# The VoxCeleb Speaker Recognition Challenge 2022 (VoxSRC-22)

NAVER | LINE

KAIST

Visual AI

VGG University of Oxford

University of Oxford

# Workshop Programme

| KST | |
|---|---|
| **17:00** | Introduction: “**VoxCeleb, VoxConverse & VoxSRC**” |
| **17:25** | Keynote speech : Junichi Yamagishi “The use of speaker embeddings in neural audio generation” |
| **18:15** | Coffee break |
| **18:25** | Announcement of Winners (Track 1,2 and 3) |
| **18:30** | Invited Talks from Track 1 and 2 |
| **19:10** | Announcement of Winners (Track 3) |
| **19:12** | Invited Talks from Track 3 |
| **19:30** | Announcement of Winners (Track 4) |
| **19:35** | Invited Talks from Track 4 |
| **19:55** | Wrap up discussion and conclusion |

# Organisers

Jaesung Huh

Andrew Brown

Joon Son Chung

Arsha Nagrani

Jee-weon Jung

Andrew Zisserman

Daniel Garcia-Romero

## Advisors

Mitch Mclaren

Doug Reynolds

# Workshop Programme

| KST | |
|---|---|
| **17:00** | Introduction: “**VoxCeleb, VoxConverse & VoxSRC**” |
| **17:25** | Keynote speech : Junichi Yamagishi “The use of speaker embeddings in neural audio generation” |
| **18:15** | Coffee break |
| **18:25** | Announcement of Winners (Track 1,2 and 3) |
| **18:30** | Invited Talks from Track 1 and 2 |
| **19:10** | Announcement of Winners (Track 3) |
| **19:12** | Invited Talks from Track 3 |
| **19:30** | Announcement of Winners (Track 4) |
| **19:35** | Invited Talks from Track 4 |
| **19:55** | Wrap up discussion and conclusion |

# Introduction

- ***Data***: *VoxCeleb and VoxConverse*

- ***Challenge Mechanics:*** *tracks, rules and metrics*

# VoxCeleb datasets

- Multi-speaker environments
- Varying audio quality and background channel noise
- Freely available
  - ○ https://mm.kaist.ac.kr/datasets/voxceleb

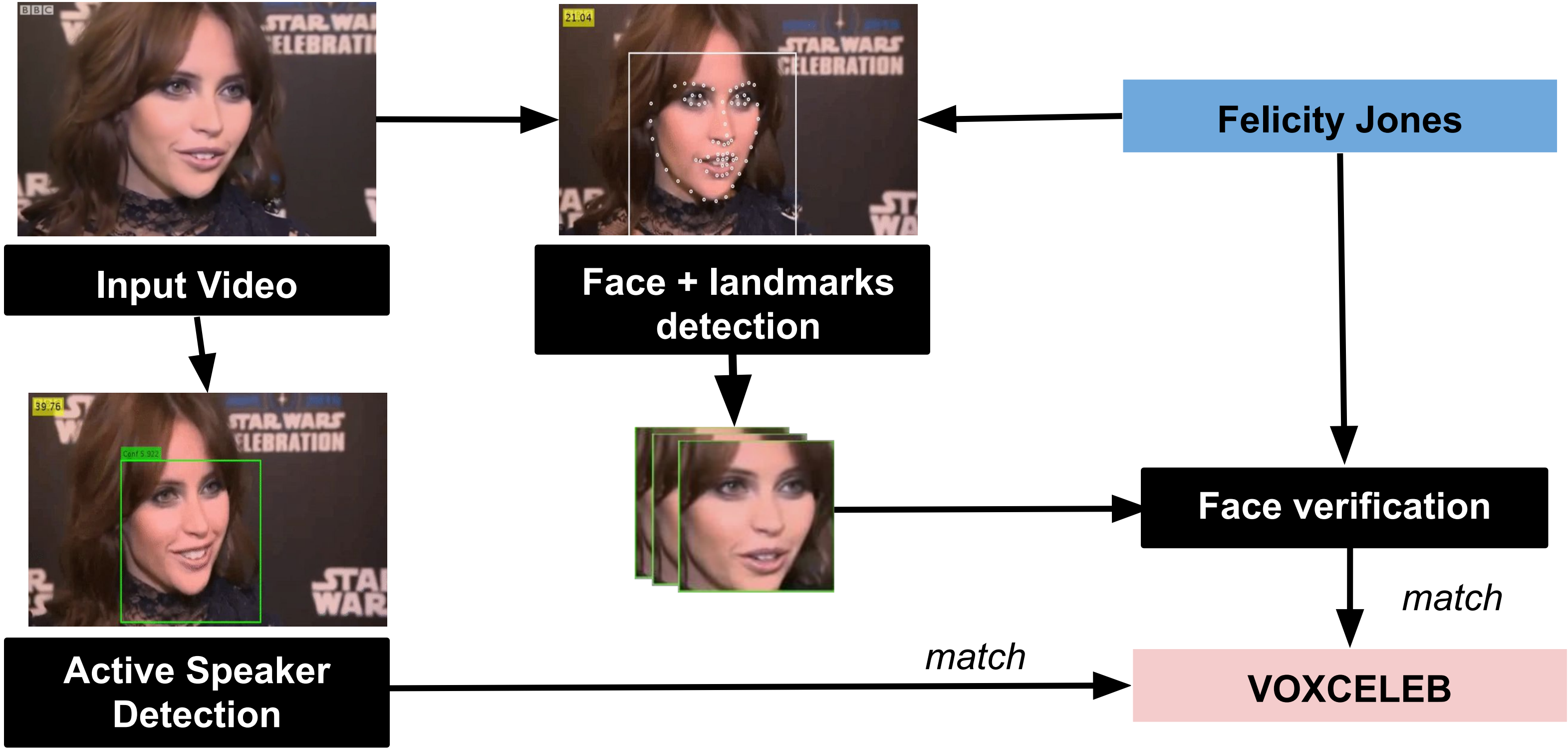
Red Carpet Interviews

Studio Interviews

Outdoor and pitch Interviews

# VoxCeleb - automatic pipeline

Transferring labels from Vision to Speech



Input Video

Face + landmarks detection

Felicity Jones

Active Speaker Detection

Face verification

match

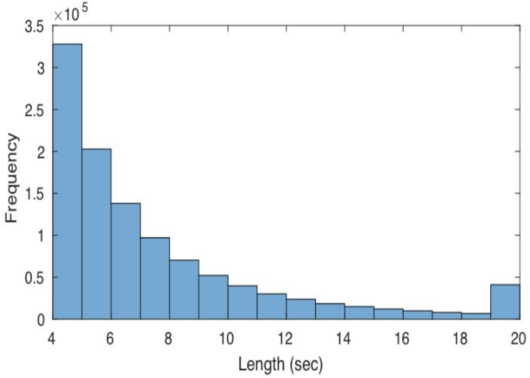
match

VOXCELEB

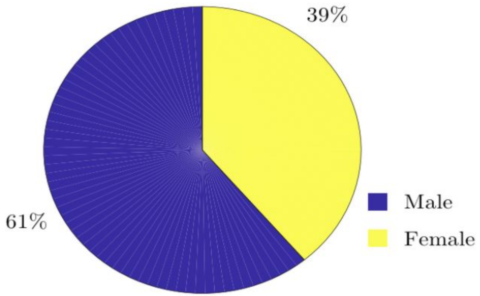
# VoxCeleb Statistics

- VoxCeleb2 dev set -> primary data for speaker verification
- Validation toolkit for scoring

