Defining Emotionally Salient Regions using Qualitative Agreement Method

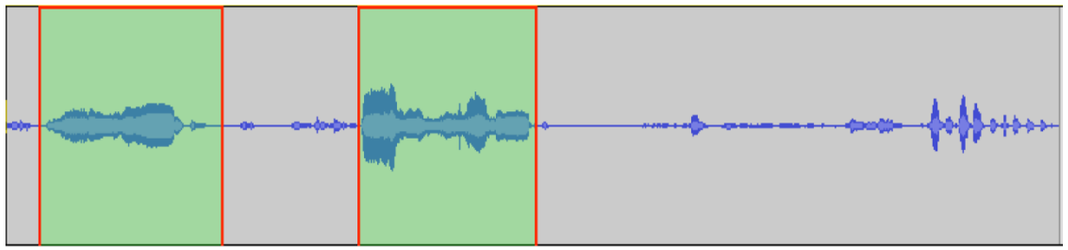
Srinivas Parthasarathy and Carlos Busso

Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas
Erik Jonsson School of Engineering and Computer Science

Sept 12, 2016

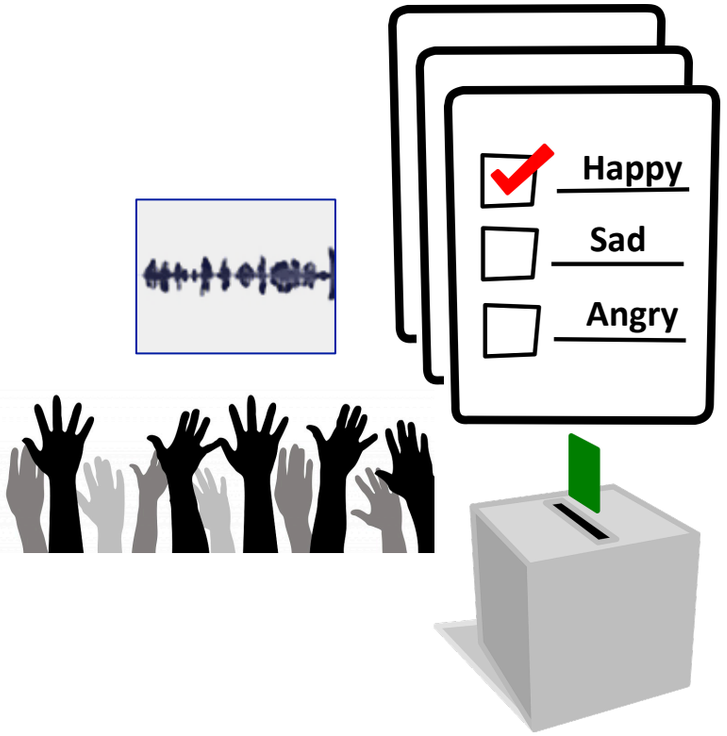
Motivation

- Expressive behavior recognition important for human computer interaction
- Human interaction is fairly neutral with few segments conveying emotion
- Need for dynamic systems that
  - are time continuous in nature
  - can detect salient regions that deviate from neutral
- Previous studies have focused on
  - continuously predicting emotional dimensional values Gunes & Schuller 2013
  - points of change of emotion Huang et al. 2015



Barriers

- Unreliable Emotional labels Cowie & Cornellius 2003, Busso et al. 2013



- Perceptual evaluation complex Cowie 2009
  - Unreliable labels affect performance of classifiers, predictors Metallinou & Narayanan 2013

- Creating labels, for salient regions, from scratch is expensive, time consuming

Goal

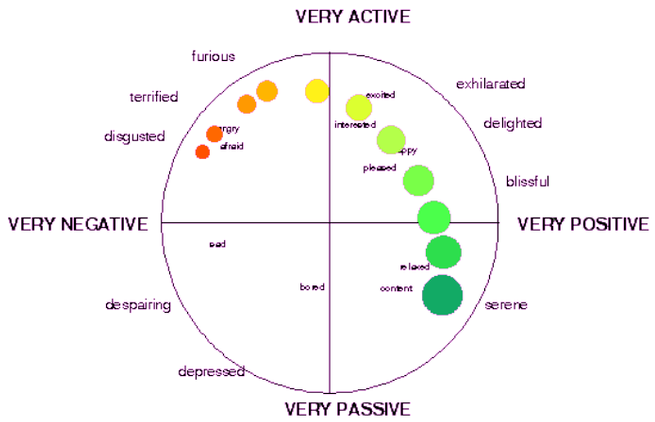
- Framework for defining reliable labels describing emotionally salient regions (hotspots)
  - Use existing perceptive evaluations (e.g. continuous time evaluations)
  - Easily extended to multiple databases
- We exploit the Qualitative Agreement (QA) method to define hotspots
  - We show that hotspots defined with QA capture individual, relative trends
  - Better than the baseline of averaging traces to form one absolute score

SEMAINE database

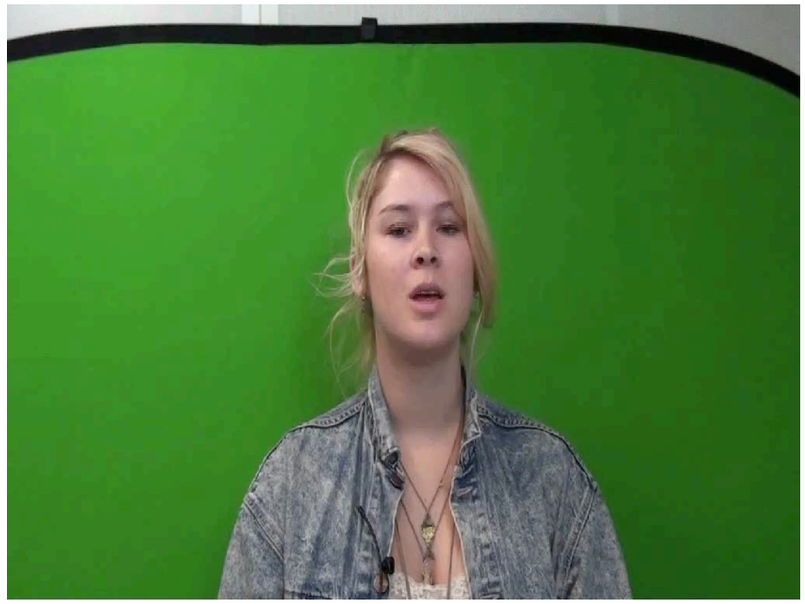
- Emotionally colored machine-human interaction McKeown et. al 2012
  - Sensitive artificial listener framework
  - Only solid SAL used (operator was played with another human)
  - 40 sessions, 10 users
- Time-continuous dimensional labels
  - Captured by FEELTRACE Cowie et al. 2000
  - We focus on arousal and valence dimensions
  - 6 evaluators for each session, evaluations range [-1,1]

