A Stepwise Analysis of Aggregated Crowdsourced Labels Describing Multimodal Emotional Behaviors

Alec Burmania and Carlos Busso

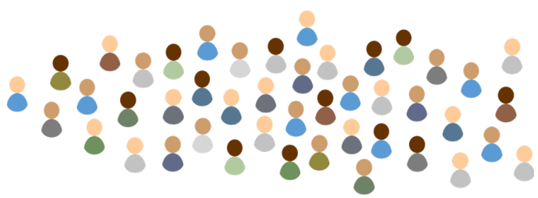
Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas
Erik Jonsson School of Engineering and Computer Science

Labels from Expressive Speech

- Emotional databases rely on labels for classification
  - Usually obtained via perceptual evaluations
- Lab Setting
  - \+ Allows researcher close control over subjects
  - \- Expensive
  - \- Small demographic distribution
  - \- Smaller corpus size
- Crowdsourcing
  - \+ Can solve some of the above issues
  - \+ Widely tested and used in perceptual evaluations
  - \- Raises issues with rater reliability

Labels from Expressive Speech

- How do we balance quality and quantity in perceptual evaluations?
  - How many labels is enough?
- Crowdsourcing makes these decisions important

Many Evaluators & Low Quality		Few Evaluators & High Quality
	or	

- What is the value of an extra evaluator?

Previous Work

- Burmania et al. (2016) explores tradeoff between quality and quantity of emotional annotations on emotion classification
  - Explore the concept of effective reliability proposed by Rosenthal [2008]

$$R_{SB} = \frac{n\kappa}{1 + (n-1)\kappa}$$

  - It is equivalent to have:
    - 15 annotators with reliability κ=0.45 (R_{SB}=92)
    - 10 annotators with reliability κ=0.54 (R_{SB}=92)

- Classification performance may be increase via design of label collection instead of maximizing inter-evaluator agreement

A. Burmania, M. Abdelwahab, and C. Busso, "Tradeoff between quality and quantity of emotional annotations to characterize expressive behaviors," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), Shanghai, China, March 2016, pp. 5190-5194.

Motivation

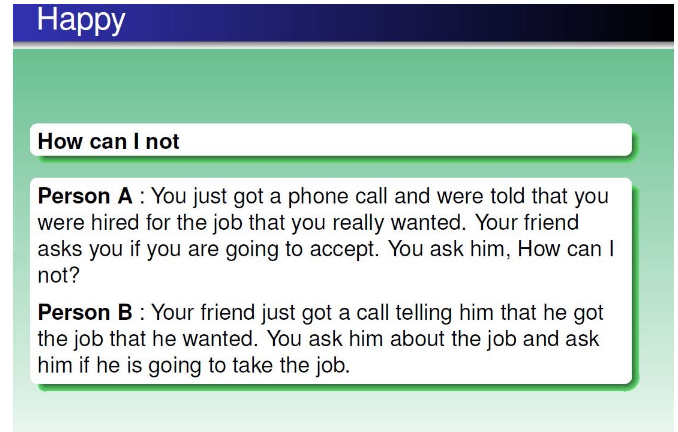
- Compare the value of additional evaluators by analyzing consensus labels

$$\underbrace{\boxed{\text{N evaluators}}}_{\text{Consensus labels}} \quad \overset{=}{\neq} \quad \underbrace{\boxed{\text{N Evaluators}} + \boxed{\text{1 new evaluator}}}_{\text{Consensus labels}}$$

- Derive guideline for subjective evaluations
  - Case study: emotional annotations of the MSP-IMPROV corpus

