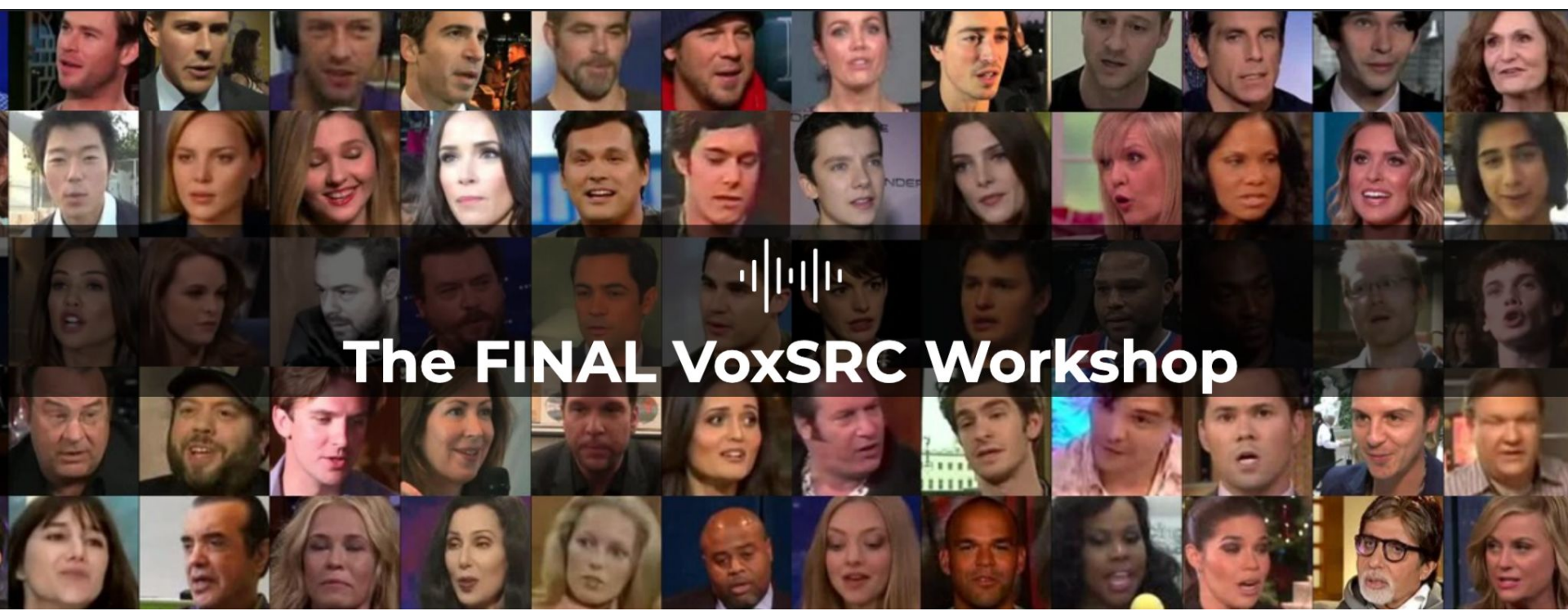


KAIST

Visual AI

VGG
University of Oxford

University of Oxford

Workshop Programme

VoxSRC

17:30	Introduction: “**VoxCeleb, VoxConverse & VoxSRC**”
17:40	Reflecting on our last five years
18:00	Keynote speech : Wei-Ning Hsu “Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations”
18:45	Coffee break
19:00	Announcement of Winners (Track 1,2 and 3)
19:05	Invited Talks from Track 1 and 2
19:30	Announcement of Winners (Track 3)
19:32	Invited Talks from Track 3
19:50	Announcement of Winners (Track 4)
19:55	Invited Talks from Track 4
20:20	Wrap up discussion and conclusion

Workshop Programme

VoxSRC

17:30 Introduction: "**VoxCeleb, VoxConverse & VoxSRC**" - **Joonson**

17:40 Reflecting our last five years - **Daniel**

18:00 Keynote speech : Wei-Ning Hsu "Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations" - **Jeeweon**

