The Sound of Respondents: How Do Emotional States Affect the Quality of Voice Answers in Smartphone Surveys?

Christoph Kern[1, 2] Jan Karem Höhne[1, 3] Stephan Schlosser[4]

[1]University of Mannheim

[2]LMU Munich

[3]RECSM - Universitat Pompeu Fabra

[4]University of Göttingen

BigSurv20 Conference

Introduction

- Continuous increase of smartphone use in web surveys
  - Smartphone rate in German Internet Panel: 4% (Sep 12), 10% (May 16), 35% (Sep 20)
- Smartphones consist of numerous built-in sensors
  - Accelerometer, camera, GPS, microphone
- Data collected from these sensors provide new avenues in web survey research
- For instance, microphones of smartphones allow the collection of voice answers
  - Substitute for text answers to closed and open questions
  - Triggering open narrations and in-depth information

Background

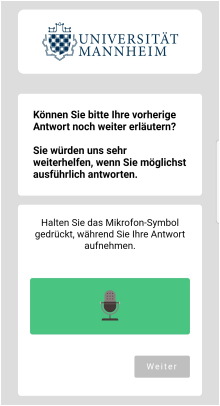
- Voice answers include tonal cues
  - Amplitudes and pitches (Frank et al. 2015, Schober et al. 2015)
- NLP research uses these cues to infer affective states
  - Models learn relationship between affective states and audio features
  - Allows to predict emotions with high accuracy (Eyben et al., 2009)
- Various directions for survey research
  - Interviewer speech in survey invitations (Conrad et al., 2013)
  - Emotions of respondents and response behavior (Heide and Gronhaug, 1991)

$\rightarrow$ Exploring the usage of pre-trained NLP models for predicting respondents' emotional states with voice data

Data

- Study conducted in the Forsa Omninet Panel (Germany)
  - Field period: December 2019 and January 2020
- One open question on attitudes towards asylum seekers
  - Administered with a request for a voice answer
  - "SurveyVoice (SVoice)" tool (Höhne et al. forthcoming)
  - Optimized survey layout
- Discrete Emotions Questionnaire (Harmon-Jones et al., 2016)
  - Self-reported emotions for validation purposes

Figure: SVoice



Data

- Sample size: 1,679
- 625 voice answers to asylum question
  - Mean length: 25.9 seconds (min: 1, max: 177)
  - Mean number of words: 59.47 words (min: 1, max: 793)

Table: Sample composition

	No voice data	Voice data
Female	0.502	0.504
Age (median cat.)	51-55	41-45
Lower education	0.229	0.178
Medium education	0.318	0.348
Higher education	0.435	0.438
In school	0.018	0.036

openEAR

Emotion and Affect Recognition Toolkit (Eyben et al., 2009)

- Includes pre-trained Support-Vector Machines
- Berlin Speech Emotion Database (Emo-DB)
  - Anger, Boredom, Disgust, Fear, Happiness, Neutral, Sadness
- Airplane Behaviour Corpus (ABC)
  - Aggressive, Cheerful, Intoxicated, Nervous, Neutral, Tired

Table: Training data (Schuller et al., 2012)

Corpus	Emotion, number of instances							Language, text, emotion	Time	#m / #f speakers
Emo-DB	angr	bore	disg	fear	happ	neut	sadn	German,	0:22	5 / 5
	127	79	38	55	64	78	53	fixed, acted		
ABC	agre	chee	into	nerv	neut	tire		German	1:15	4 / 4
	95	105	33	93	79	25		fixed, induced		

openEAR

→ Loop over all voice files

1. Obtain predicted probabilities for emotional states every ∼2 seconds (= one frame)
2. Average predicted probabilities over time (i.e., all frames)
3. Assign emotional state with the highest mean probability

Figure: openEAR in action (example)

```
LibSVM  'arousal' result (@ time: 98.389271) :  ~~> -0.12 <~~

LibSVM  'valence' result (@ time: 98.389271) :  ~~> -0.45 <~~

LibSVM  'emodbEmotion' result (@ time: 98.389271) :  ~~> neutral <~~
    prob. class 'anger':        0.030644
    prob. class 'boredom':      0.222644
    prob. class 'disgust':      0.017379
    prob. class 'fear':         0.092013
    prob. class 'happiness':    0.049911
    prob. class 'neutral':      0.568772
    prob. class 'sadness':      0.018638

LibSVM  'abcAffect' result (@ time: 98.389271) :  ~~> tired <~~
    prob. class 'agressiv':     0.010015
    prob. class 'cheerful':     0.004780
    prob. class 'intoxicated':          0.115504
    prob. class 'nervous':      0.001586
    prob. class 'neutral':      0.017269
    prob. class 'tired':        0.850846

LibSVM  'avicInterest' result (@ time: 98.389271) :  ~~> loi3 <~~
    prob. class 'loi1':         0.258098
    prob. class 'loi2':         0.311916
    prob. class 'loi3':         0.429985
```