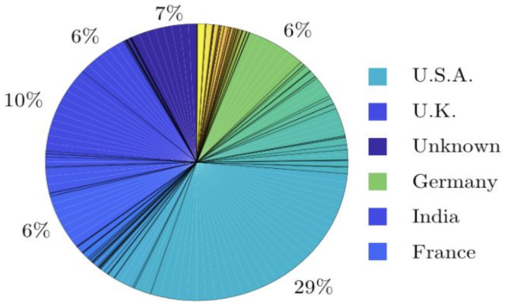
| | Train | Validation |
|---|---|---|
| # Speakers | 5,994 | 1,251 |
| # Utterances | 1,092,009 | 153,516 |



Utterance Lengths



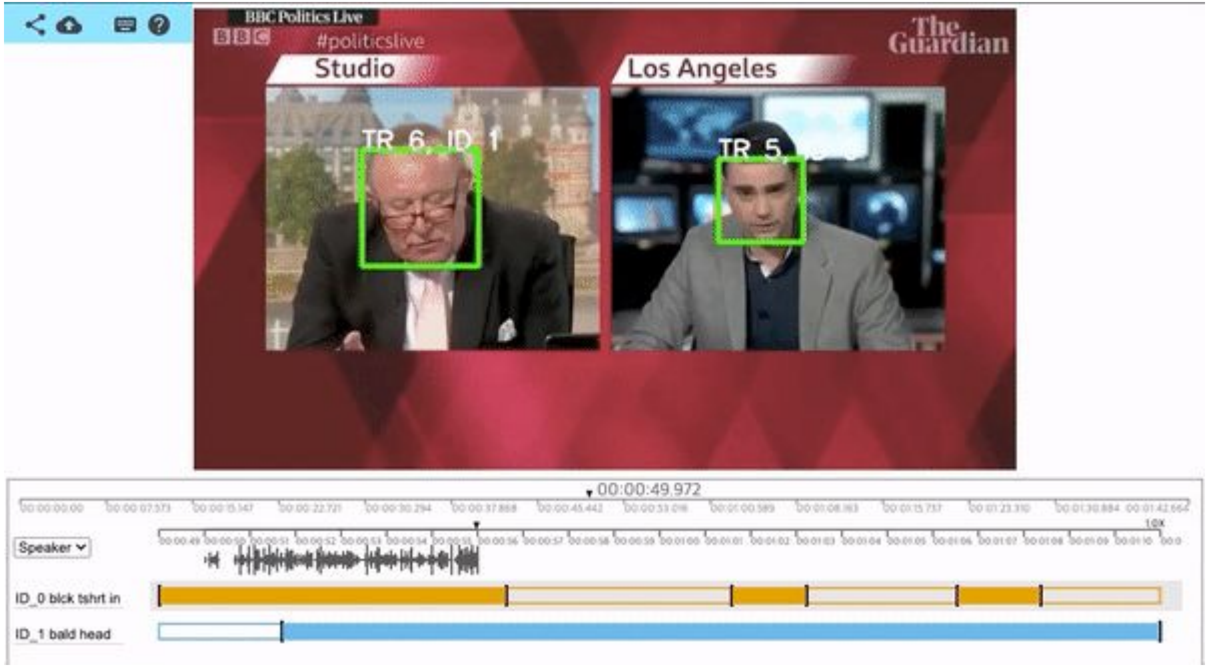
Gender Distribution



Nationality Distribution

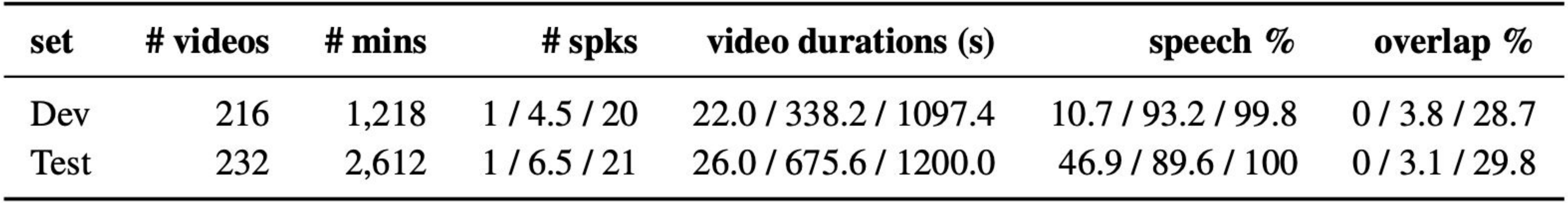
# Audio speaker diarization

- Solving “who spoke when” in multi-speaker video.



# Diarization - The VoxConverse dataset

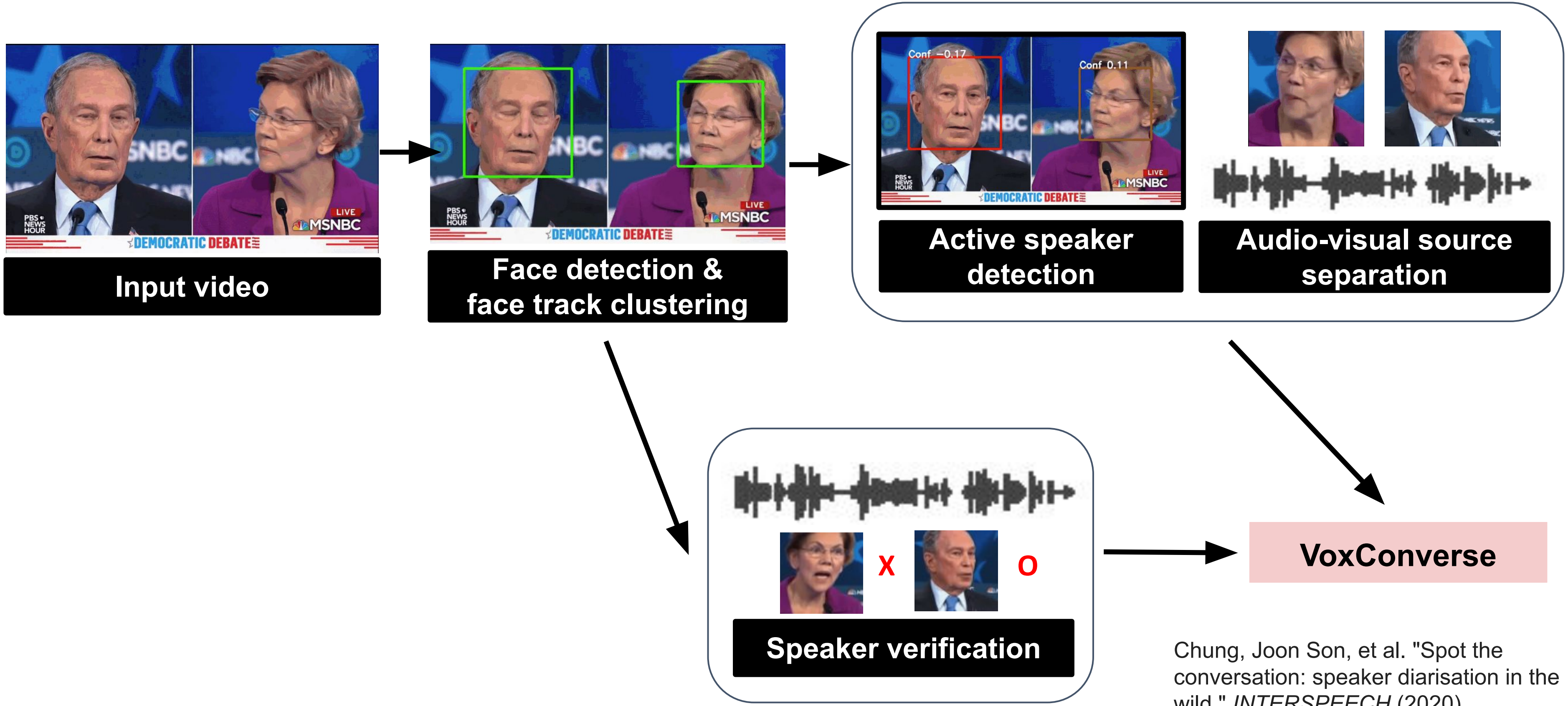
- Videos from YouTube
- Mostly debates, talk shows, news segments