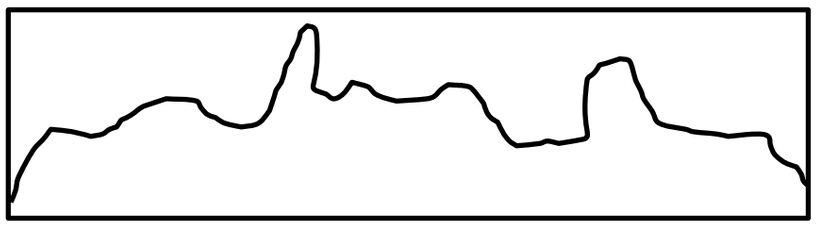
User

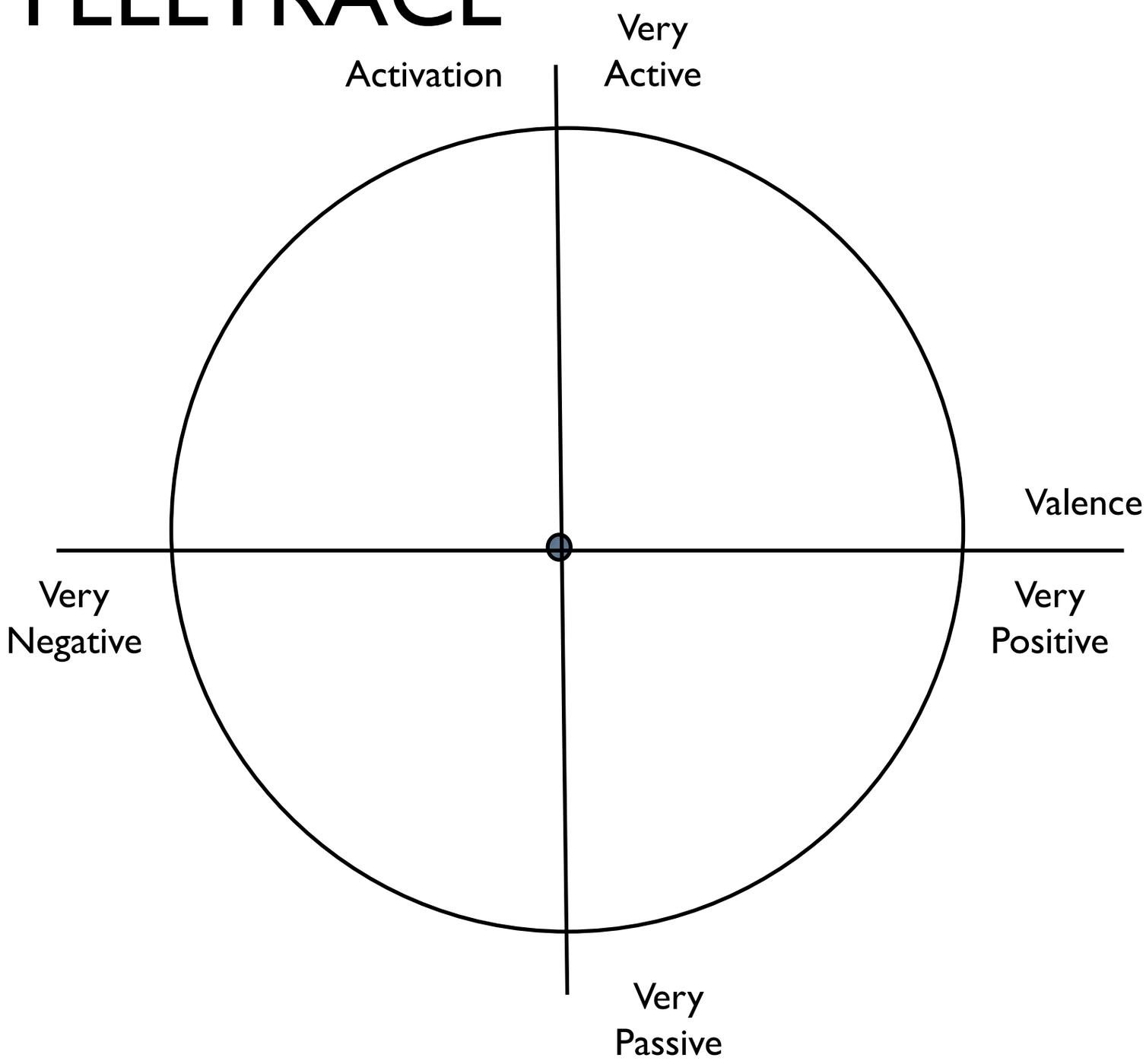
Operator



FEELTRACE






Activation
Very
Active
Valence
Very
Negative
Very
Positive
Very
Passive

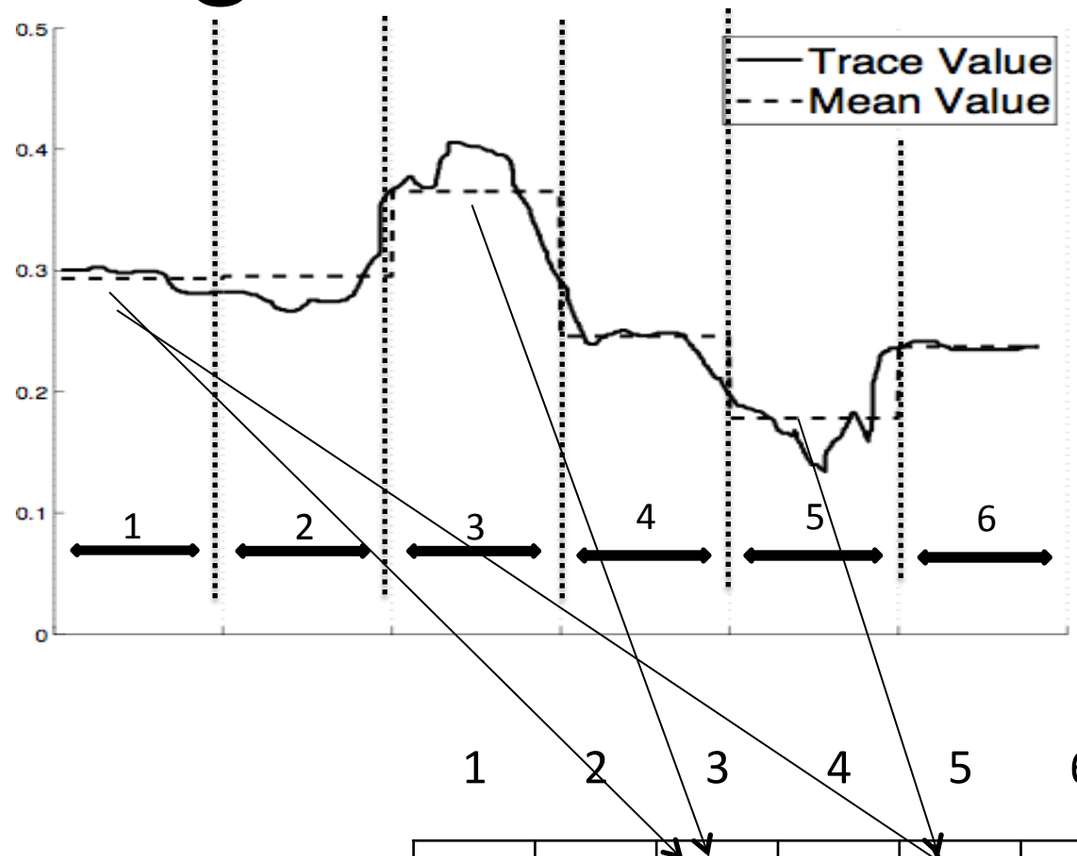
Hotspot Definition

- Hotspots defined as segments having high or low levels of emotional attribute
  - Eg. Valence hotspots – Very negative or very positive emotions

- Proposed method for definition ? **Qualitative Agreement (QA)**
- QA - Promising results for ranking emotions Parthasarathy et al. 2016

