# Preference-Learning with Qualitative Agreement for Sentence Level Emotional Annotations


UT Dallas MSP Multimodal Signal Processing Laboratory

THE UNIVERSITY OF TEXAS AT DALLAS

Srinivas Parthasarathy, Carlos Busso

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

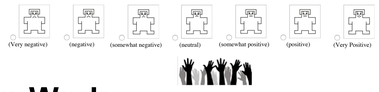


INTERSPEECH 2018
SEPTEMBER 2-6 | HYDERABAD, INDIA
HYDERABAD INTERNATIONAL CONVENTION CENTRE


## Motivation

### Background

- Ground-truth labels for emotion descriptors collected as perceptual evaluations
  - Inconsistencies between annotators
  - Poor inter-annotator agreement
- How can we capture consistent information provided by raters?



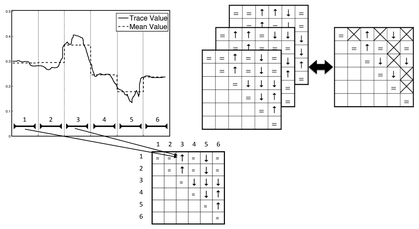

### Our Work

- Qualitative Agreement (QA) to derive ordinal labels at sentence level
  - Preference-learning or regression methods to rank-order derived labels
  - Benefits of QA-based labels over averaging scores


## QA for Sentence Level Annotations

### QA approach

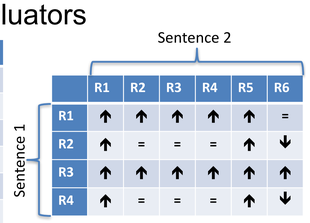
- Emotional traces




### Proposed approach

- Sentence level evaluations from individual evaluators

| | Sentence 1 | Sentence 2 |
|---|---|---|
| Rater 1 | 4.0 | 2.0 |
| Rater 2 | 3.0 | 3.0 |
| Rater 3 | 5.0 | 3.0 |
| Rater 4 | 3.0 | 3.0 |
| Rater 5 | - | 2.0 |
| Rater 6 | - | 4.0 |




- Pairwise comparisons between all pairs of annotators
  - $t_{threshold} = 1$ (attribute scores are integers)

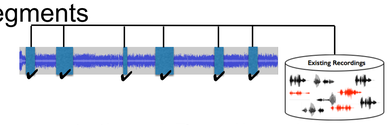
$$b_i - b_j \geq t_{threshold}$$
$$b_j - b_i \geq t_{threshold}$$
$$|b_i - b_j| < t_{threshold}$$

- We set a threshold to set clear preferences ($X$=60%, other settings showed similar performance)
  - Averaging – Sentence 1 : 3.75 , Sentence 2 : 2.83
  - QA – Sentence 1 : 62.5 % preferences
- We do not need same annotators for all the sentences
- We do not need same number of annotators per sentence

## Database and Features

### The MSP-Podcast Corpus

- Multiple sentences from speakers appearing in various podcasts (2.75s – 11s)
- Annotated on Amazon Mechanical Turk
  - Arousal, valence, dominance
- V1.1: 22,630 turns with emotional labels (38h56m)
  - Test set: 7,181 segments from 50 speakers
  - Dev set: 2,614 segments from 20 speakers
  - Train set: 12,835 segments




### Acoustic Features

- Interspeech 2013 Computational Paralinguistic Challenge feature set (6,373 features)