| set | # videos | # mins | # spks | video durations (s) | speech % | overlap % |
|---|---|---|---|---|---|---|
| Dev | 216 | 1,218 | 1 / 4.5 / 20 | 22.0 / 338.2 / 1097.4 | 10.7 / 93.2 / 99.8 | 0 / 3.8 / 28.7 |
| Test | 232 | 2,612 | 1 / 6.5 / 21 | 26.0 / 675.6 / 1200.0 | 46.9 / 89.6 / 100 | 0 / 3.1 / 29.8 |

http://www.robots.ox.ac.uk/~vgg/data/voxconverse/

# Automatic audio-visual diarization method



Input video

Face detection & face track clustering

Active speaker detection

Audio-visual source separation

Speaker verification

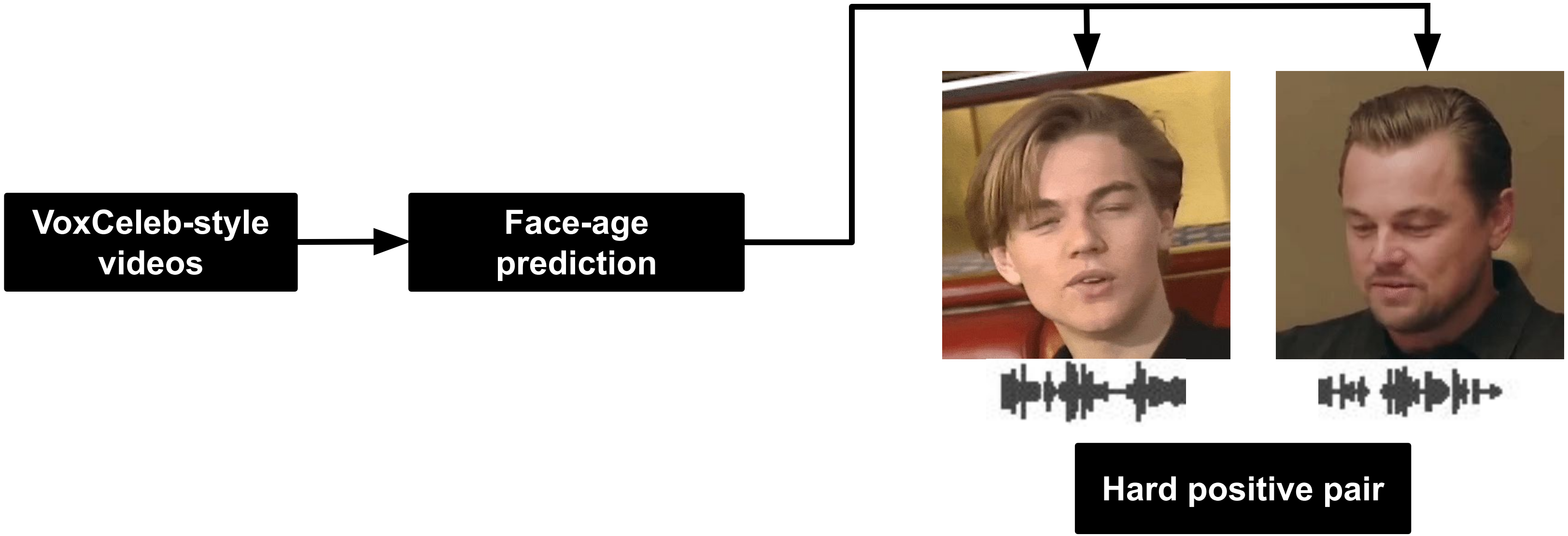
VoxConverse

Chung, Joon Son, et al. "Spot the conversation: speaker diarisation in the wild." *INTERSPEECH* (2020).

# The VoxCeleb Speaker Recognition Challenge

# New challenging settings for Speaker Verification

- **Harder positives:** We focus on how speech segments taken from the same speaker at different ages impact speaker verification systems



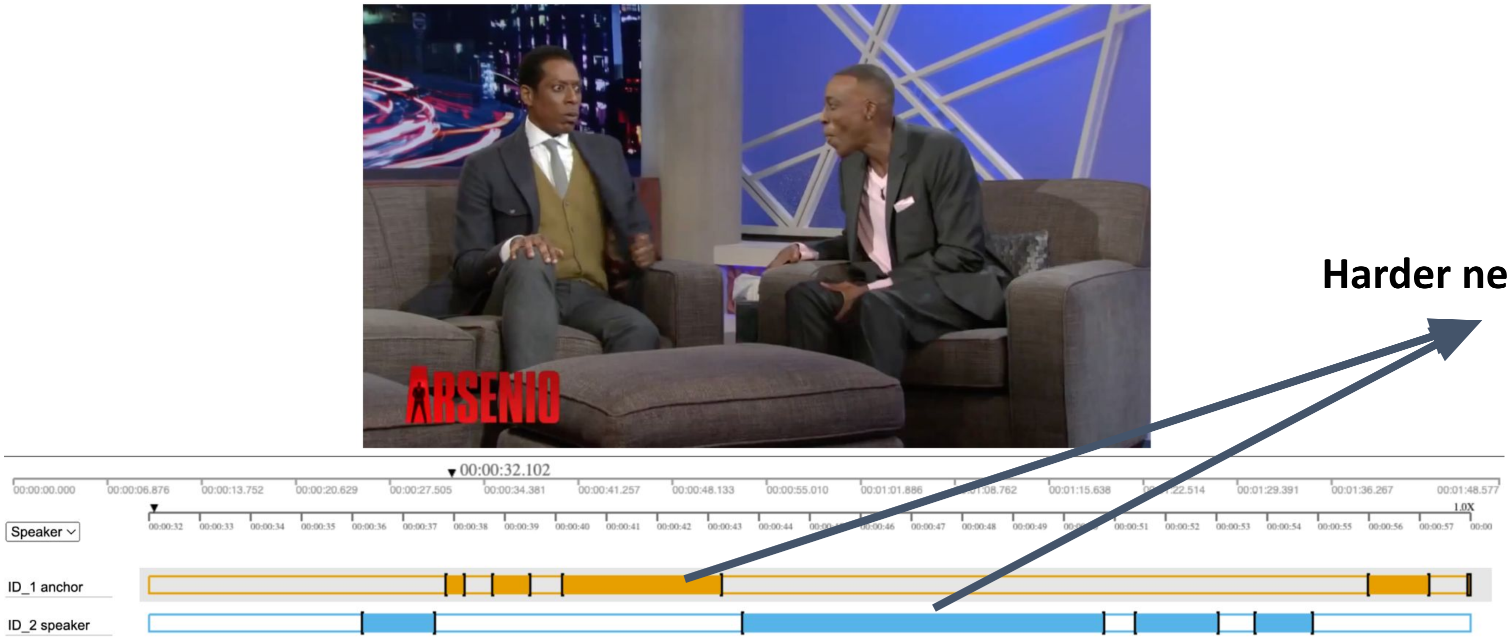
VoxCeleb-style videos

Face-age prediction

Hard positive pair

# New challenging settings for Speaker Verification

- **Harder negatives:** We focus on how speaker verification systems perform when speech segments from different speakers have the same background noise in **VoxConverse** dataset



**Harder negatives**

# New Semi-supervised domain adaptation track

- Propose a problem of how models, pre-trained on a large set of data with labels in a source domain, can adapt to a new target domain given:
    - a large set of unlabeled data from the target domain
    - a small set of labeled data from the target domain.