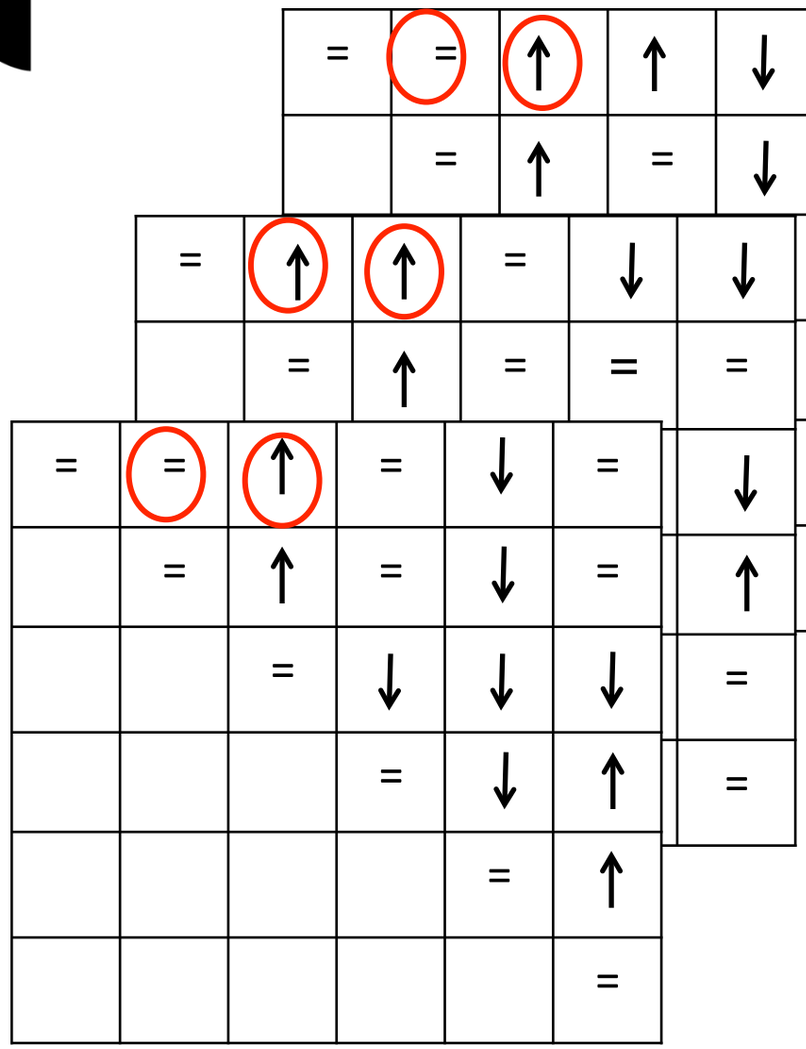
Qualitative Agreement

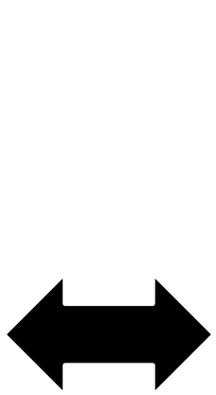
- Proposed by Cowie and McKeown 2010
- Divide trace into discretized bins
  - Mean value (b_i) of trace assigned to the bin
- Form Individual Matrix (IM)
  - Rise $b_j - b_i > t_{threshold}$
  - Fall $b_i - b_j > t_{threshold}$
  - Equal $|b_i - b_j| < t_{threshold}$

