

Emotion Recognition from Voice in the Wild

Oren Wright

Carnegie Mellon University Software Engineering Institute Copyright 2019 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon[®] is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM19-1106

Research Review 2019

Toward Machine-Emotional Intelligence









Proposed and Future



Passive Biometrics at the SEI

Real-Time Heartrate Extraction (2016)

Facial Micro-Expression Analysis (2017)



Used with permission of the Poker Channel: youtube.com/user/sergeypoker/

© 2019 Carnegie Mellon University

Research Review 2019

Voice Forensics at CMU Language Technologies Institute



U.S. Coast Guard photo by Eric D. Woodall

Profiling Hoax Callers R. Singh, et al., 2016

This person is:

- White
- Brought up in the U.S.A.
- Approx. 175 cm tall
- Approx. 75 kg
- Approx. 40 years old
- Not in any trouble
- Not on a boat
- In a warehouse of some kind
- Using homemade equipment
- Sitting on a metal chair upon a concrete floor

Emotion Recognition from Voice



- Voice is a complex process that presents bio-markers
- Bio-marker analysis enabled by
 micro-articulometry
- Made possible by 30 years of automatic speech recognition technology at CMU

Mission Applications



- Security checkpoints and encounters
- Interrogations
- Intelligence profiling
- Media analysis and exploitation
- Detection of stress, PTSD
- Human-machine teaming

Research Review 2019

Emotion Recognition from Voice





Database Construction



Broekens & Brinkman, 2009



nrabian & Russell, 1974

© 2019 Carnegie Mellon University

CMU-SER Database

- Largest ever in-the-wild speech emotion database
 - Over 29,000 annotated audio clips, totaling over 54 hours of voice recordings
 - Over **324,000** unique annotations
- Open source tools
 - Voice processing and exemplar creation
 - Crowdsourcing platform with Amazon Mechanical Turk



Al for Super-Intelligent Hearing: Deducing Human Emotion Status from Voice

Rita Singh

Language Technologies Institute School of Computer Science Carnegie Mellon University
 Introduction
 Motivation
 Methodology
 Results
 Perception
 Future Work
 Questions

 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •</

Motivation: Voice and Emotions



Methodology Introduction Motivation Results Perception Future Work Questions \bigcirc \bigcirc \bigcirc \bigcirc

Motivation: Discrete Emotions



Fearful



Angry



Sad





Disgusted



Surprised



Al for Super-Intelligent Hearing SEI Research Review 2019

Happy



Arousal (calm vs. excited)

arousal, dominance)

 Introduction
 Motivation
 Methodology
 Results
 Perception
 Future Work
 Questions

 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •
 •</

Methodology: Affect Button



Introduction

Motivation

Methodology

 $\bullet \bullet \circ \circ$

Results

Perception

Future Work

 \bigcirc

Questions

Methodology: Data Collection & Crowdsourcing

- Data selection (internet archive, news channels, etc.)
- Data cleaning (noisy utterances, music, manual annotations, etc.)
- Pre-hoc quality control
- Challenges of crowdsourcing:
 - Re-sampling
 - Filtering

Clip 1/5		?		
REPLAY AUDIO	[PLAY/PAL	JSE	
Audio clip progress				1
			YES	NO
Clear and audible speech? ⑦			YES	NO
Clear and audible speech? ⑦	0	1	YES ②	NO © > 2
Clear and audible speech? ⑦ Number of People Audible ⑦	0	1	YES © 2 ©	NO 0 >2
Clear and audible speech? ⑦ Number of People Audible ⑦	0 O Male	1 © Female	YES 2 Both	NO > 2 0 Non

 Introduction
 Motivation
 Methodology
 Results
 Perception
 Future Work
 Questions

 Introduction
 Image: Comparison of the second sec

Methodology: Model I (speaker-specific embeddings)





Methodology: Model II (multimodal embedding network)