MSP-IMPROV Corpus

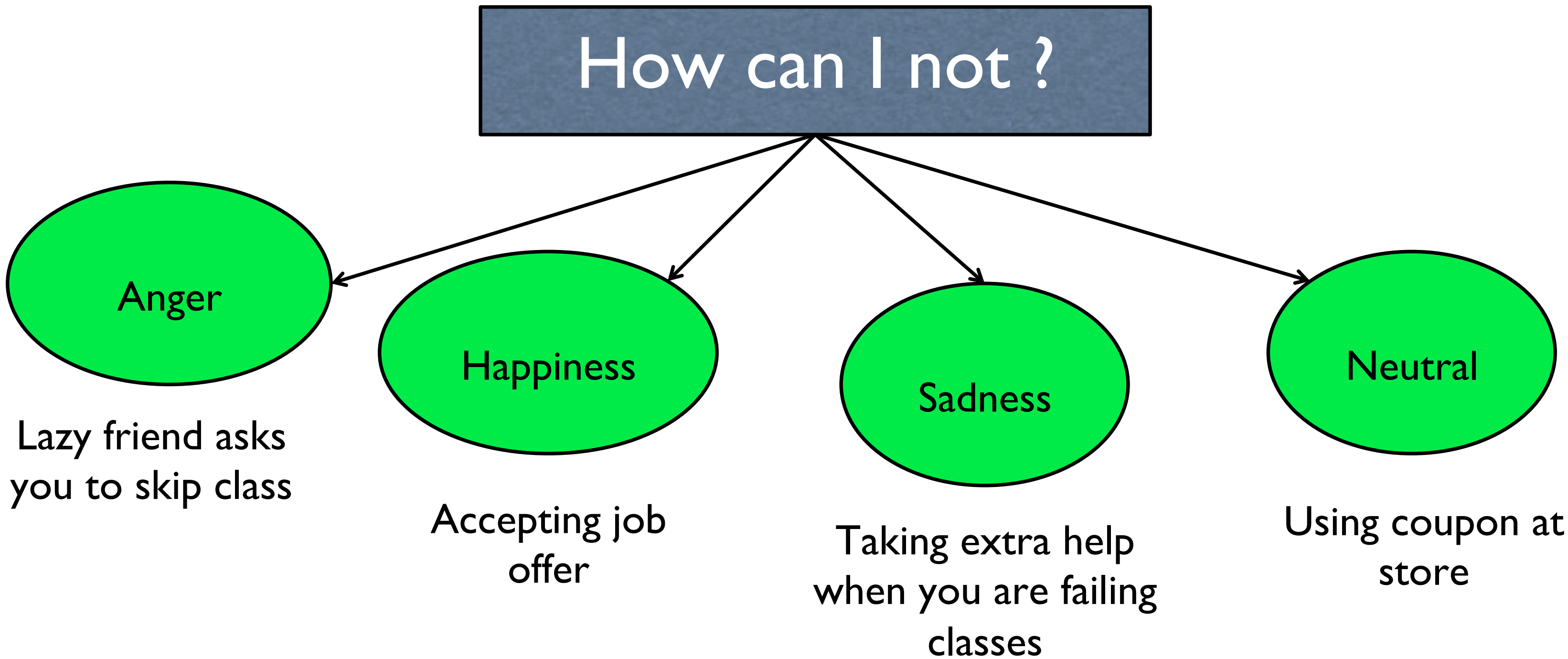
- Recordings of 12 subjects improvising scenes in pairs (>9 hours, 8,438 turns) [Busso et al, 2017]
- Actors are assigned context for a scene that they are supposed to act out
- Collected for corpus of fixed lexical content but different emotions
- Data Sets
  - Target – Recorded Sentences with fixed lexical content (648)
  - Improvisation – Scene to produce target
  - Interaction – Interactions between scenes



An example scene.

