The VoicePrivacy 2020 Challenge

Odyssey 2020

Subjective evaluation-2

Presenters: **Anais Chanclu & Benjamin O'Brien**

Natalia Tomashenko [1]
Brij M.L. Srivastava [2]
Xin Wang [3]
Emmanuel Vincent [2]
Andreas Nautsch [4]
Junichi Yamagishi [3,5]
Nicholas Evans [4]
Jose Patino [4]
Jean-François Bonastre [1]
Paul-Gauthier Noé [1]
Massimiliano Todisco [4]
Mohamed Maouche [2]
Benjamin O'Brien [6]
Anais Chanclu [1]

[1] LIA – University of Avignon – France
[2] Inria – France
[3] NII – Tokyo – Japan
[4] Audio Security and Privacy Group, EURECOM – France
[5] University of Edinburgh – UK
[6] LPL – Aix-Marseille University – France

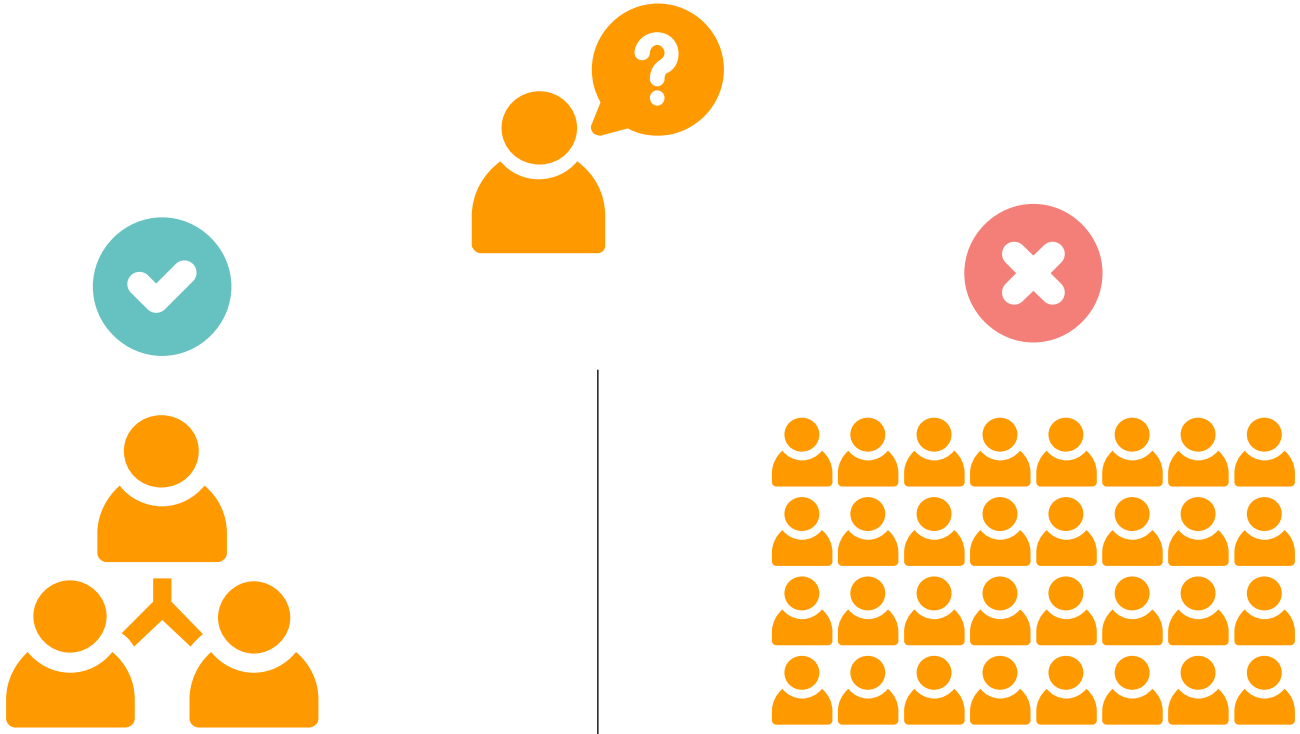
4th November 2020

Hypotheses

How do humans perfom voice identification?