18:45 Coffee break

19:00 Announcement of Winners (Track 1,2 and 3) - **Jaesung**

- Invited Talks from Track 1 and 2 - 6+3 minutes per talk * **3 – we need names of presenters**

19:30 Announcement of Winners (Track 3) - **Jaesung**

- Invited Talks from Track 3 - 6+3 minutes per talk * **2**

19:50 Announcement of Winners (Track 4) - **Jaesung**

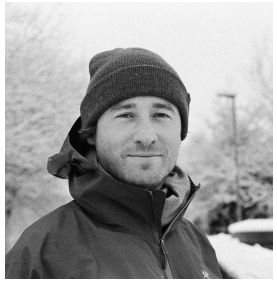
- Invited Talks from Track 4 - 6+3 minutes per talk * **3**

20:20 Wrap up discussion and conclusion - **Everyone**?

VoxSRC

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

VoxSRC

17:30	Introduction: “**VoxCeleb, VoxConverse & VoxSRC**”
17:40	Reflecting our last five years
18:00	Keynote speech : Wei-Ning Hsu “Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations”
18:45	Coffee break
19:00	Announcement of Winners (Track 1,2 and 3)
19:05	Invited Talks from Track 1 and 2
19:30	Announcement of Winners (Track 3)
19:32	Invited Talks from Track 3
19:50	Announcement of Winners (Track 4)
19:55	Invited Talks from Track 4
20:20	Wrap up discussion and conclusion

Introduction of VoxSRC

- Two major tasks
  - Speaker verification : identifying whether two speakers are the same person.
  - Speaker diarisation : solving ‘who spoke when’.

VoxCeleb datasets

VoxSRC

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

Red Carpet Interviews

Studio Interviews

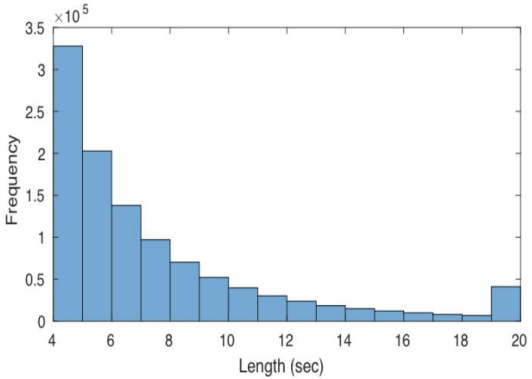
Outdoor and pitch Interviews

VoxCeleb Statistics

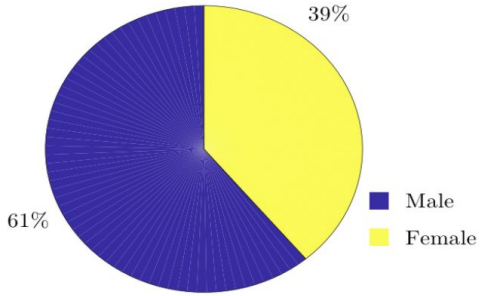
VoxSRC

- VoxCeleb2 dev set -> primary data for speaker verification
- Validation toolkit for scoring
- > 10k downloads

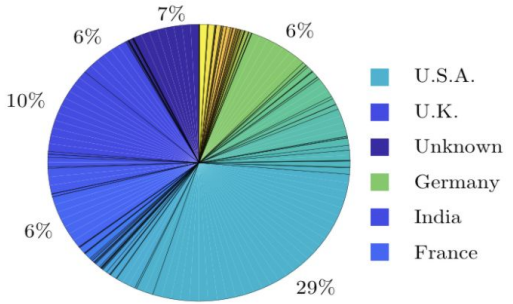
	Train	Validation
# Speakers	5,994	1,251
# Utterances	1,092,009	153,516