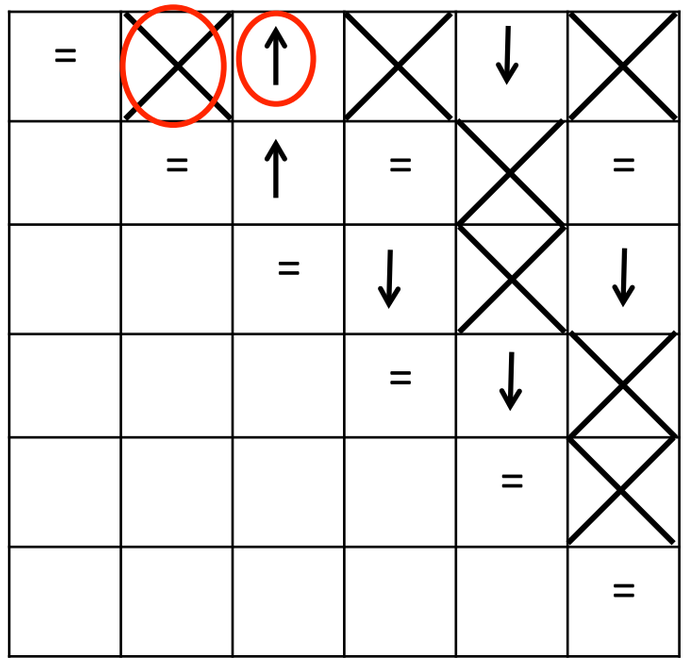


	1	2	3	4	5	6
1	=	=	↑	=	↓	=
2		=	↑	=	↓	=
3			=	↓	↓	↓
4				=	↓	↑
5					=	↑
6						=

Qualitative Agreement



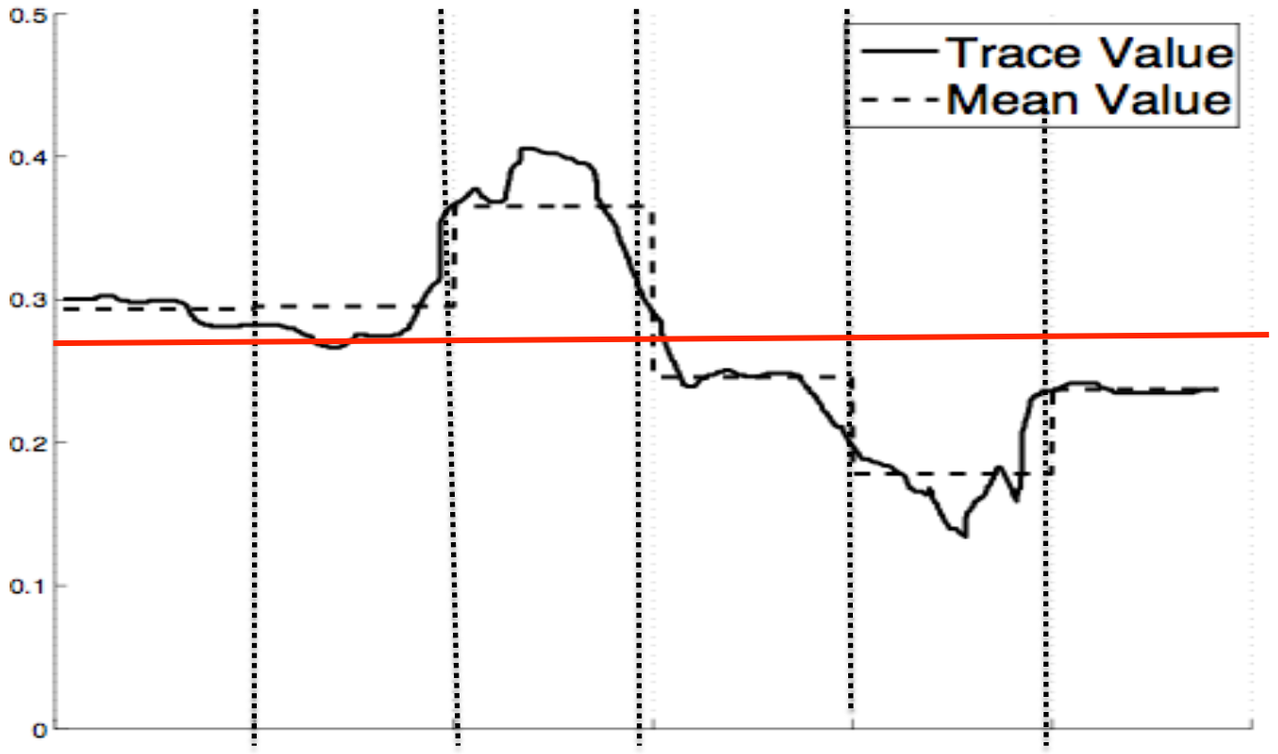


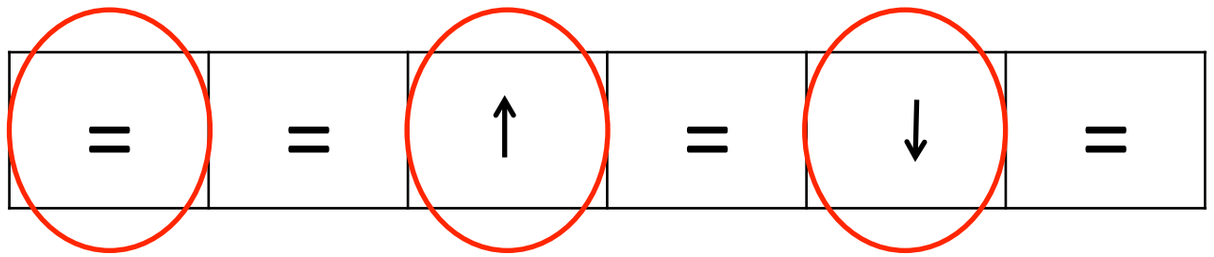


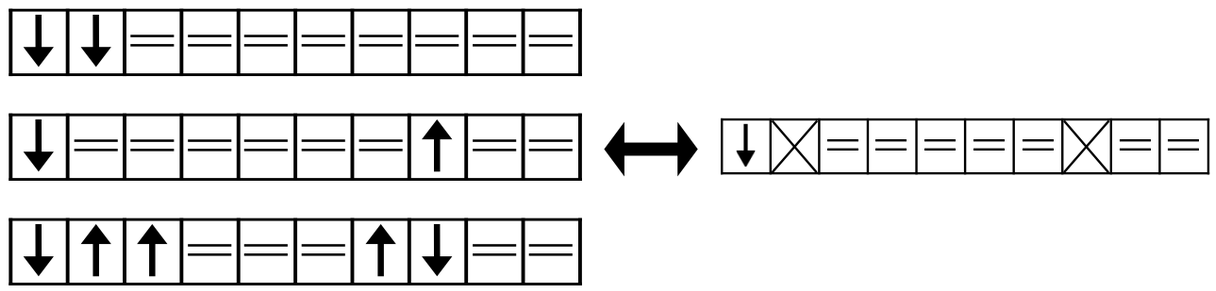
- Combine different individual matrices to form a consensus matrix (CM) to find agreement between raters
  - If X% agree on trend in IM set that to CM
  - Otherwise not considered

Qualitative Agreement – Hotspot Detection

- How to adapt QA for hotspot detection ?
- Compare with median value instead of bins
- Form individual vector (IV) for each rater
  - High $b_i - b_{median} > t_{threshold}$
  - Low $b_{median} - b_i > t_{threshold}$
  - Neutral $|b_i - b_{median}| < t_{threshold}$
- Consensus Vector (CV) – X % agreement



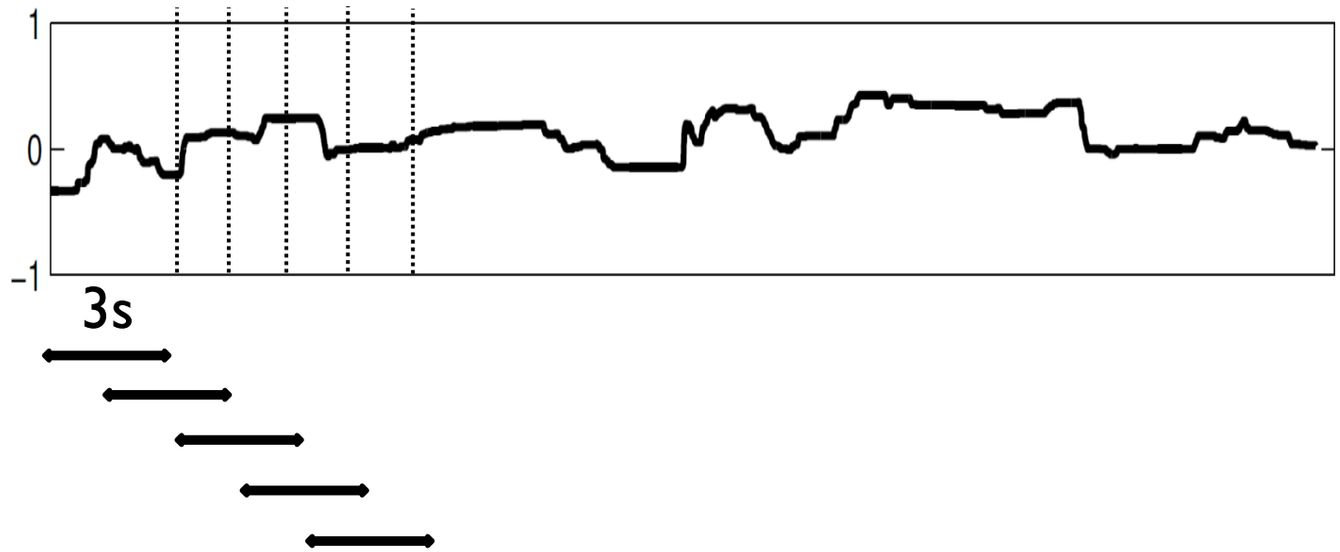




Parameters – Length of Bin

$$b_i - b_{median} > t_{threshold}$$

- Length of bin (L) set to 3s

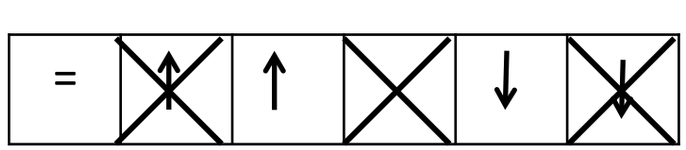


- Successive bins shifted by 250ms with 2.75s overlap
  - Gives reliable, continuous bins for hotspots, regression tasks

Parameters – Agreement Consensus

$$b_i - b_{median} > t_{threshold}$$

- Agreement – 66% (4 out of 6 raters)

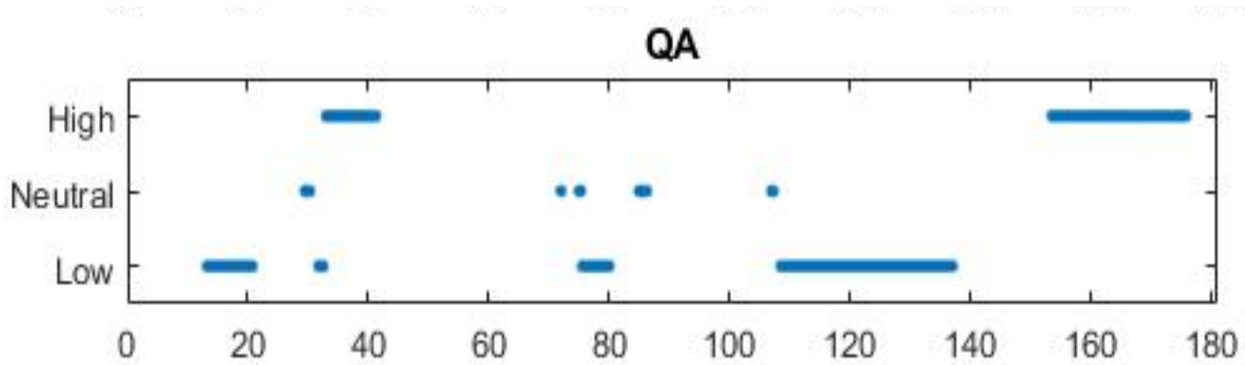


Parameters - $t_{threshold}$

$$b_i - b_{median} > t_{threshold}$$

$t_{threshold}$ = 0.0[illegible]