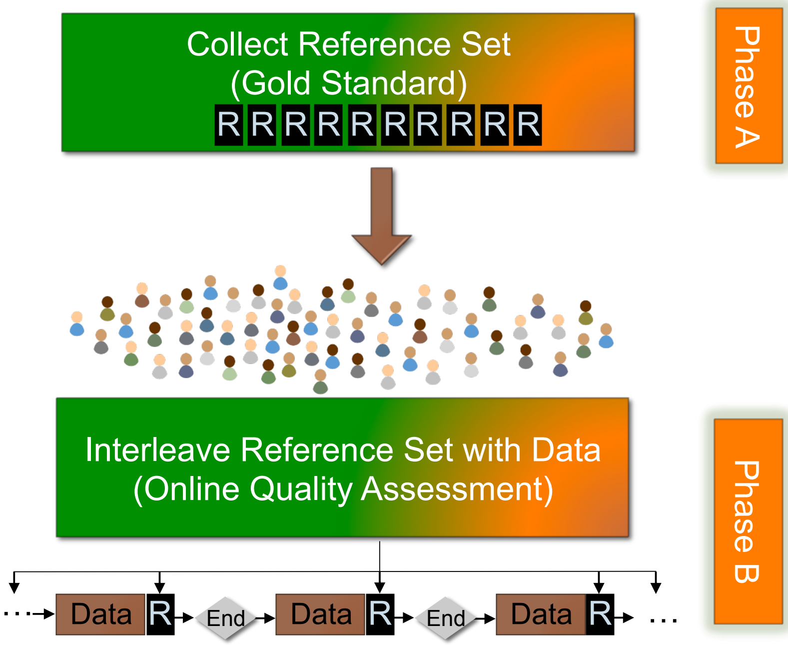
C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," IEEE Transactions on Affective Computing, vol. 8, no. 1, pp. 119-130 January-March 2017.

MSP-IMPROV Corpus

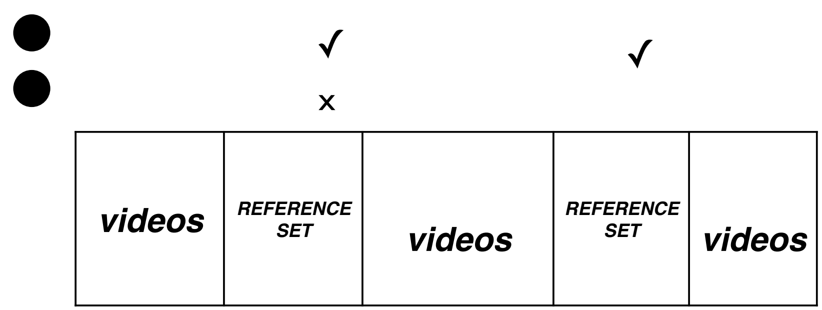


Perceptual Evaluation

- Verify if a worker is spamming in real time
- We will focus on a five class problem (angry, sad, neutral, happy, other)
- Reference set includes target sentences (648)



Trace performance in real time



A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," IEEE Transactions on Affective Computing, vol. 7, no. 4, pp. 374-388, October-December 2016.

Rater Quality

Constant sample size

	5 Raters		10 Raters		15 Raters		20 Raters		25 Raters	
Δθ	# sent	κ	# sent	κ	# sent	κ	# sent	κ	# sent	κ
5	638	**0.572**	525	0.558	246	0.515	52	0.488	0	-
10	643	0.532	615	0.522	466	0.501	207	0.459	26	0.455
15	648	0.501	643	0.495	570	0.483	351	0.443	112	0.402
20	648	0.469	648	0.471	619	0.463	510	0.451	182	0.414
25	648	0.452	648	0.450	643	**0.450**	561	0.440	247	0.416
30	648	0.438	648	0.433	648	0.436	609	0.431	298	0.410
35	648	0.425	648	0.433	648	0.426	619	0.424	346	0.403
40	648	0.420	648	0.427	648	0.425	629	0.423	356	0.402
90	648	**0.422**	648	0.419	648	0.422	629	**0.419**	381	0.409

Increasing agreement due to filter

Decreasing samples meeting size criteria

Label Groups

- We consider two sets of labels based on kappa agreement:
  - High agreement group (n=12)
  - Moderate agreement group (n=20)