Results - Descriptive Statistics

Table: Distribution of predicted emotion classes

(a) Emo-DB

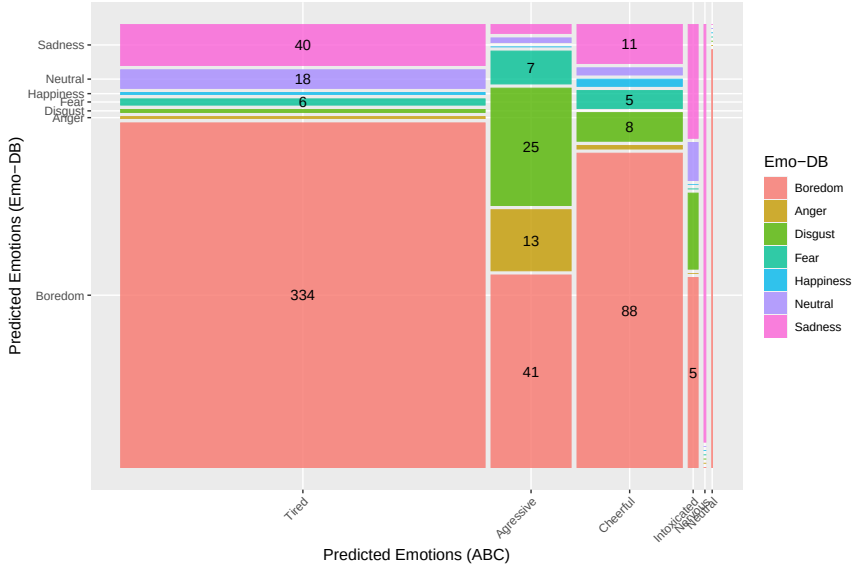
	all		$\geq$ 2 frames	
Class	n	Prop.	n	Prop.
Boredom	469	0.750	453	0.768
Sadness	58	0.093	54	0.092
Disgust	38	0.061	31	0.053
Neutral	22	0.035	18	0.031
Fear	18	0.029	16	0.027
Anger	16	0.026	15	0.025
Happiness	4	0.006	3	0.005

(b) ABC

	all		$\geq$ 2 frames	
Class	n	Prop.	n	Prop.
Tired	405	0.648	393	0.666
Cheerful	117	0.187	106	0.180
Aggressive	89	0.142	78	0.132
Intoxicated	11	0.018	10	0.017
Nervous	2	0.003	2	0.003
Neutral	1	0.002	1	0.002

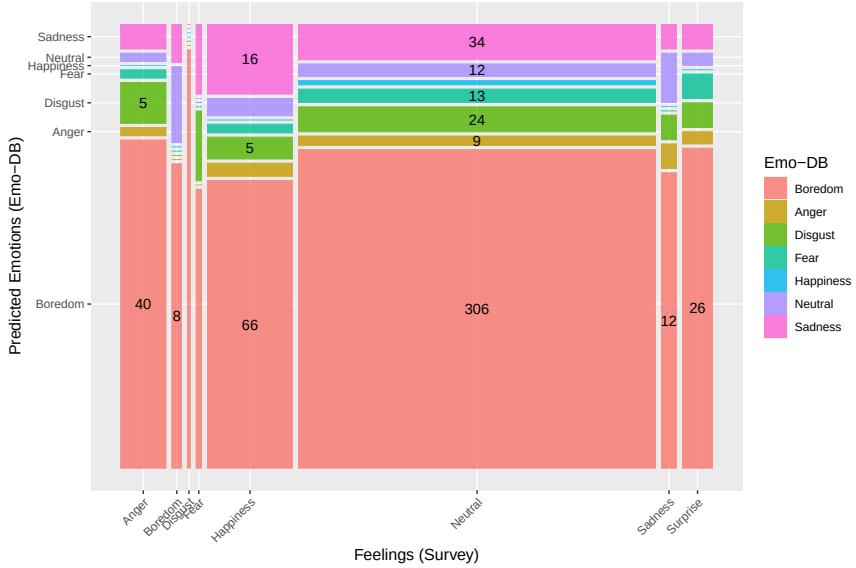
Results - Consistency Check

Figure: Comparison of predicted emotion classes (Emo-DB vs. ABC)