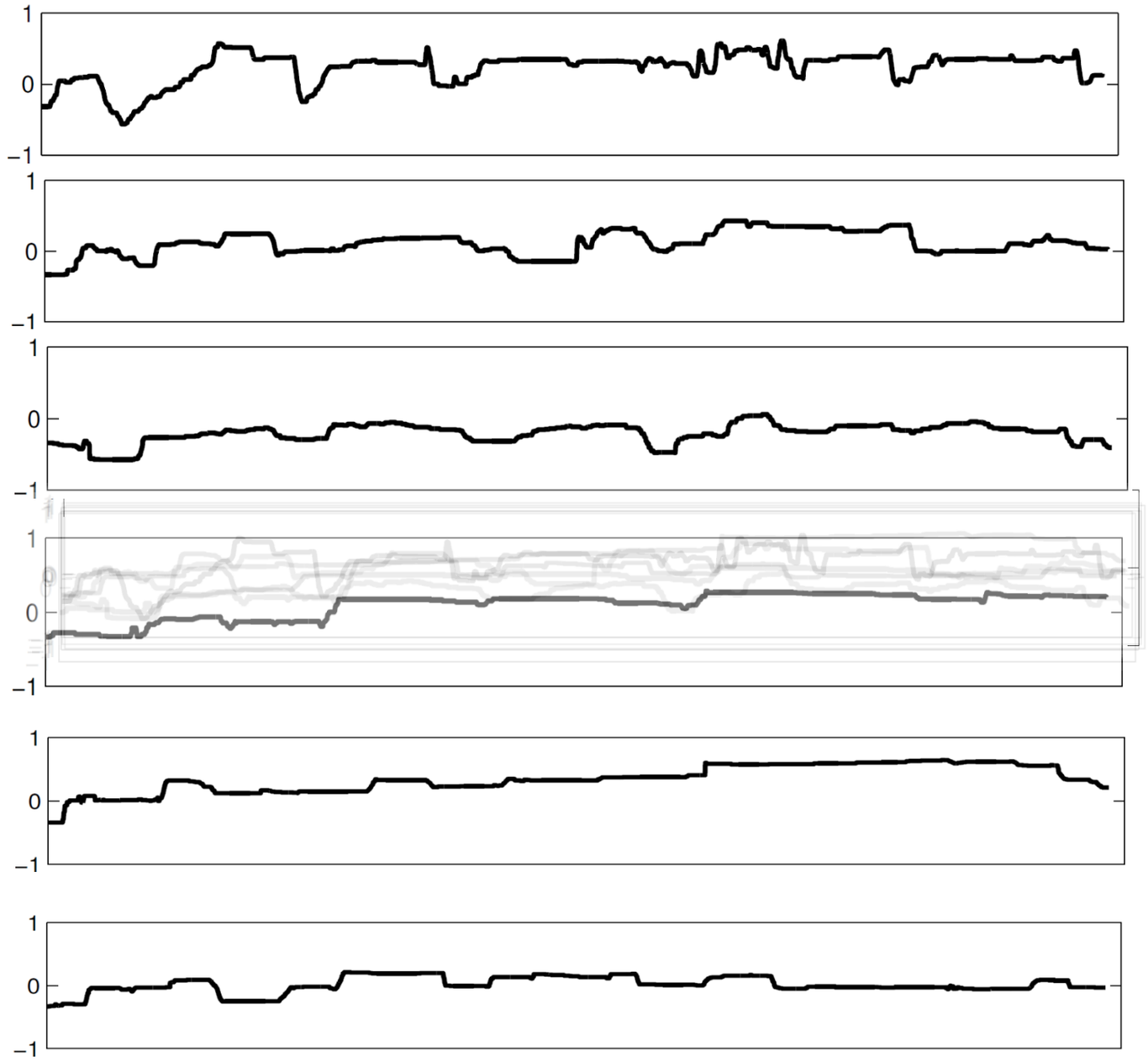
- $t_{threshold}$ = [0.025, 0.050, 0.075, 0.100, 0.125, 0.150, 0.175, 0.200]



- For low $t_{threshold}$ we have more high, low regions
- As $t_{threshold}$ is raised more neutral regions

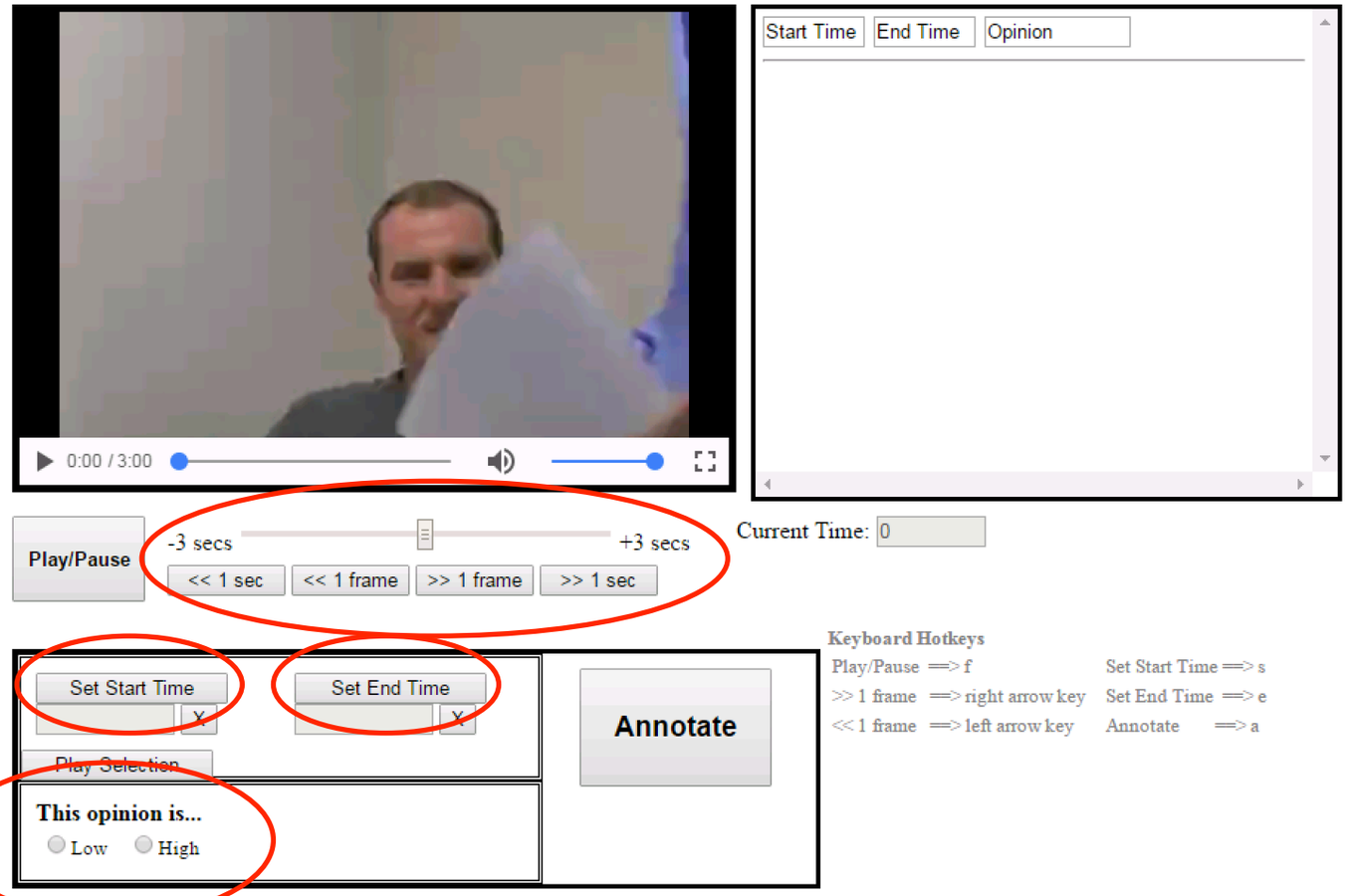
Baseline – Hotspot detection

- 6 evaluations considered individually
- Traces are averaged instead of QA
- Bin length and $t_{threshold}$ same parameters as QA.
- Unlike QA Individual trends are not considered



Hotspot ground truth

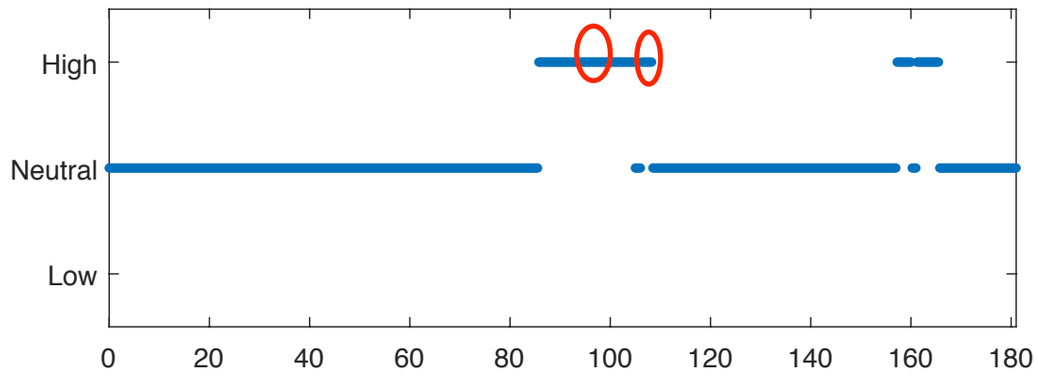
- Ground-truth established from scratch by perceptual evaluation
  - 16 sessions (8 arousal, 8 valence)
  - Evenly divided between 4 characters covering different emotions
  - Task – Select hotspot segments marking regions evaluator perceived as emotionally high or low, rest neutral, after watching entire clip