• New York (5:30 - 7.30) • UK (10:30 -12.30) • France (11:30 -13.30) • India (16:00 -18:00) • China (18:30 - 20.30) • Japan (19:30 - 21:30)

Voice identification by humans

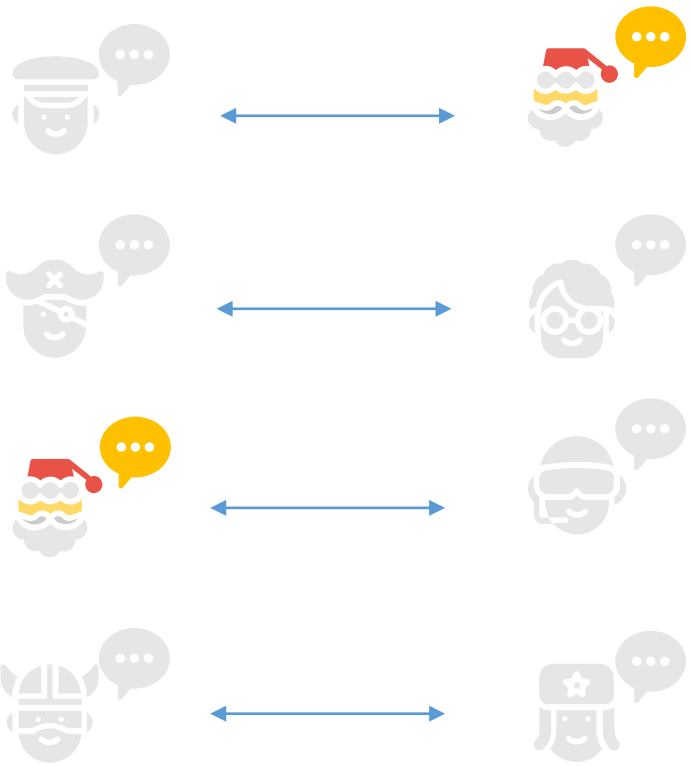


• New York (5:30 - 7.30) • UK (10:30 -12.30) • France (11:30 -13.30) • India (16:00 -18:00) • China (18:30 - 20.30) • Japan (19:30 - 21:30)

Voice identification by humans

• New York (5:30 - 7.30) • UK (10:30 -12.30) • France (11:30 -13.30) • India (16:00 -18:00) • China (18:30 - 20.30) • Japan (19:30 - 21:30)

Voice identification by humans



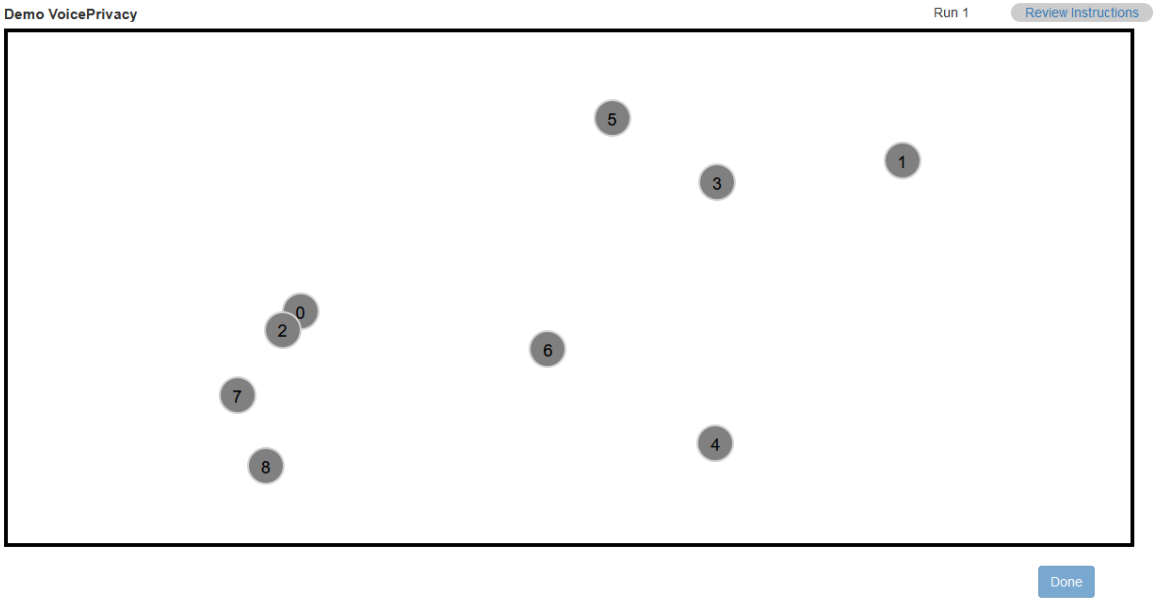
• New York (5:30 - 7.30) • UK (10:30 -12.30) • France (11:30 -13.30) • India (16:00 -18:00) • China (18:30 - 20.30) • Japan (19:30 - 21:30)

Protocol

How do humans cluster speakers?