Results - Prediction vs. Survey Response

Figure: Comparison of predicted (Emo-DB) and reported emotions



Results - Prediction vs. Survey Response

Table: Regression models with y = number of words of response, $\mathbf{x}$ = Emo-DB predictions

	(1)	(2)
Anger	22.338 (17.004)	31.231 (17.543)
p	0.190	0.076
Disgust	−7.176 (11.281)	−3.343 (11.437)
p	0.525	0.771
Fear	−14.246 (16.064)	−11.079 (16.203)
p	0.376	0.495
Happiness	−20.662 (33.583)	−4.912 (33.556)
p	0.539	0.884
Neutral	−30.276 (14.590)	−32.968 (14.504)
p	0.039	0.024
Sadness	−11.205 (9.310)	−11.693 (9.259)
p	0.230	0.208
Constant	63.412 (3.092)	75.543 (19.747)
Demographic controls		Yes
Observations	624	622
r^2	0.014	0.073

Reference: Boredom

Results - Prediction vs. Survey Response

Table: Regression models with y = number of words of response, $\mathbf{x}$ = ABC predictions

	(1)	(2)
Aggressive	15.361 (7.742)	16.491 (7.887)
p	0.048	0.037
Cheerful	−22.978 (6.941)	−18.963 (7.035)
p	0.001	0.008
Intoxicated	−34.114 (20.204)	−27.444 (20.154)
p	0.092	0.174
Nervous	−56.342 (46.866)	−41.899 (46.738)
p	0.230	0.371
Neutral	−28.842 (66.196)	−43.679 (65.933)
p	0.664	0.508
Constant	63.842	75.008
Demographic controls		Yes
Observations	624	622
r^2	0.035	0.083

Reference: Tired

Discussion

Summary

- Using survey-based voice data with two emotion models (Emo-DB, ABC) leads to ∼largely consistent predictions
- Most common predicted emotion is boredom/ tired
- Some correlation between predicted emotion and number of words of survey response

Next steps

- Re-classify into neutral/ bored vs. everything else?
- Obtain predictions for additional items with voice response data
- Various research directions
  - Analyze effects on superficial responding, correlate with sentiment of response, ...

Thanks!

c.kern@uni-mannheim.de

References I

Conrad, F., Broome, J., Benki, J., Kreuter, F., Groves, R., Vannette, D., and Mcclain, C. (2013). Interviewer speech and the success of survey invitations. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 176:191–210.

Eyben, F., Wöllmer, M., and Schuller, B. (2009). OpenEAR – introducing the Munich open-source emotion and affect recognition toolkit. Paper presented at the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, 2009, 1–6.

Frank, M. G., Griffin, D. J., Svetieva, E., and Maroulis, A. (2015). Nonverbal elements of the voice. In Kostic, A. and Chadee, D., editors, *The Social Psychology of Nonverbal Communication*, pages 92–113. London, UK: Palgrave Macmillian.

Harmon-Jones, C., Bastian, B., and Harmon-Jones, E. (2016). The discrete emotions questionnaire: A new tool for measuring state self-reported emotions. *PLoS ONE*, 11:1–25.

Heide, M. and Gronhaug, K. (1991). Respondents' moods as a biasing factor in surveys: An experimental study. *Advances in Consumer Research*, 18:566–575.

Höhne, J. K., Gavras, K., and Qureshi, D. D. (forthcoming). SurveyVoice (SVoice): A comprehensive guide for recording voice answers in surveys. Zenodo.

References II

Schober, M. F., Conrad, F. G., Antoun, C., Ehlen, P., Fail, S., Hupp, A. L., Johnston, M., Vickers, L., Yan, H. Y., and Zhang, C. (2015). Precision and disclosure in text and voice interviews on smartphones. *PloS One*, 10:1–20.

Schuller, B., Zhang, Z., Weninger, F., and Burkhardt, F. (2012). Synthesized speech for model training in cross-corpus recognition of human emotion. *International Journal of Speech Technology*, 15:313–323.