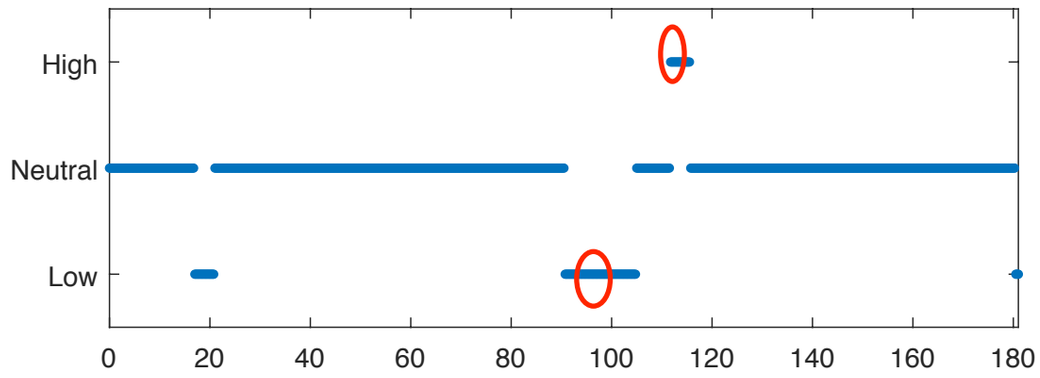


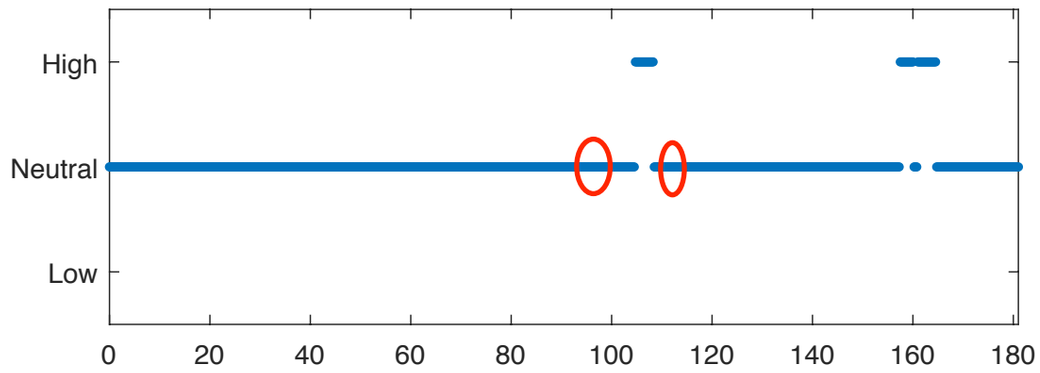
OCTAB Toolkit Park et al. 2012

Hotspot ground truth

- 3 evaluators
  - Fuse annotations by simple majority (2 out of 3)
  - Segments without agreement – no label
- Independently for arousal and valence







Hotspot ground truth

- Percentage hotspot
  - Around 5% of total traces annotated as hotspot
- Consistency – Fleiss Kappa
  - Used for measuring agreement between raters. [-1,1] corresponding to perfect disagreement and agreement
- Overall K and Region-wise K for Low, Neutral, High region
- Low values of K indicates the complexity of the task
- Time demanding

Dimension	Percentage of Ground truth hotspots			
	Low	Neutral	High	WA
Arousal	1.7%	93.4%	3.5%	1.4%
Valence	2.2%	95.6%	1.6%	0.6%

Dimension	Region-wise K			Overall K
	Low	Neutral	High	
Arousal	0.0651	0.1375	0.1938	0.1355
Valence	0.0778	0.1145	0.2256	0.1212

Results

- Proposed definition of hotspots compared to ground truth hotspot
- Process similar to Voice Activity Detection
  - Evaluation done with metrics used for VAD
- Hit Rate – Recall of neutral and emotional regions

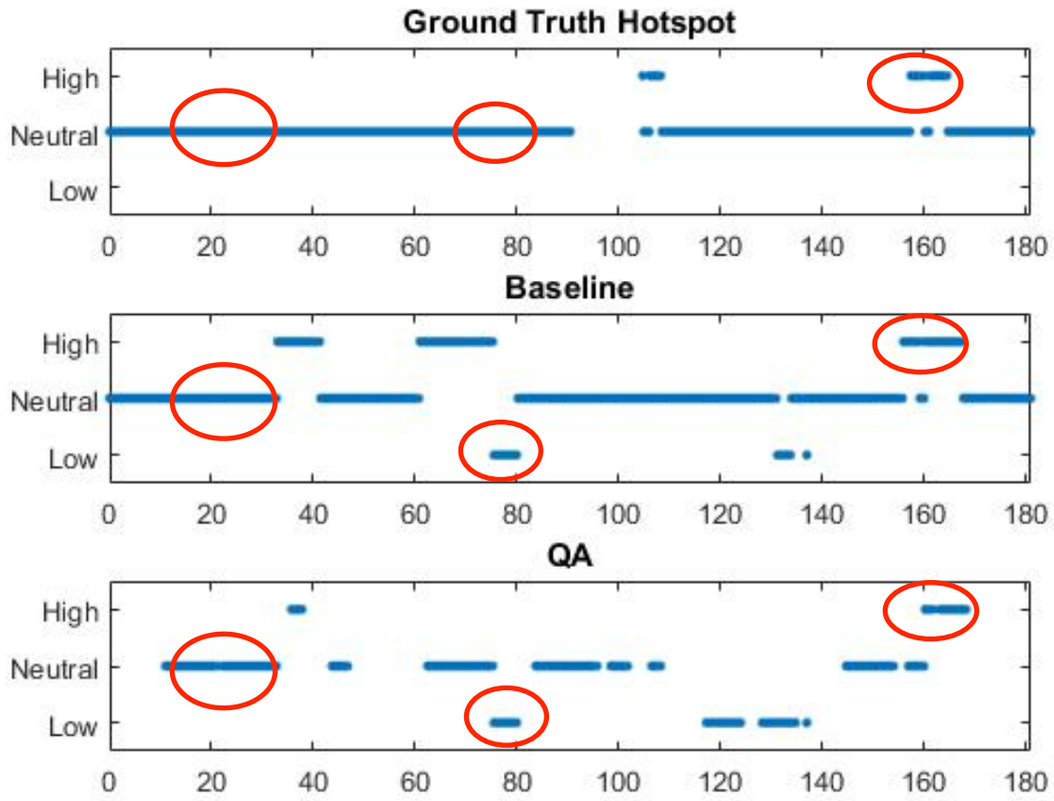
$$H_{h,l} = \frac{N_{high,low}^{pred}}{N_{high,low}^{ref}}$$