# New Semi-supervised domain adaptation track

- Domain adaptation in speaker verification from one language in a source domain (English), to a different language in a target domain (Chinese).
- Source domain : VoxCeleb2
- Target domain : CN-Celeb

**We would like to thank CN-Celeb authors for providing such a valuable dataset!**

# VoxSRC-2022 tracks

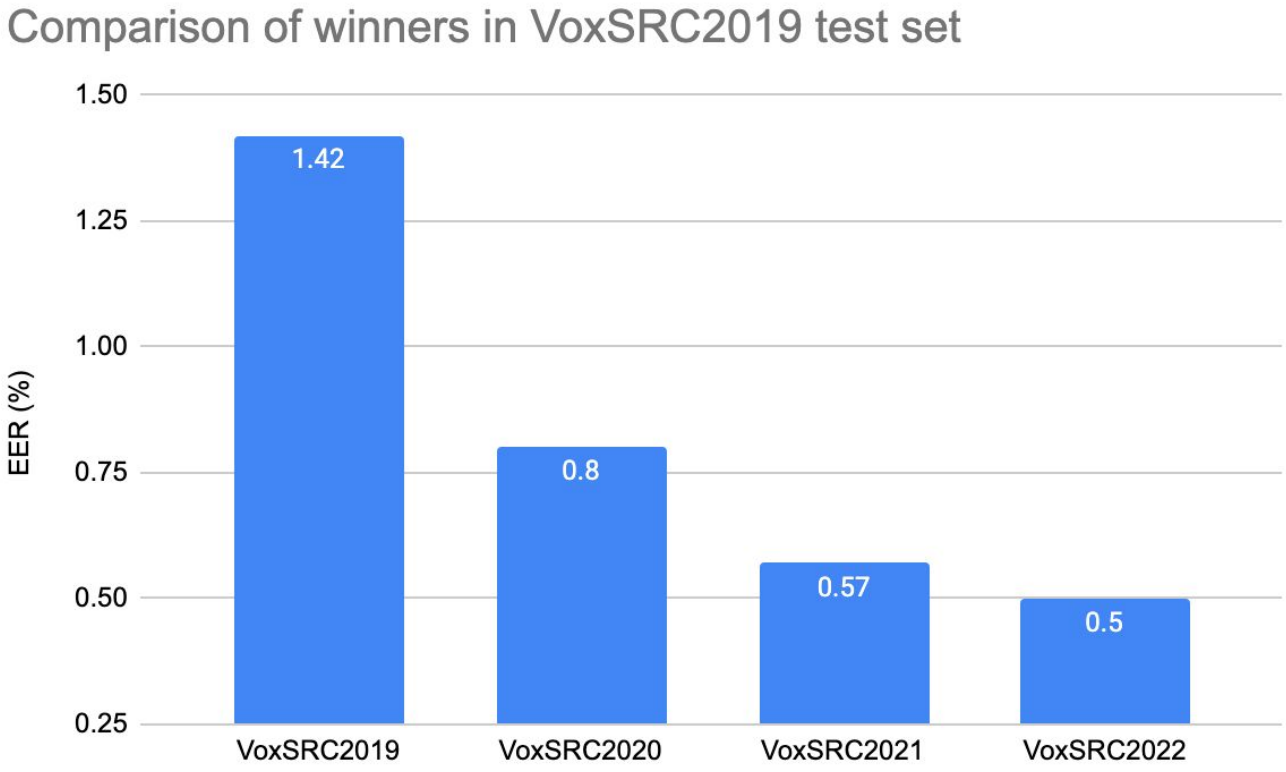
- **Track 1 :** Supervised speaker verification (closed)
- **Track 2 :** Supervised speaker verification (open)
- **Track 3 :** *Semi-Supervised Domain Adaptation* (closed)
  - Domain adaptation task on language domain
- **Track 4 :** Speaker *diarization* (open)
  - Solving “who spoke when” in multi-speaker video.
  - Speaker overlap, challenging background conditions

# Mechanics

- Metrics (Tracks 1-3)
  - **DCF (Tracks 1 & 2), EER (Track 3)**
  - Following NIST-SRE 2018
- Metrics (Track 4)
  - **DER**, JER
  - Overlapping speech counted, collar of 0.25s
- Only 1 submission per day, 10 in total
- Submissions via CodaLab

# Progress over time

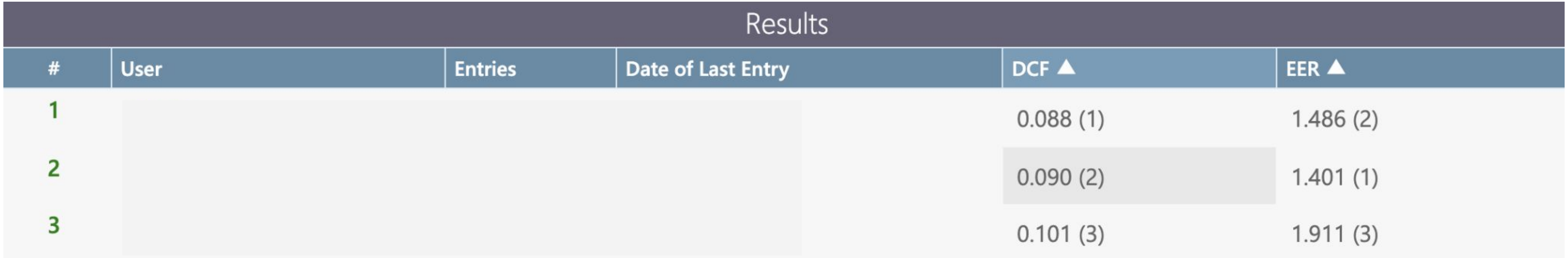
VoxSRC



**Lower is better**

# Performance gap between Track 1 & 2

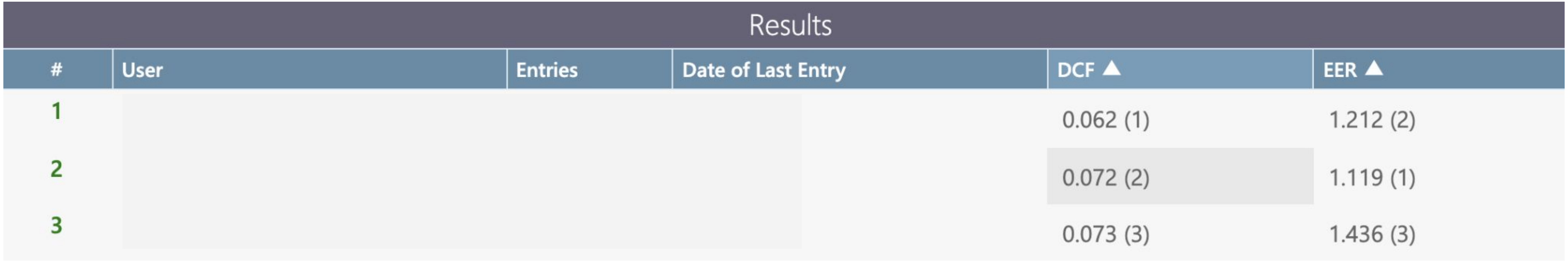
## Track 1 winners (Closed track)



Results

| # | User | Entries | Date of Last Entry | DCF ▲ | EER ▲ |
|---|---|---|---|---|---|
| 1 | | | | 0.088 (1) | 1.486 (2) |
| 2 | | | | 0.090 (2) | 1.401 (1) |
| 3 | | | | 0.101 (3) | 1.911 (3) |