	5 Raters		10 Raters		15 Raters		20 Raters		25 Raters	
Δθ	# sent	κ	# sent	κ	# sent	κ	# sent	κ	# sent	κ
5	638	**0.572**	525	0.558	246	0.515	52	0.488	0	-
10	643	0.532	615	0.522	466	0.501	207	0.459	26	0.455
15	648	0.501	643	0.495	570	0.483	351	0.443	112	0.402
20	648	0.469	648	0.471	619	0.463	510	0.451	182	0.414
25	648	0.452	$\boxed{648}$	$\boxed{0.450}$	643	**0.450**	561	0.440	247	0.416
30	648	0.438	648	0.433	648	0.436	609	0.431	298	0.410
35	648	0.425	648	0.433	648	0.426	619	0.424	346	0.403
40	648	0.420	648	0.427	648	0.425	629	0.423	356	0.402
90	648	**0.422**	648	0.419	648	0.422	$\boxed{629}$	$\boxed{\mathbf{0.419}}$	381	0.409

High Agreement Condition

Moderate Agreement Condition

Label Aggregation

Happiness
Happiness
Sadness
Happiness
Sadness
→ Happiness

- Aggregation of votes is done using majority vote
- Each vote is equally weighted
- Votes are iteratively added chronologically as they were collected
- Due to majority vote, we establish the following transitions:
  - EmoA→EmoA (No Change)
  - EmoA→NA (No Agreement – a tie has been established)
  - NA→EmoA (A tie is broken)
  - NA→NA (tie remains a tie)

We cannot transition from one emotion to another!

Before

Happiness
Happiness
Sadness
Happiness
Sadness
Sadness
→ No Agreement

After

Happiness
Happiness
Sadness
Happiness
Sadness
Sadness
+ Anger
→ No Agreement

Experiments

- Trends in labels will be evaluated iteratively for each added label
- We consider:

Label Stability

Label Changes

Frequency of Change

Adding more than one evaluator

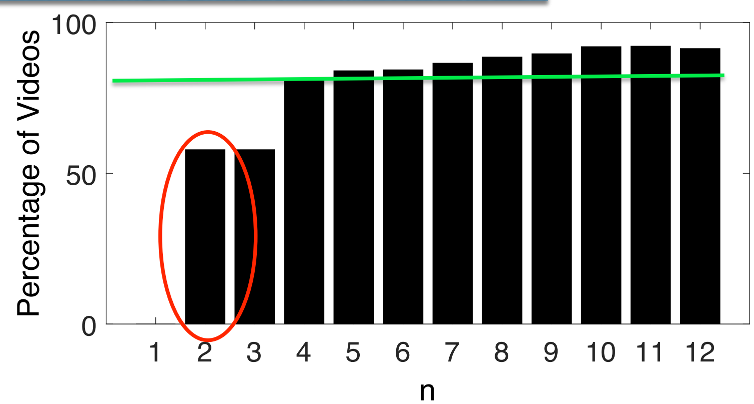
Five class problem (angry, sad, neutral, happy, other)!

Label Stability

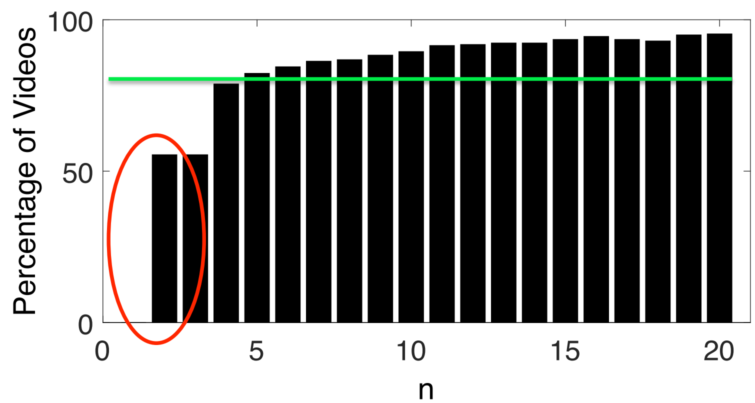
Percentage of videos with the same aggregated labels before and after adding an additional evaluator