$$H_{neu} = \frac{N_{neu}^{pred}}{N_{neu}^{ref}}$$

$$H_{ov} = \frac{H_{h,l} + H_{neu}}{2}$$

Results

- Emphasis on recall on both high, low as well as neutral regions

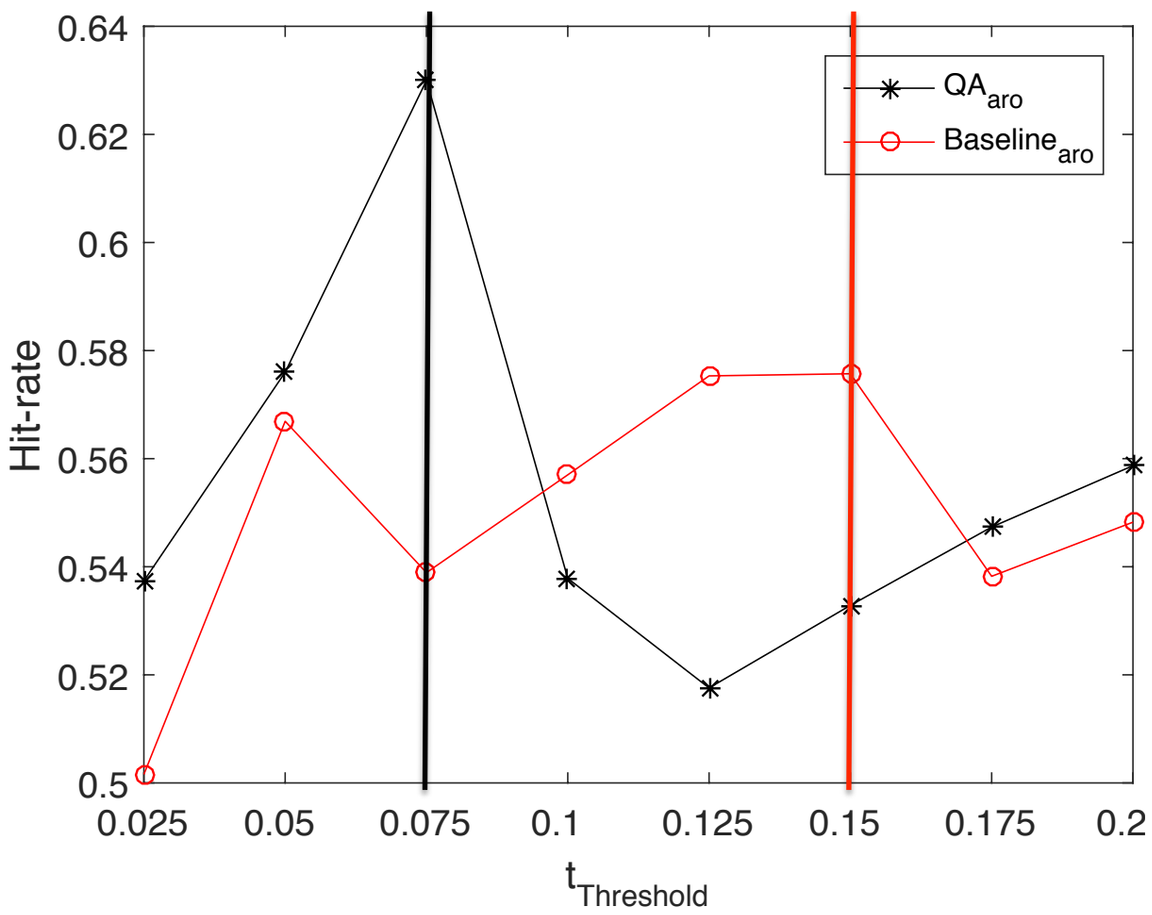


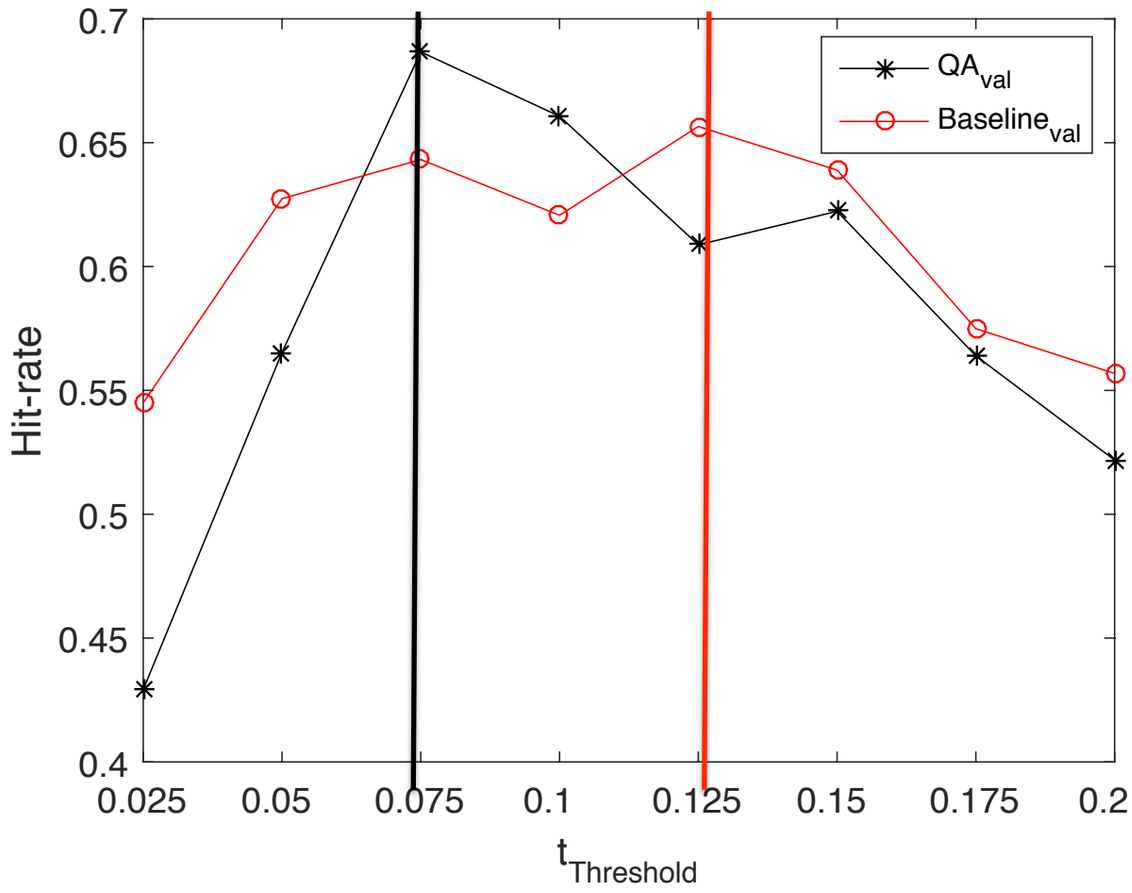
- False hotspot detection affects $H_{neu} = \frac{N_{neu}^{pred}}{N_{neu}^{ref}}$

- Good definition increases both rate of both recalls, captured by $H_{ov} = \frac{H_{h,l} + H_{neu}}{2}$

Best Definition?

- Which threshold gives best hitrates ?