## Track 2 winners (Open track)



Results

| # | User | Entries | Date of Last Entry | DCF ▲ | EER ▲ |
|---|---|---|---|---|---|
| 1 | | | | 0.062 (1) | 1.212 (2) |
| 2 | | | | 0.072 (2) | 1.119 (1) |
| 3 | | | | 0.073 (3) | 1.436 (3) |

# Workshop Programme

| KST | |
|---|---|
| **17:00** | Introduction: “**VoxCeleb, VoxConverse & VoxSRC**” |
| **17:25** | Keynote speech : Junichi Yamagishi “The use of speaker embeddings in neural audio generation” |
| **18:15** | Coffee break |
| **18:25** | Announcement of Winners (Track 1,2 and 3) |
| **18:30** | Invited Talks from Track 1 and 2 |
| **19:10** | Announcement of Winners (Track 3) |
| **19:12** | Invited Talks from Track 3 |
| **19:30** | Announcement of Winners (Track 4) |
| **19:35** | Invited Talks from Track 4 |
| **19:55** | Wrap up discussion and conclusion |

Keynote:

# The use of speaker embeddings in neural audio generation

Junichi Yamagishi

National Institute of Informatics

# Q & A

# Workshop Programme

VoxSRC

| KST | |
|---|---|
| **17:00** | Introduction: “**VoxCeleb, VoxConverse & VoxSRC**” |
| **17:25** | Keynote speech : Junichi Yamagishi “The use of speaker embeddings in neural audio generation” |
| **18:15** | Coffee break |
| **18:25** | Announcement of Winners (Track 1,2 and 3) |
| **18:30** | Invited Talks from Track 1 and 2 |
| **19:10** | Announcement of Winners (Track 3) |
| **19:12** | Invited Talks from Track 3 |
| **19:30** | Announcement of Winners (Track 4) |
| **19:35** | Invited Talks from Track 4 |
| **19:55** | Wrap up discussion and conclusion |

# Coffee Break

Restarting at  18:25 Korea time

# Workshop Programme

| KST | |
|---|---|
| **17:00** | Introduction: “**VoxCeleb, VoxConverse & VoxSRC**” |
| **17:25** | Keynote speech : Junichi Yamagishi “The use of speaker embeddings in neural audio generation” |
| **18:15** | Coffee break |
| **18:25** | Announcement of Winners (Track 1,2 and 3) |
| **18:30** | Invited Talks from Track 1 and 2 |
| **19:10** | Announcement of Winners (Track 3) |
| **19:12** | Invited Talks from Track 3 |
| **19:30** | Announcement of Winners (Track 4) |
| **19:35** | Invited Talks from Track 4 |
| **19:55** | Wrap up discussion and conclusion |

# VoxSRC Winners - Track1 Speaker verification [closed training set]

VoxSRC

- 1st Team ravana (ID R&D lab)

  Rostislav Makarov, Alexander Alenin, iVan Iakovlev, Anton Okhotnikov, Nikita Torgashov

- 2nd Team KristonAI (Kriston AI Lab)

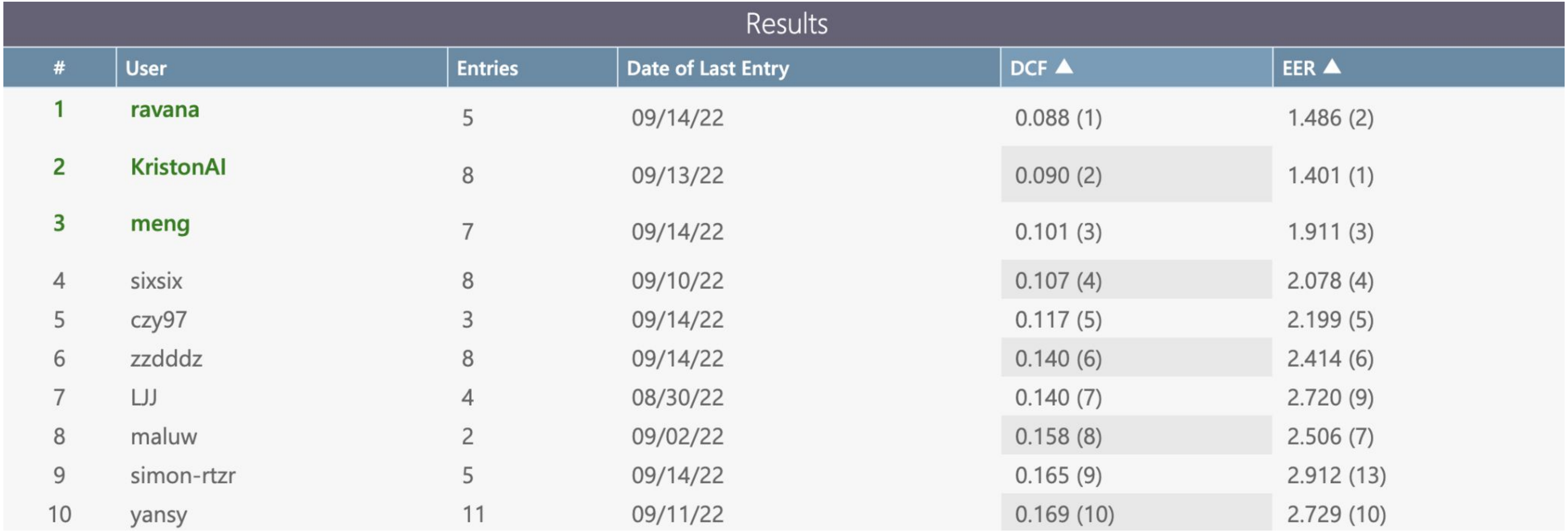
  Qutang Cai, Guoqiang Hong, Zhijian Ye, Ximin Li, Haizhou Li

- 3rd Team SJTU-AISPEECH (Shanghai Jiao Tong University, AISpeech)

  Zhengyang Chen, Bing Han, Xu Xiang, Houjun Huang, Bei Liu, Yanmin Qian

77 participants, 39 teams submitted, 198 total submissions

# Track 1 - Speaker verification (closed)



Results

| # | User | Entries | Date of Last Entry | DCF ▲ | EER ▲ |
|---|---|---|---|---|---|
| 1 | ravana | 5 | 09/14/22 | 0.088 (1) | 1.486 (2) |
| 2 | KristonAI | 8 | 09/13/22 | 0.090 (2) | 1.401 (1) |
| 3 | meng | 7 | 09/14/22 | 0.101 (3) | 1.911 (3) |
| 4 | sixsix | 8 | 09/10/22 | 0.107 (4) | 2.078 (4) |
| 5 | czy97 | 3 | 09/14/22 | 0.117 (5) | 2.199 (5) |
| 6 | zzdddz | 8 | 09/14/22 | 0.140 (6) | 2.414 (6) |
| 7 | LJJ | 4 | 08/30/22 | 0.140 (7) | 2.720 (9) |
| 8 | maluw | 2 | 09/02/22 | 0.158 (8) | 2.506 (7) |
| 9 | simon-rtzr | 5 | 09/14/22 | 0.165 (9) | 2.912 (13) |
| 10 | yansy | 11 | 09/11/22 | 0.169 (10) | 2.729 (10) |

77 participants, 39 teams submitted, 198 total submissions

# VoxSRC Winners - Track2 Speaker verification [open training set]

VoxSRC

- 1st Team ravana (ID R&D lab)

  Rostislav Makarov, Alexander Alenin, iVan Iakovlev, Anton Okhotnikov, Nikita Torgashov

- 2nd Team KristonAI (Kriston AI Lab)