- EmoA ➔ EmoA
- NA ➔ NA

- Observations
  - After 4 evaluators, labels are stable
  - n=6, less than 10% of labels change
  - Similar trends for high and moderate agreement conditions



Moderate Agreement



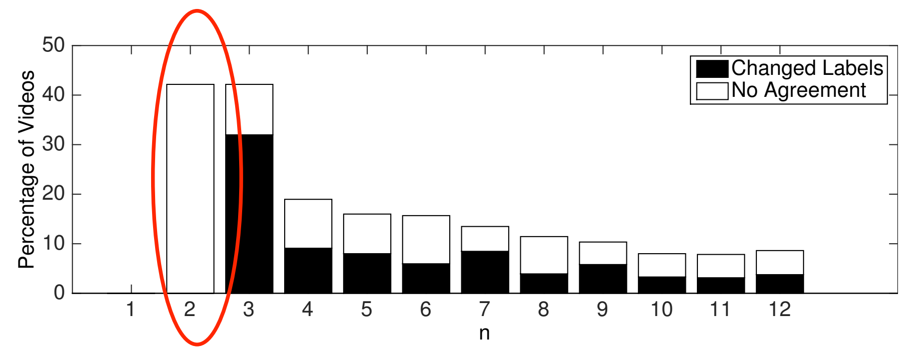
High Agreement

Label Changes

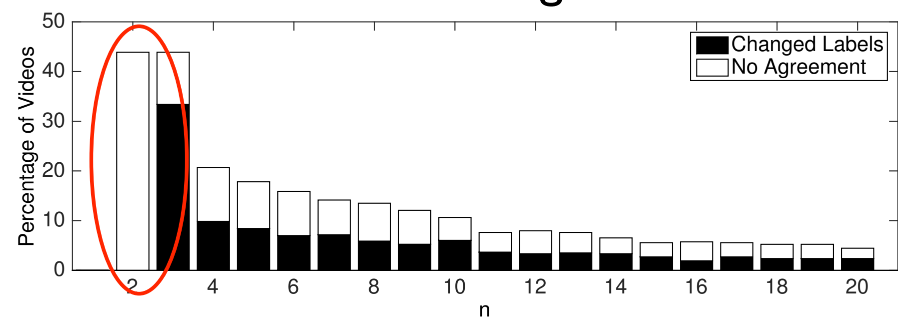
Percentage of the videos in which their labels changed as we add one extra evaluator

- Inverse plots
- NA → EmoA
- EmoA → NA

- Observations
  - n=2, 40-44% agreement is lost
  - n=3, most of the ties are solved