AI for Super-Intelligent Hearing SEI Research Review 2019

Introduction

Motivation____

Methodology

Results Perception

 \bigcirc

Future Work

 \bigcirc

Questions

Results

speaker (data: C	speaker specific embedding network (data: CMU-SER, task: regression)				
	MAE		RMSE	Ratings on the Big Five model Agricultioness Conscientiousness	
Baseline	0.26		0.31	Neuroficiam Denniess to Experience	
Proposed Model	0.16		0.20	Extraversion General Traits	
multim (data: IEI	odal embec MOCAP, tas	dding ne sk: class	etwork sification)	Leadersvip Sociability Behavioral Traits	
			Accuracy	Mildheos Aggrosskoheos	
Baseline		30.8%			
Proposed Model		53%			



Snippet of a lecture by Christopher Manning marked as "neutral" in terms of emotion by our demo

Introduction

Motivation Methodology

Results



()

Questions ()

Perception

Expression versus Perception •

- Are emotions expressed through lacksquarespeech universal in how they are perceived?
- Studying gender differences in • perception of emotion
- Challenges of crowdsourced data in lacksquarerelation to statistical analysis

Detecting gender differences in perception of emotion in crowdsourced data

Shahan Ali Memon **Hira Dhamyal** Carnegie Mellon University Pittsburgh, USA {samemon,hyd}@cs.cmu.edu

Oren Wright Daniel Justice Vijavkumar Palat William Boler Software Engineering Institute Carnegie Mellon University Pittsburgh, USA

Bhiksha Raj **Rita Singh** Carnegie Mellon University Pittsburgh, USA {bhiksha,rsingh}@cs.cmu.edu

{owright,dljustice,vpalat,wmboler}@sei.cmu.edu

ABSTRACT

Do men and women perceive emotions differently? Popular convictions place women as more emotionally perceptive than men. Empirical findings, however, remain inconclusive. Most prior studies focus on visual modalities. In addition, almost all of the studies are limited to experiments within controlled environments. Generalizability and scalability of these studies has not been sufficiently established. In this paper, we study the differences in perception of emotion between genders from speech data in the wild, annotated through crowdsourcing. While we limit ourselves to a single modality (i.e. speech), our framework is applicable to studies of emotion perception from all such loosely annotated data in general. Our paper addresses multiple serious challenges related to making statistically viable conclusions from crowdsourced data. Overall, the contributions of this paper are two fold: a reliable novel framework for perceptual studies from crowdsourced data; and the demonstration of statistically significant differences in

and interpret emotions [30]. It is important to emphasize that perception, while correlated to emotional expression, is different from it. This distinction can be best explained by the modified version of Brunswik's lens model proposed in [37]. There are three main stages in this model: the encoding, the transmission, and the decoding of emotions. Encoding is the process where individual conveys their internal state by modifying their communicative channel. Decoding is the process where another individual makes an inference about the state of the first individual. The cues that are encoded and the cues that are decoded may differ based on the noise in the transmission. When studying expression, the focus is on how the emotions were encoded, and the primary subject of study is the encoder. On the other hand, perception deals with how the emotions were interpreted or decoded, and, hence, the focus is on the decoder. To create emotionally intelligent machines that can interact with humans, understanding how humans perceive emotions is a crucial first step. Not only Introduction Motivation Methodology Results Perception Future Work Questions

Future Work: Phonetic Embeddings

phoneme segementer



AI for Super-Intelligent Hearing SEI Research Review 2019

Introduction

Motivation

Methodology

Results Perception

Future Work

Questions

Questions?

Rita Singh (LTI, ECE) (rsingh@cs.cmu.edu)

The Amazing Team

Hira Dhamyal (LTI) Shahan Ali Memon (LTI) Bhiksha Raj (LTI, ECE) Richard Stern (ECE, LTI)

Oren Wright (SEI) **Daniel Justice (SEI)** Vijaykumar Palat (SEI) William Boller (SEI)

Al for Super-Intelligent Hearing SEI Research Review 2019