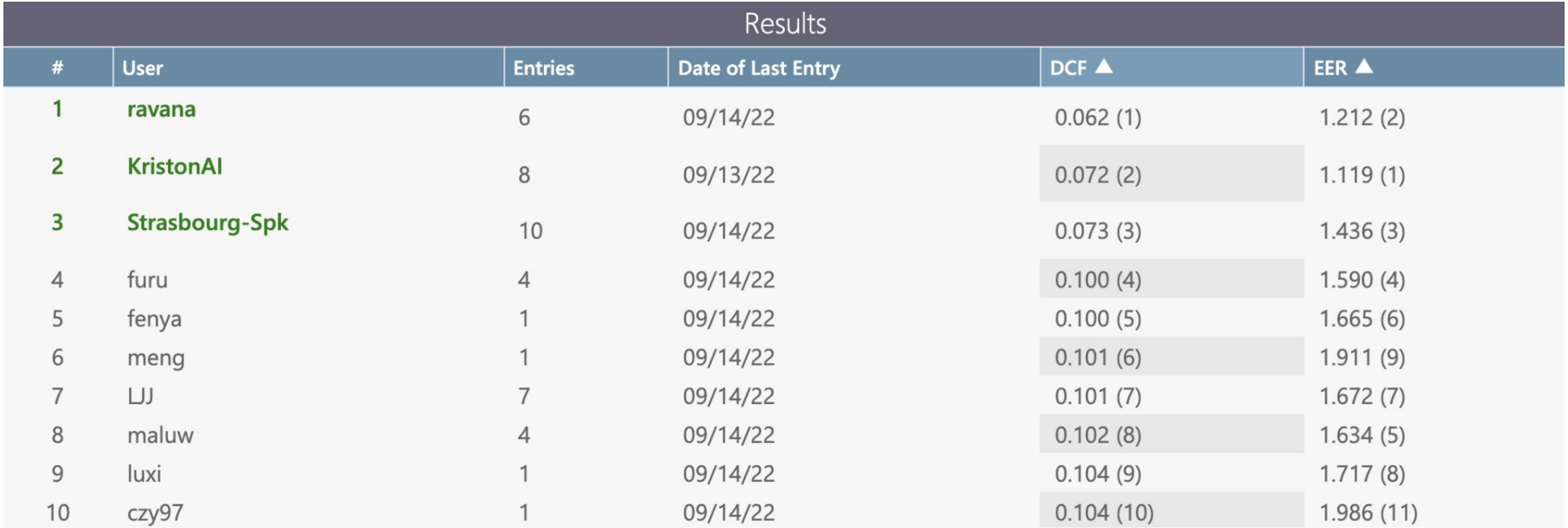
  Qutang Cai, Guoqiang Hong, Zhijian Ye, Ximin Li, Haizhou Li

- 3rd Team Strasbourg-spk (Microsoft)

  Gang Liu, Tianyan Zhou, Yong Zhao, Yu Wu, Zhuo Chen, Yao Qian, Jian Wu

67 participants, 35 teams submitted, 166 total submissions

# Track 2 - Speaker verification (open)



Results

| # | User | Entries | Date of Last Entry | DCF ▲ | EER ▲ |
|---|---|---|---|---|---|
| 1 | ravana | 6 | 09/14/22 | 0.062 (1) | 1.212 (2) |
| 2 | KristonAI | 8 | 09/13/22 | 0.072 (2) | 1.119 (1) |
| 3 | Strasbourg-Spk | 10 | 09/14/22 | 0.073 (3) | 1.436 (3) |
| 4 | furu | 4 | 09/14/22 | 0.100 (4) | 1.590 (4) |
| 5 | fenya | 1 | 09/14/22 | 0.100 (5) | 1.665 (6) |
| 6 | meng | 1 | 09/14/22 | 0.101 (6) | 1.911 (9) |
| 7 | LJJ | 7 | 09/14/22 | 0.101 (7) | 1.672 (7) |
| 8 | maluw | 4 | 09/14/22 | 0.102 (8) | 1.634 (5) |
| 9 | luxi | 1 | 09/14/22 | 0.104 (9) | 1.717 (8) |
| 10 | czy97 | 1 | 09/14/22 | 0.104 (10) | 1.986 (11) |

67 participants, 35 teams submitted, 166 total submissions

VoxSRC

# Workshop Programme

| KST | |
|---|---|
| **17:00** | Introduction: “**VoxCeleb, VoxConverse & VoxSRC**” |
| **17:25** | Keynote speech : Junichi Yamagishi “The use of speaker embeddings in neural audio generation” |
| **18:15** | Coffee break |
| **18:25** | Announcement of Winners (Track 1,2 and 3) |
| **18:30** | Invited Talks from Track 1 and 2 |
| **19:10** | Announcement of Winners (Track 3) |
| **19:12** | Invited Talks from Track 3 |
| **19:30** | Announcement of Winners (Track 4) |
| **19:35** | Invited Talks from Track 4 |
| **19:55** | Wrap up discussion and conclusion |

# Talks by the winners

1. Team ravana (ID R&D lab) - tracks 1,2
2. Team KristonAI (Kriston AI) - tracks 1,2
3. Team meng - track 1
4. Team Starsbourg-Spk - track 2

# Team ravana

In-person talk

# Team KristonAI

# Team meng

# Team Strasbourg-spk

# Workshop Programme

| KST | |
|---|---|
| **17:00** | Introduction: “**VoxCeleb, VoxConverse & VoxSRC**” |
| **17:25** | Keynote speech : Junichi Yamagishi “The use of speaker embeddings in neural audio generation” |
| **18:15** | Coffee break |
| **18:25** | Announcement of Winners (Track 1,2 and 3) |
| **18:30** | Invited Talks from Track 1 and 2 |
| **19:10** | Announcement of Winners (Track 3) |
| **19:12** | Invited Talks from Track 3 |
| **19:30** | Announcement of Winners (Track 4) |
| **19:35** | Invited Talks from Track 4 |
| **19:55** | Wrap up discussion and conclusion |

# Track 3: Semi-supervised domain adaptation

- Domain adaptation task in speaker verification from one language in a source domain, to a different language in a target domain.
- Primary metric : EER (%)

# VoxSRC Winners - Track 3
# Semi-supervised domain adaptation

VoxSRC

- 1st Team zzdddz (Chinese Academy of Science)

  Zhenduo Zhao, Zhuo Li, Wenchao Wang

- 2nd Team sixsix (Duke Kunshan University, Tencent AI)