Utterance Lengths



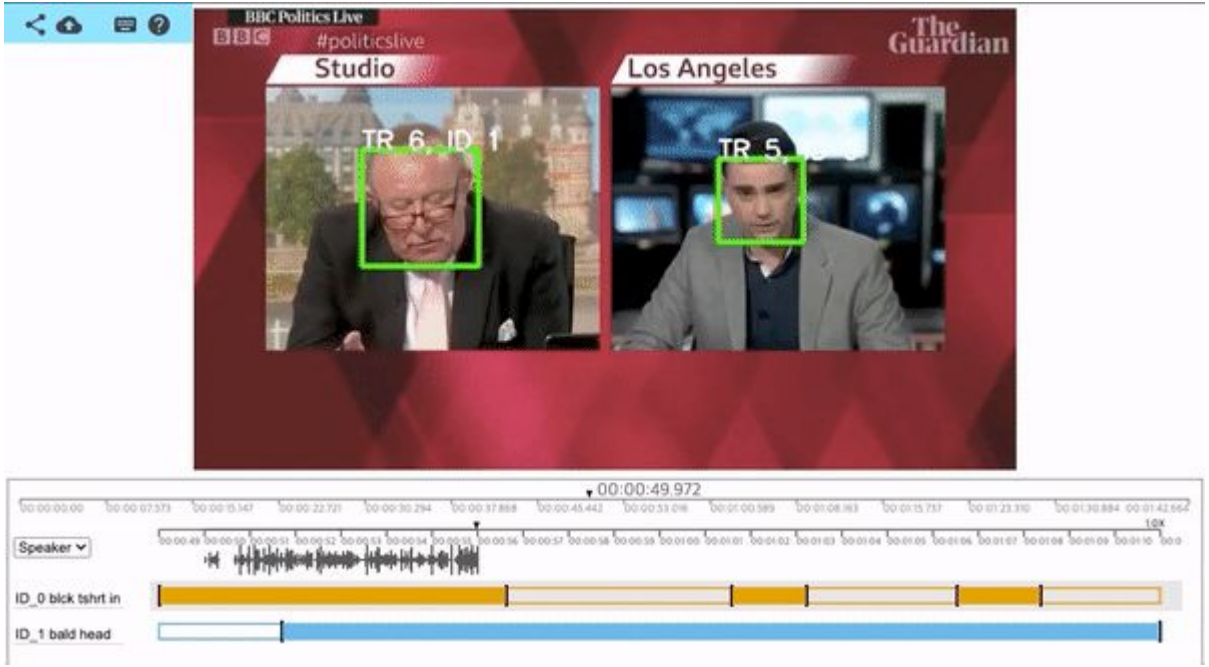
Gender Distribution



Nationality Distribution

Audio speaker diarization

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



Diarization - The VoxConverse dataset

VoxSRC

- 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/

VoxSRC-2023 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)
  - **minDCF (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

Workshop Programme

VoxSRC

17:30	Introduction: “VoxCeleb, VoxConverse & VoxSRC”
17:40	**Reflecting over last five years**
18:00	Keynote speech : Wei-Ning Hsu “Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations”
18:45	Coffee break
19:00	Announcement of Winners (Track 1,2 and 3)
19:05	Invited Talks from Track 1 and 2
19:30	Announcement of Winners (Track 3)
19:32	Invited Talks from Track 3
19:50	Announcement of Winners (Track 4)
19:55	Invited Talks from Track 4
20:20	Wrap up discussion and conclusion