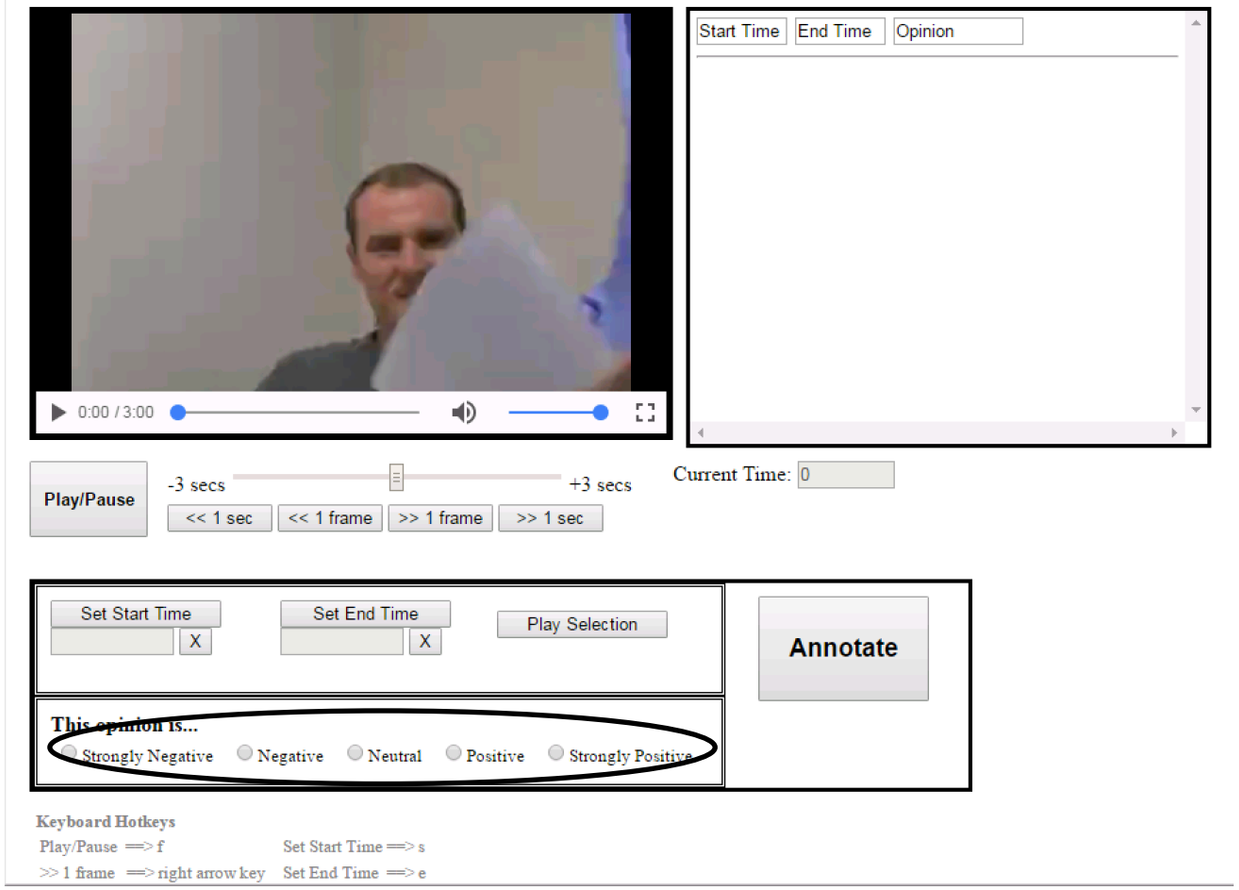


Hit-rate	
Arousal	
Baseline	0.58
QA	0.63

Hit-rate	
Valence	
Baseline	0.66
QA	0.69

Aposteriori Evaluation

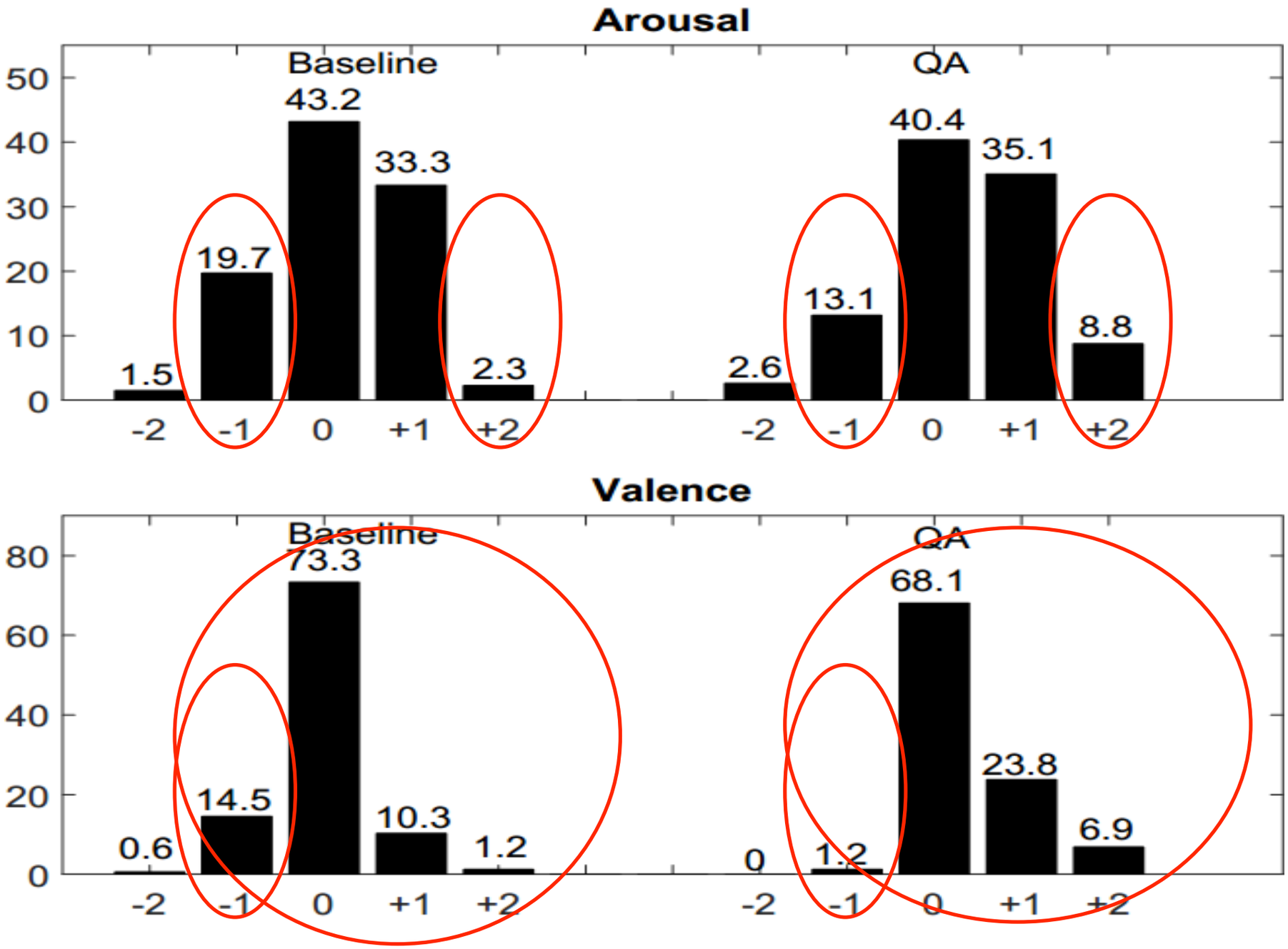
- Second set of evaluations on defined hotspots
- For each dialogue, proposed hotspots for QA, baseline evaluated posteriorly
  - Rate each hotspot once for QA and once for baseline
  - Best thresholds for baseline and QA used
  - 5 likert scale (-2 strongly disagree, 2 strongly agree)



Aposteriori Evaluation

Aposteriori Evaluation

- Reviewers find QA hotspots better



Conclusions

- Definition of emotionally salient regions over continuous time evaluations
- Two methods explored with various parameters
  - Baseline averaging
  - QA
- Hotspots defined through QA closer to ground truth and more agreeable posteriorly

Thanks for your attention!

[1] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, February 2013.

[2] Z. Huang, J. Epps, and E. Ambikairajah, "An investigation of emotion change detection from speech," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 1329–1333.

[3] R.Cowie and R.Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, April 2003.

[4] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human- computer interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford University Press, November 2013, pp. 110– 127.

[5] R. Cowie, "Perceiving emotion: towards a realistic understanding of the task," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3515–3525, December 2009.

[6] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013.

Thanks for your attention!

[7] G.McKeown, M.Valstar, R.Cowie, M.Pantic,and M.Schroder, "The SEMAINE database: Annotated multimodal r ecords of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, January-March 2012.

[8] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder, "'FEELTRACE': An instrum ent for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and E motion*. Newcastle, Northern Ireland, UK: ISCA, September 2000, pp. 19–24.

[9] S. Parthasarathy, R. Cowie, C. Busso, "Using Agreement on Direction of Change to Build Rank-Based Emotion Classifiers", To Appear, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016

[10] S. Park, G. Mohammadi, R. Artstein, and L. P. Morency, "Crowd- sourcing micro-level multimedia annotations : The challenges of evaluation and interface," in *ACM Multimedia 2012 workshop on Crowdsourcing for multimedia (Cr owdMM)*, Nara, Japan, Octo- ber 2012, pp. 29–34.

[11] R. Cowie and G. McKeown, "Statistical analysis of data from initial labelled database and recommendations fo r an economical coding scheme," Belfast, Northern Ireland, UK, September 2010, SEMAINE Report D6b. [Online] . Available: http://semaine-project.eu