• New York (5:30 - 7.30) • UK (10:30 -12.30) • France (11:30 -13.30) • India (16:00 -18:00) • China (18:30 - 20.30) • Japan (19:30 - 21:30)

1 session, 3 trials



Recordings are presented to the listeners. Their objective is to classify the recordings into clusters which supposedly correspond to different speakers. However, a part of the recordings are anonymised which affects the voice quality.

• New York (5:30 - 7.30) • UK (10:30 -12.30) • France (11:30 -13.30) • India (16:00 -18:00) • China (18:30 - 20.30) • Japan (19:30 - 21:30)

Data

The data come from the VCTK corpus which is composed of English sentences read by native speakers.

Speakers

A selection of 15 female speakers and 15 male speakers has been made. The 15 speakers of each gender are divided into two groups:

1. 9 reference speakers
2. 6 distractors

Utterances

We only use the 24 first utterances pronounced.

To reduce the trial duration, the utterances are split after 3 seconds of speech. The reduced recordings are called "chunks".

Trials

Trial composition

- 16 different chunks
- 3 reference speakers
- 1 distractor
- 2-6 chunks per speaker

1 control trial

None of the chunks have been anonymised.

2 evaluation trials

- 7 anonymized chunks for the reference speakers
- 1 anonymized chunk for the distractor
- All anonymised chunks have been processed by the same system

• New York (5:30 - 7.30) • UK (10:30 -12.30) • France (11:30 -13.30) • India (16:00 -18:00) • China (18:30 - 20.30) • Japan (19:30 - 21:30)

Listeners

13

26

9

24

• New York (5:30 - 7.30) • UK (10:30 -12.30) • France (11:30 -13.30) • India (16:00 -18:00) • China (18:30 - 20.30) • Japan (19:30 - 21:30)

Results

Based on 39 listeners

• New York (5:30 - 7.30) • UK (10:30 -12.30) • France (11:30 -13.30) • India (16:00 -18:00) • China (18:30 - 20.30) • Japan (19:30 - 21:30)

Tier Analysis

Tier I

Compare performance **between** trials (Control, Evaluation)
→ original + anonymous recordings

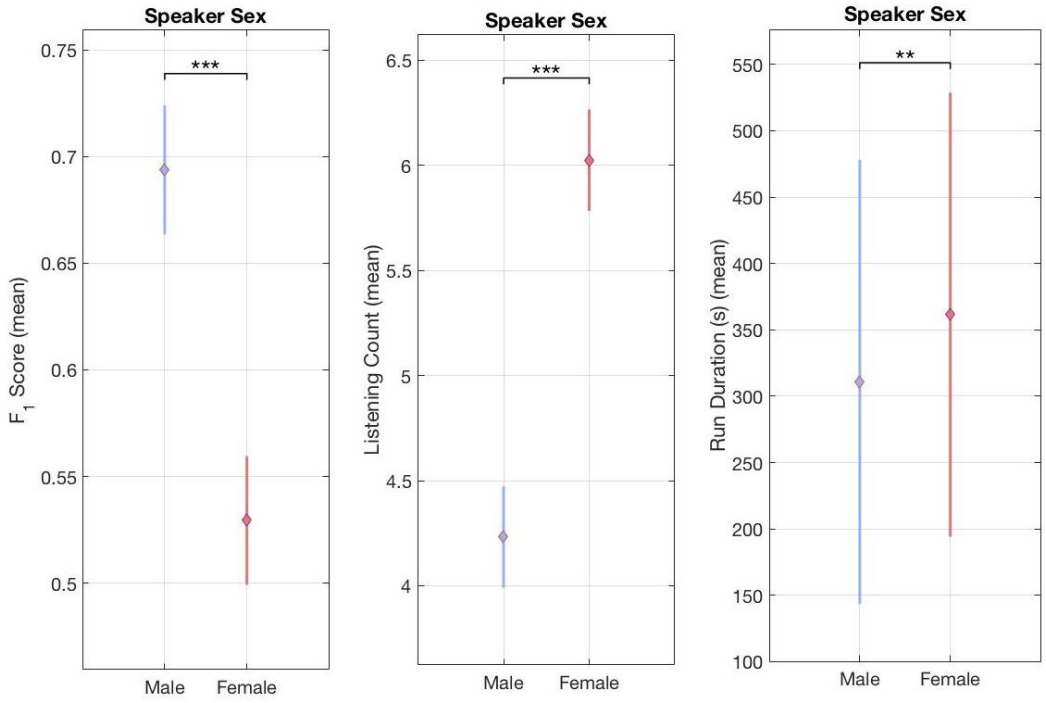
Tier II

Compare performance **between** anonymous recordings (Evaluation)
→ *only* anonymous recordings