## Experimental Evaluation

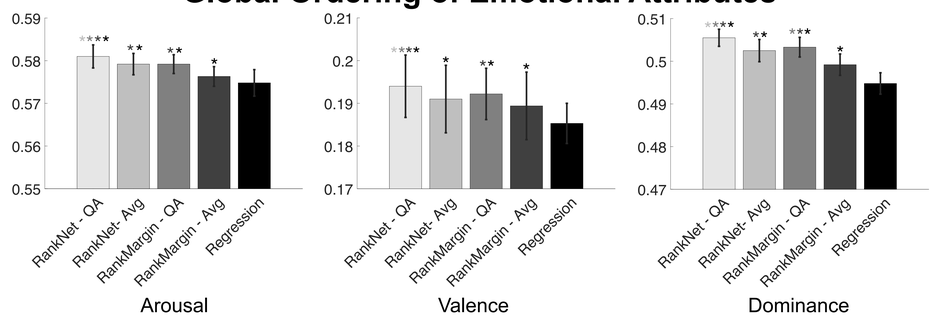
### Preference-Learning Methods

- Given: Pair of samples $x_i$, $x_j$ and corresponding labels $y_i, y_j$
- Goal: Learn function $f$ mapping $\mathbf{x}$ to scores $S$ such that if $y_i > y_j$, then $S_i > S_j$
- Given preference $y_i > y_j$
  - RankMargin: $L_{RankMargin} = \max(0, \alpha + S_j - S_i)$
  - RankNet: $L_{RankNet} = \log(1 + \exp^{-(S_i - S_j)})$
  - Regression: Concordance correlation coefficient

### Implementation

- $f$: two layer neural network, 256 nodes, ReLU activation, dropout = 0.5, ADAM optimization
- 200k random pairs for training preference learning

### Global Ordering of Emotional Attributes




- We evaluate the methods to rank-order emotional attributes
- Evaluation Metric : Kendall's $\mathcal{T}$ between two ordered lists
  - $\mathcal{T}$ = -1, complete disagreement, $\mathcal{T}$ = 1, complete agreement
- Ground-truth is the order provided by the average based labels
- Preference-learning > Regression, QA > Avg
- RankNet-QA > RankMargin-QA

## Discussions

### Local Ordering of Emotional Attributes

- Retrieval of extreme emotional utterances

| Emo | Model | Hi-10 | Hi-20 | Lo-10 | Lo-20 |
|---|---|---|---|---|---|
| Arousal | RN-QA | $0.273^{\dagger}$ | $0.300^{*\dagger}$ | $0.297^{*}$ | $0.332^{*\dagger}$ |
| | RN-Avg | $0.274^{\dagger}$ | $0.298^{\dagger}$ | 0.286 | 0.325 |
| | RMa-QA | $0.272^{\dagger}$ | $0.298^{*\dagger}$ | $0.295^{*}$ | $0.329^{*}$ |
| | RM-Avg | $0.272^{\dagger}$ | $0.296^{\dagger}$ | 0.288 | 0.326 |
| | Reg | 0.262 | 0.287 | 0.293 | 0.330 |
| Valence | RN-QA | 0.111 | $0.162^{*}$ | $0.061^{\dagger}$ | 0.044 |
| | RN-Avg | 0.109 | 0.158 | 0.058 | $0.051^{\dagger}$ |
| | RM-QA | $0.117^{*}$ | 0.159 | 0.060 | $0.045^{\dagger}$ |
| | RM-Avg | 0.104 | 0.157 | 0.058 | $0.050^{\dagger}$ |
| | Reg | 0.122 | 0.166 | 0.056 | 0.043 |
| Dominance | RN-QA | $0.159^{*}$ | $0.200^{*\dagger}$ | 0.241 | 0.287 |
| | RN-Avg | 0.149 | 0.190 | 0.238 | 0.285 |
| | RM-QA | $0.160^{*}$ | $0.198^{*\dagger}$ | $0.244^{*}$ | $0.287^{*}$ |
| | RM-Avg | 0.148 | 0.190 | 0.238 | 0.284 |
| | Reg | 0.156 | 0.194 | 0.249 | 0.291 |

### Conclusions

- QA method for constructing relative labels from sentence level annotations
- QA-based labels are more reliable than average labels for global and local emotion rankings


This work was funded by NSF CAREER award IIS-1453781