Moderate Agreement

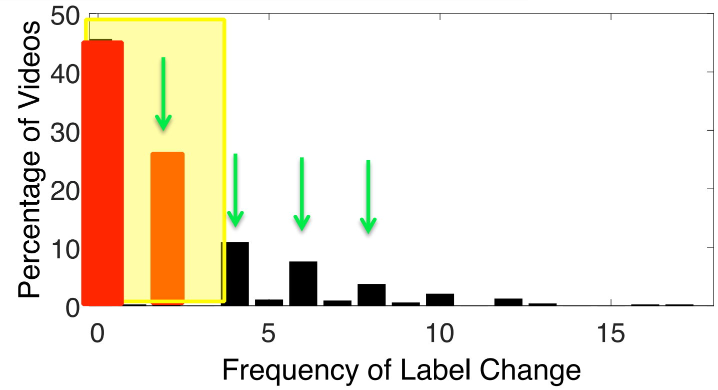


High Agreement

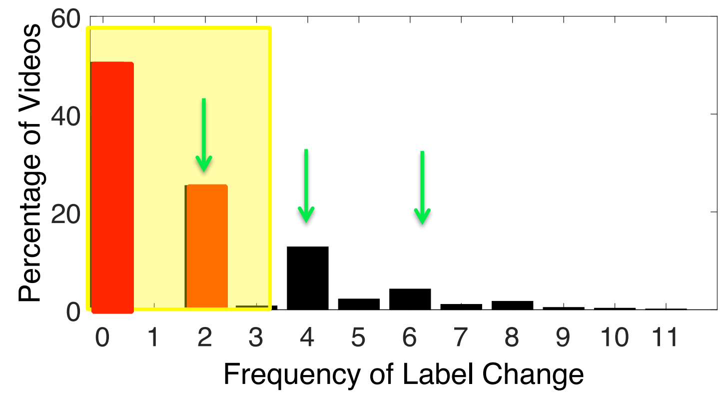
Change Frequency

Percentage of the videos in which their aggregated labels changed *m* times as we incrementally add evaluators

- Example, ~25% change labels 2 times
- Observations
  - 45% to 50% never change labels
  - Trend on even values of *m* indicate that ties are usually broken
  - About 75% sentences change labels less than 4 times
  - About 10% of the sentences change labels multiple times



Moderate Agreement



High Agreement

Adding More than One Evaluator

- How different are the aggregated labels when we add more than one evaluator?
  - 3 versus 5, 5 versus 20
- This analysis does not follow the incremental stepwise approach
  - Snapshots different values of *n*
- We consider:
  - 3, 5, 9, and 20 annotators
- We have an additional case:
  - EmoA ➔ EmoB (from one emotion to another)

