for Speech Emotion Recognition

RankNet Framework[1]

This study relies on the RankNet-based implementation for preference learning

$$P_{ij} = \frac{1}{1+e^{-\sigma(s_i - s_j)}}$$



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

Baseline ordinal labels:

- ABS (Baseline): Preference labels obtained using a difference between consensus score. Trained using all possible pairs.
- QA (Qualitative Agreement): Preference labels obtaining using QA method. Trained using randomly selected 200K pairs.

Experimental Results

KT: Kendall's Tau coefficient

Trained using (rows) / Tested using (columns)

KT	ABS	QA	CL1	CL2	CL3
	Arousal				
ABS	0.482	0.496	0.489	0.494	0.497
QA	0.491	0.512	0.481	0.486	0.485
CL1	0.501	0.521	0.526	0.534	0.535
CL2	**0.504**	**0.527**	**0.533**	**0.539**	0.537
CL3	0.498	0.513	0.518	0.535	**0.539**

KT	ABS	QA	CL1	CL2	CL3
	Valence				
ABS	0.301	0.292	0.284	0.289	0.292
QA	0.311	0.316	0.298	0.304	0.302
CL1	0.308	0.321	0.331	0.334	0.331
CL2	**0.315**	**0.330**	0.346	0.348	0.341
CL3	0.314	0.329	**0.349**	**0.351**	**0.345**

KT	ABS	QA	CL1	CL2	CL3
	Dominance				
ABS	0.380	0.376	0.364	0.369	0.373
QA	0.388	0.393	0.382	0.393	0.397
CL1	**0.398**	0.395	0.417	0.419	0.426
CL2	0.395	0.402	**0.428**	**0.430**	0.424
CL3	0.389	**0.406**	0.416	0.426	**0.432**

- ➢ Proposed CL resulted in a better performance, even on the test sets obtained using baseline label methods (ABS, QA).
- ➢ Among the proposed CL schemes, CL2 performed best for arousal, and dominance.
- ➢ CL3 leads to better performance in valence.

Corpora

The MSP-PODCAST corpus (Emotional corpus collected at UT Dallas)

- We used 1.10 version of the corpus, which is sourced from various audio-sharing websites with creative commons licenses
- Includes 63,076 segments of audio for training, (10,999, and 16,903) segments for development and testing
- We have only used attributes (arousal, valence, and dominance) labels in this work

Features

- We used pre-trained wav2vec2-large-robust[2] model from the HuggingFace library. We pruned top 12 transformer blocks and fine-tuned with MSP-PODCAST train set
- Considered average pooled vector across all frames as the sentence level representation

CONCLUSIONS

- Considered ordinal labels using consecutive annotations from annotators, resulting in less noisy and reliable labels.
- Explored trade-off between quality and quantity in the implementation of the proposed ordinal labels.

Future Work

- In the future, we want to explore similar strategies to deal with ordinal labels for categorical emotions.

References:
[1] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," (ICML 2005)
[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representa- tions," (NeurIPS 2020)

This work was supported by NSF under Grant CNS-2016719