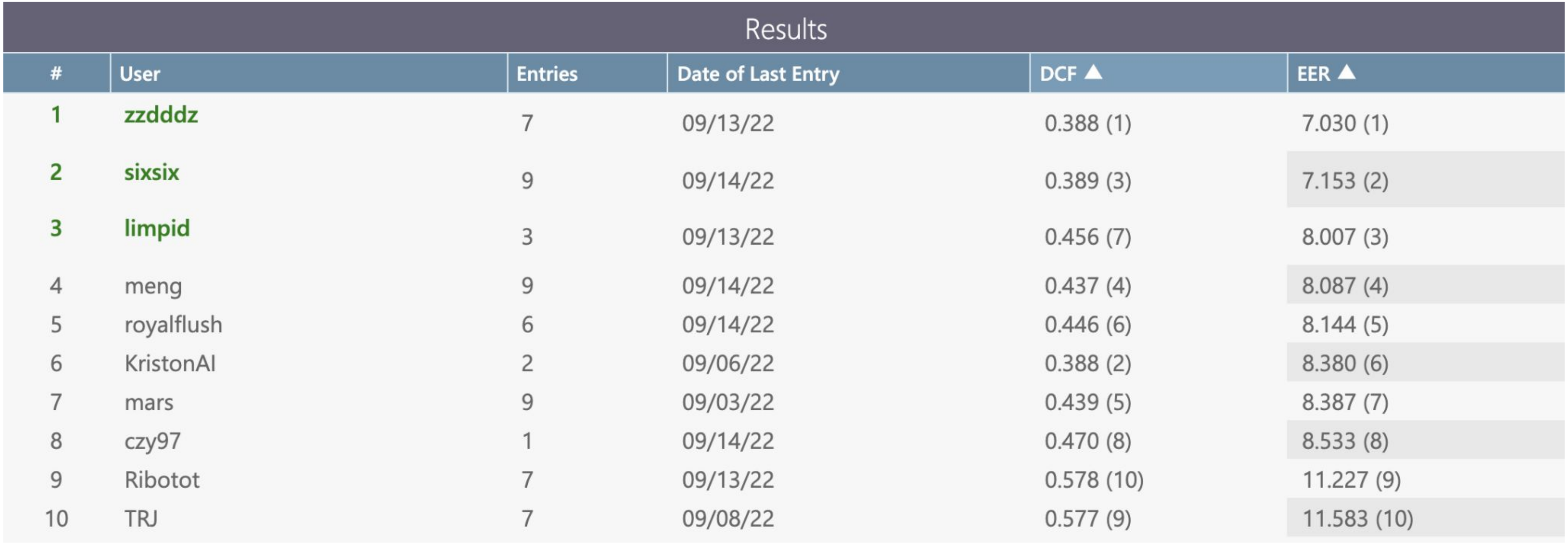
  Xiaoyi Qin, Na Li, Yuke Lin, Yiwei Ding, Chao Weng, Dan Su, Ming Li

- 3rd Team SJTU-AISPEECH (Shanghai Jiao Tong University, AISpeech)

  Zhengyang Chen, Bing Han, Xu Xiang, Houjun Huang, Bei Liu, Yanmin Qian

42 participants, 12 teams submitted, 89 total submissions

# Track 3 - Semi Supervised Speaker Verification



Results

| # | User | Entries | Date of Last Entry | DCF ▲ | EER ▲ |
|---|---|---|---|---|---|
| 1 | zzdddz | 7 | 09/13/22 | 0.388 (1) | 7.030 (1) |
| 2 | sixsix | 9 | 09/14/22 | 0.389 (3) | 7.153 (2) |
| 3 | limpid | 3 | 09/13/22 | 0.456 (7) | 8.007 (3) |
| 4 | meng | 9 | 09/14/22 | 0.437 (4) | 8.087 (4) |
| 5 | royalflush | 6 | 09/14/22 | 0.446 (6) | 8.144 (5) |
| 6 | KristonAI | 2 | 09/06/22 | 0.388 (2) | 8.380 (6) |
| 7 | mars | 9 | 09/03/22 | 0.439 (5) | 8.387 (7) |
| 8 | czy97 | 1 | 09/14/22 | 0.470 (8) | 8.533 (8) |
| 9 | Ribotot | 7 | 09/13/22 | 0.578 (10) | 11.227 (9) |
| 10 | TRJ | 7 | 09/08/22 | 0.577 (9) | 11.583 (10) |

42 participants, 12 teams submitted, 89 total submissions

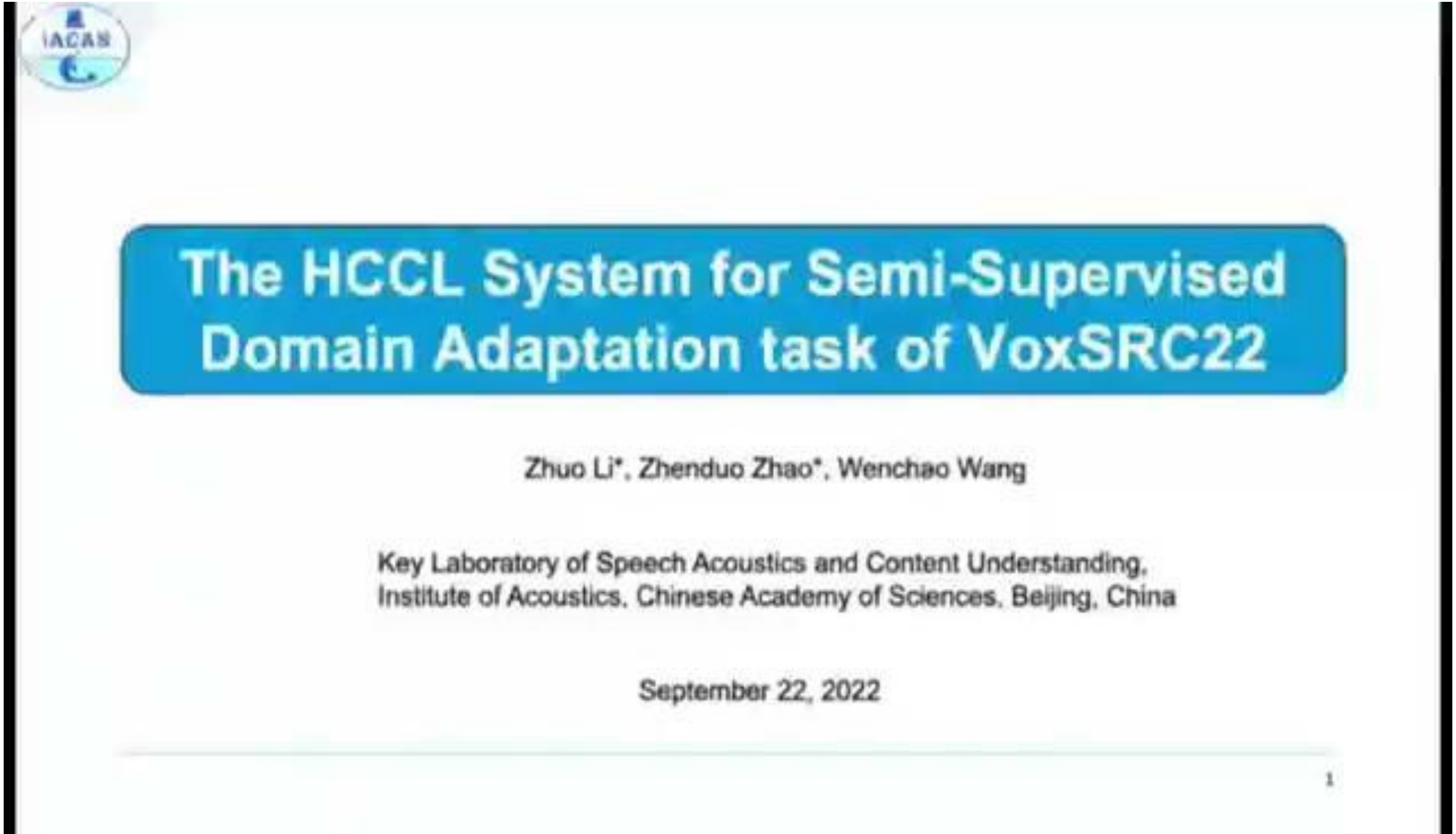
# Workshop Programme

| KST | |
|---|---|
| **17:00** | Introduction: “**VoxCeleb, VoxConverse & VoxSRC**” |
| **17:25** | Keynote speech : Junichi Yamagishi “The use of speaker embeddings in neural audio generation” |
| **18:15** | Coffee break |
| **18:25** | Announcement of Winners (Track 1,2 and 3) |
| **18:30** | Invited Talks from Track 1 and 2 |
| **19:10** | Announcement of Winners (Track 3) |
| **19:12** | Invited Talks from Track 3 |
| **19:30** | Announcement of Winners (Track 4) |
| **19:35** | Invited Talks from Track 4 |
| **19:55** | Wrap up discussion and conclusion |