Before

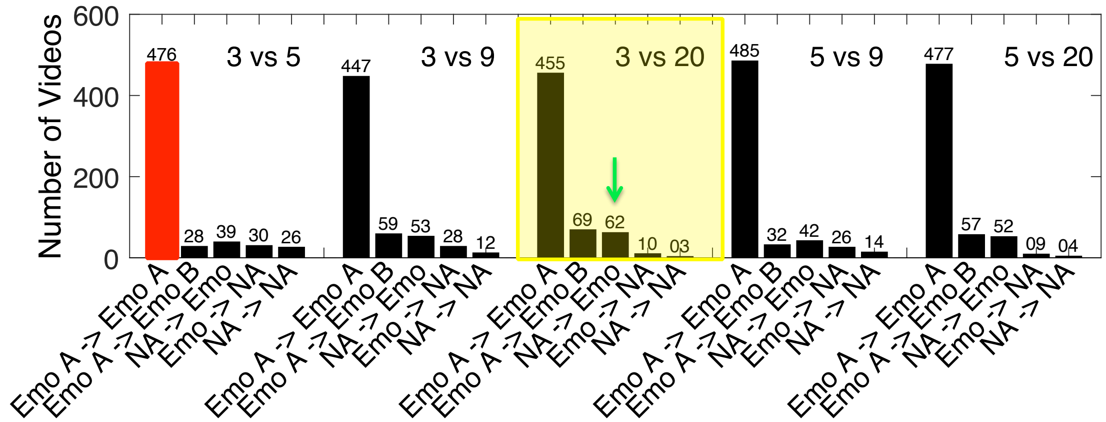
Happiness
Sadness
Happiness
→ Happiness

After

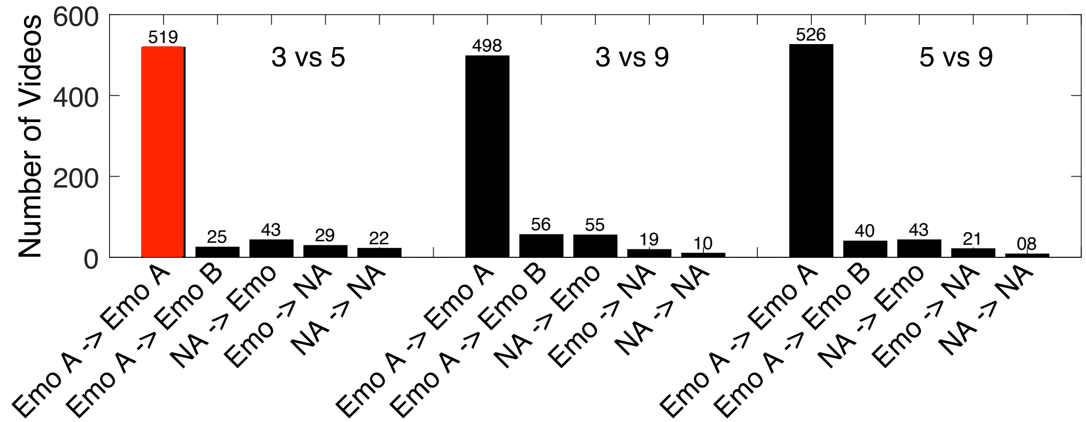
Happiness
Sadness
Happiness
Sadness
Sadness
→ Sadness

Adding More than One Evaluator

Moderate Agreement



High Agreement



- Observations:
  - Labels are very stable, even 3 versus 20 (76% overlap in labels)
  - Only few labels benefits from extra evaluations
  - Higher agreement case shows more stability

Discussion

- There is a reduced value in additional annotations
  - It helps about 10% of the labels
- We can save resources by tracking consistency of evaluations
  - Five evaluators per sentence resolve most of the ambiguities
  - We observe this trend for moderate and high inter-evaluator agreement

- Zhang et al. [2015] proposed to stop evaluation when agreement is reached
  - If n=5 and three people agree, stop the evaluation

Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions," in International conference on Multimodal interaction (ICMI 2015), Seattle, WA, USA, November 2015, pp. 275–278.

Discussion

- An important exception is when consensus labels are not the goal
  - Training with soft-margin [Lotfian and Busso, 2017]
  - Study of emotion perception

- Emotion perceptual evaluations are complex cognitive tasks
  - We expect higher label stability for simpler behavioral tasks

R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in International Conference on Affective Computing and Intelligent Interaction (ACII 2017), San Antonio, TX, USA, October 2017.

Limitation and Future Work

- Generalizing the patterns in other databases
  - Larger or small numbers of classes
  - Different corpora
  - Inter-evaluator agreement variability
- Use of other aggregation techniques
  - Entropy based techniques

Questions?

Interested in the MSP-IMPROV database?
Come visit us at msp.utdallas.edu and click "Resources"

This work was funded by NSF CAREER award IIS-1453781

References

Burmania, A., S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using on line quality assessment," IEEE Transactions on Affective Computing, vol. 7, no. 4, pp. 374–388, October-December 2016.

Burmania, A., M. Abdelwahab, and C. Busso, "Tradeoff between quality and quantity of emotional annotation s to characterize expressive behaviors," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), Shanghai, China, March 2016, pp. 5190–5194.

Busso, C., S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPRO V: An acted corpus of dyadic interactions to study emotion perception," IEEE Transactions on Affective Computing, vol. 8, no. 1, pp. 67–80, January-March 2017.

Lotfian, R. and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in International Conference on Affective Computing and Intelligent Interaction (ACII 2017), San Antonio, TX, USA, October 2017.

Rosenthal, R. "Conducting judgment studies: Some methodological issues," in The new handbook of methods in nonverbal behavior research, J. Harrigan, R. Rosenthal, and K. R. Scherer, Eds., pp. 199–234. Oxford University Press, Oxford, UK, May 2008.

Zhang, Y., E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions," in International conference on Multimodal interaction (ICMI 2015), Seattle, WA, USA, November 2015, pp. 275–278.