Anova factors: Trial type, speaker gender
Response Variables (RV): F_1 Score, Listening Count, Listening Duration, Move Duration

Tier I Findings

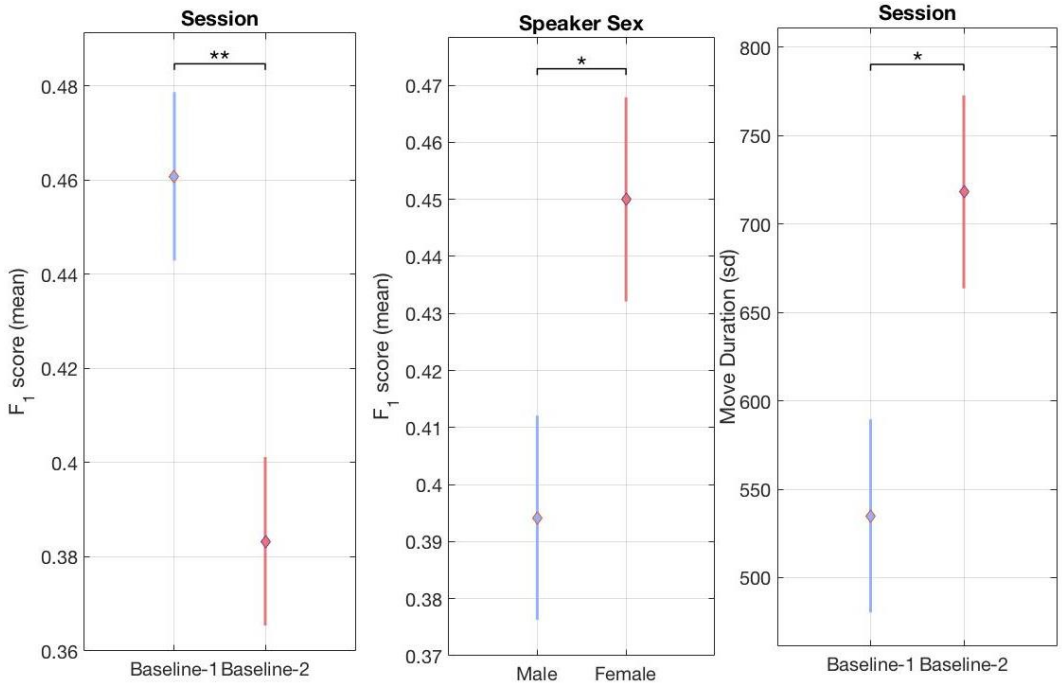
		Trial Type	Speaker Gender	Trial x Gender
F_1 Score	Mean		$p < 0.001$	
	Std		$p < 0.001$	
Listening Count	Mean		$p < 0.001$	
	Std		$p < 0.001$	
Listening Duration	Mean		$p < 0.01$	
	Std		$p < 0.01$	
Move Duration	Mean			
	Std			



• New York (5:30 - 7.30) • UK (10:30 -12.30) • France (11:30 -13.30) • India (16:00 -18:00) • China (18:30 - 20.30) • Japan (19:30 - 21:30)

Tier II Findings

		Trial Type	Speaker Gender	Session x Gender
F_1 Score	Mean	$p < 0.01$	$p < 0,05$	
	Std			
Listening Count	Mean			
	Std			
Listening Duration	Mean			
	Std			
Move Duration	Mean			
	Std	$p < 0.05$		



• New York (5:30 - 7.30) • UK (10:30 -12.30) • France (11:30 -13.30) • India (16:00 -18:00) • China (18:30 - 20.30) • Japan (19:30 - 21:30)

Clustering homogeneity measure

Female speakers

Type of trial	Mean	Std
Control	89.14%	10.91
Baseline-1	60.76%	10.18
Baseline-2	59.38%	12.28

Male speakers

Type of trial	Mean	Std
Control	87.50%	11.71
Baseline-1	62.50%	11.20
Baseline-2	63.44%	10.89

$$homogeneity = \frac{number\ of\ chunks\ in\ the\ corresponding\ clusters}{total\ number\ of\ chunks}$$

Thanks for your attention!

Do you have any questions?

• New York (5:30 - 7.30) • UK (10:30 -12.30) • France (11:30 -13.30) • India (16:00 -18:00) • China (18:30 - 20.30) • Japan (19:30 - 21:30)