# Talks by the winners

1. Team zzdddz
2. Team sixsix

Team
zzdddz



# The HCCL System for Semi-Supervised Domain Adaptation task of VoxSRC22

Zhuo Li*, Zhenduo Zhao*, Wenchao Wang

Key Laboratory of Speech Acoustics and Content Understanding,
Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

September 22, 2022

1

# Team sixsix

# Workshop Programme

| KST | |
|---|---|
| **17:00** | Introduction: “**VoxCeleb, VoxConverse & VoxSRC**” |
| **17:25** | Keynote speech : Junichi Yamagishi “The use of speaker embeddings in neural audio generation” |
| **18:15** | Coffee break |
| **18:25** | Announcement of Winners (Track 1,2 and 3) |
| **18:30** | Invited Talks from Track 1 and 2 |
| **19:10** | Announcement of Winners (Track 3) |
| **19:12** | Invited Talks from Track 3 |
| **19:30** | Announcement of Winners (Track 4) |
| **19:35** | Invited Talks from Track 4 |
| **19:55** | Wrap up discussion and conclusion |

# Track 4: Speaker Diarization

- Solving “who spoke when”
- VoxConverse - data from debates, talk shows from YouTube
- Primary metric - Diarization Error Rate (DER)

# 🥇 VoxSRC Winners - Track 4 Speaker diarisation

VoxSRC

- 1st Team dkusmiip (Duke University)

  Weiqing Wang, Xiaoyi Qin, Ming Cheng, Yucong Zhang, Kangyu Wang, Ming Li

- 2nd Team KristonAI (KristonAI lab)

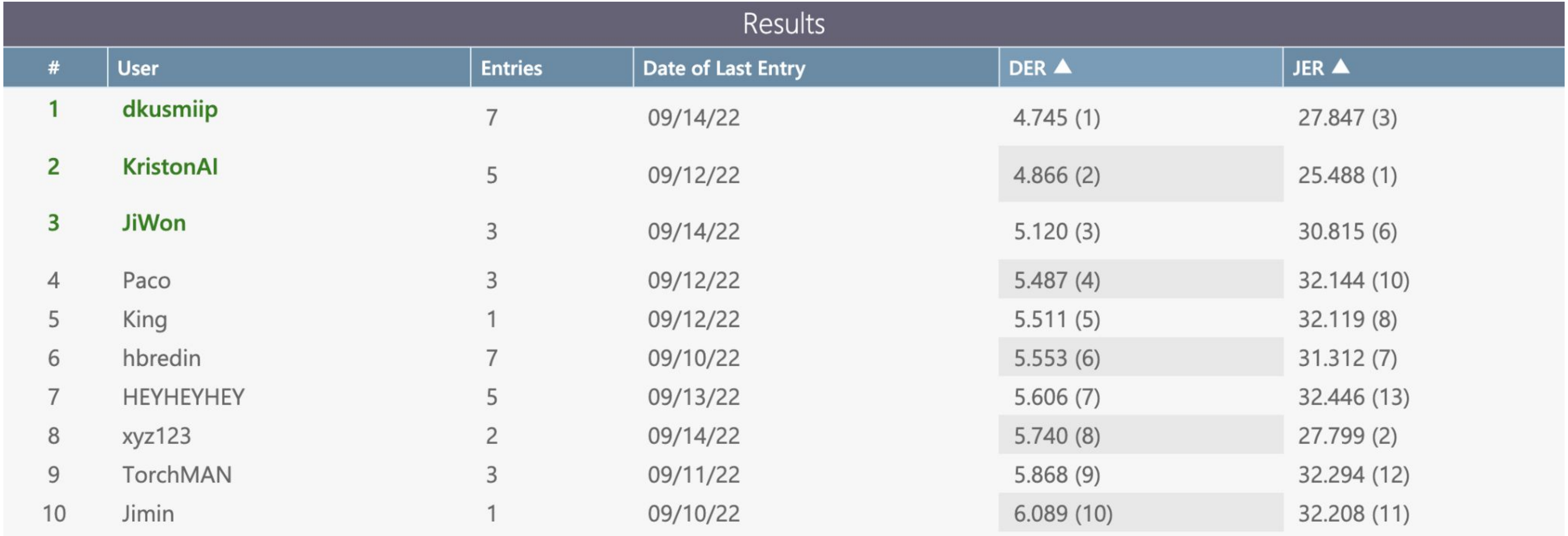
  Qutang Cai, Guoqiang Hong, Zhijian Ye, Ximin Li, Haizhou Li

- 3rd Team AiTeR (GIST)

  Dongkeon Park, Yechan Yu, Keyeongwan Park, Jiwon Kim, Hongkook Kim

44 participants, 17 teams submitted, 101 total submissions

# Track 4 - Speaker Diarization (open)



Results

| # | User | Entries | Date of Last Entry | DER ▲ | JER ▲ |
|---|---|---|---|---|---|
| 1 | dkusmiip | 7 | 09/14/22 | 4.745 (1) | 27.847 (3) |
| 2 | KristonAI | 5 | 09/12/22 | 4.866 (2) | 25.488 (1) |
| 3 | JiWon | 3 | 09/14/22 | 5.120 (3) | 30.815 (6) |
| 4 | Paco | 3 | 09/12/22 | 5.487 (4) | 32.144 (10) |
| 5 | King | 1 | 09/12/22 | 5.511 (5) | 32.119 (8) |
| 6 | hbredin | 7 | 09/10/22 | 5.553 (6) | 31.312 (7) |
| 7 | HEYHEYHEY | 5 | 09/13/22 | 5.606 (7) | 32.446 (13) |
| 8 | xyz123 | 2 | 09/14/22 | 5.740 (8) | 27.799 (2) |
| 9 | TorchMAN | 3 | 09/11/22 | 5.868 (9) | 32.294 (12) |
| 10 | Jimin | 1 | 09/10/22 | 6.089 (10) | 32.208 (11) |

44 participants, 17 teams submitted, 101 total submissions

# Workshop Programme

VoxSRC

| KST | |
|---|---|
| **17:00** | Introduction: “**VoxCeleb, VoxConverse & VoxSRC**” |
| **17:25** | Keynote speech : Junichi Yamagishi “The use of speaker embeddings in neural audio generation” |
| **18:15** | Coffee break |
| **18:25** | Announcement of Winners (Track 1,2 and 3) |
| **18:30** | Invited Talks from Track 1 and 2 |
| **19:10** | Announcement of Winners (Track 3) |
| **19:12** | Invited Talks from Track 3 |
| **19:30** | Announcement of Winners (Track 4) |
| **19:35** | Invited Talks from Track 4 |
| **19:55** | Wrap up discussion and conclusion |

# Talks by the winners

1. Team dkusmiip (Duke Kunshan University)
2. Team KristonAI (KristonAI)
3. Team JiWON (GIST)

# Team dkusmiip

Team KristonAI

# Team AiTeR

In-person talk

# Workshop Programme

VoxSRC

| KST | |
|---|---|
| **17:00** | Introduction: “**VoxCeleb, VoxConverse & VoxSRC**” |
| **17:25** | Keynote speech : Junichi Yamagishi “The use of speaker embeddings in neural audio generation” |
| **18:15** | Coffee break |
| **18:25** | Announcement of Winners (Track 1,2 and 3) |
| **18:30** | Invited Talks from Track 1 and 2 |
| **19:10** | Announcement of Winners (Track 3) |
| **19:12** | Invited Talks from Track 3 |
| **19:30** | Announcement of Winners (Track 4) |
| **19:35** | Invited Talks from Track 4 |
| **19:55** | Wrap up discussion and conclusion |

# Thank you!

Please email us at voxsrc@gmail.com

Feedback, suggestions welcome!

See you all at VoxSRC-2023