Reflecting over last five years

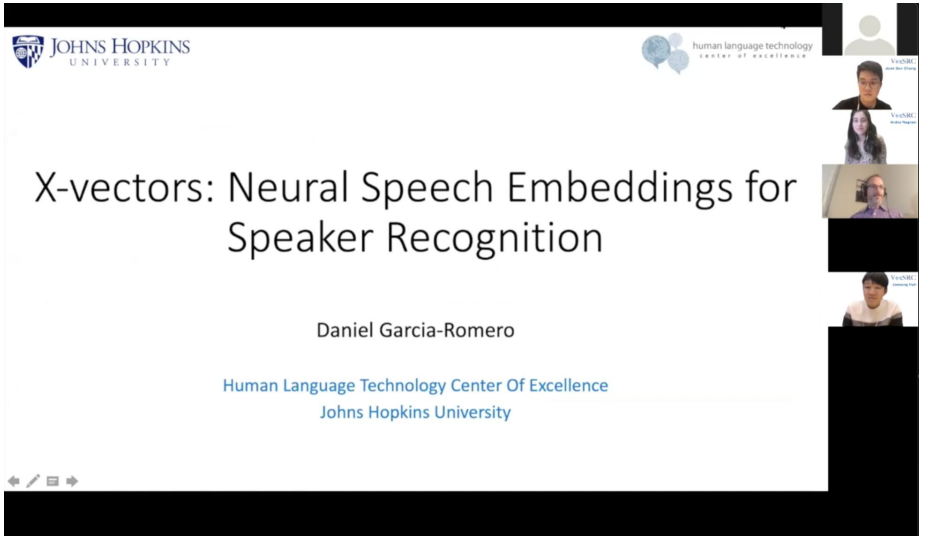
VoxSRC 2019 : the beginning

- In-person workshop in Graz, Austria
- Started with two speaker verification tracks (closed and open)
- Complemented existing evaluations (NIST, SITW)
  - Open access to train/test data
  - Online leaderboards
  - Focus on improving discrimination
    - Calibration-insensitive metrics
      - EER and minDCF
  - Positive/Negative pairs of short duration

The first VoxSRC workshop (2019)

VoxSRC 2020 : workshop during pandamic

- First online workshop due to COVID-19
- Introduced two new tracks
  - Self-supervised speaker verification
  - Speaker diarization
- Introduced new “movie” domain on speaker verification



< Daniel giving the keynote speech>

To make challenging verification test sets….

- **Year 2020 : Speaker verification at the movies**

VoxCeleb

VoxMovies

VoxSRC 2021

- Maintain four tracks
- Another online workshop
- A special focus on multi-lingual speaker verification

To make challenging verification test sets….

- **Year 2021 : Multilingual focus**

English - Spanish

English - French

VoxSRC 2022 : back to normal

- Hybrid workshop with Interspeech 2022
- Remove self-supervised speaker verification track
- Introduce a new track : semi-supervised domain adaptation
  - Large labeled data in source domain (VoxCeleb)
  - Target domain (CN-Celeb)
    - Large unlabeled data
    - Small sample with labels

VoxSRC 2022 workshop

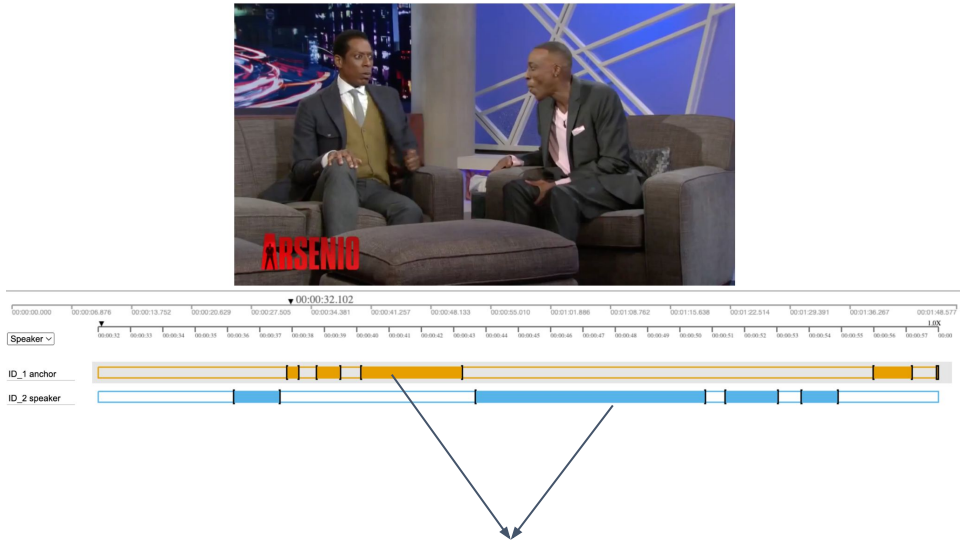
To make challenging verification test sets….

- **Year 2022 : New hard positive pairs and hard negative pairs**

Leonardo
in age 24

Leonardo
in age 47

< Hard positive Pairs >



< Hard negative pairs >

VoxSRC 2023

- Hybrid workshop with Interspeech 2023
- Same tasks and tracks as 2022
- We updated the test paris by…

To make challenging verification test sets….

- **Year 2023 : We include everything**

VoxMovies

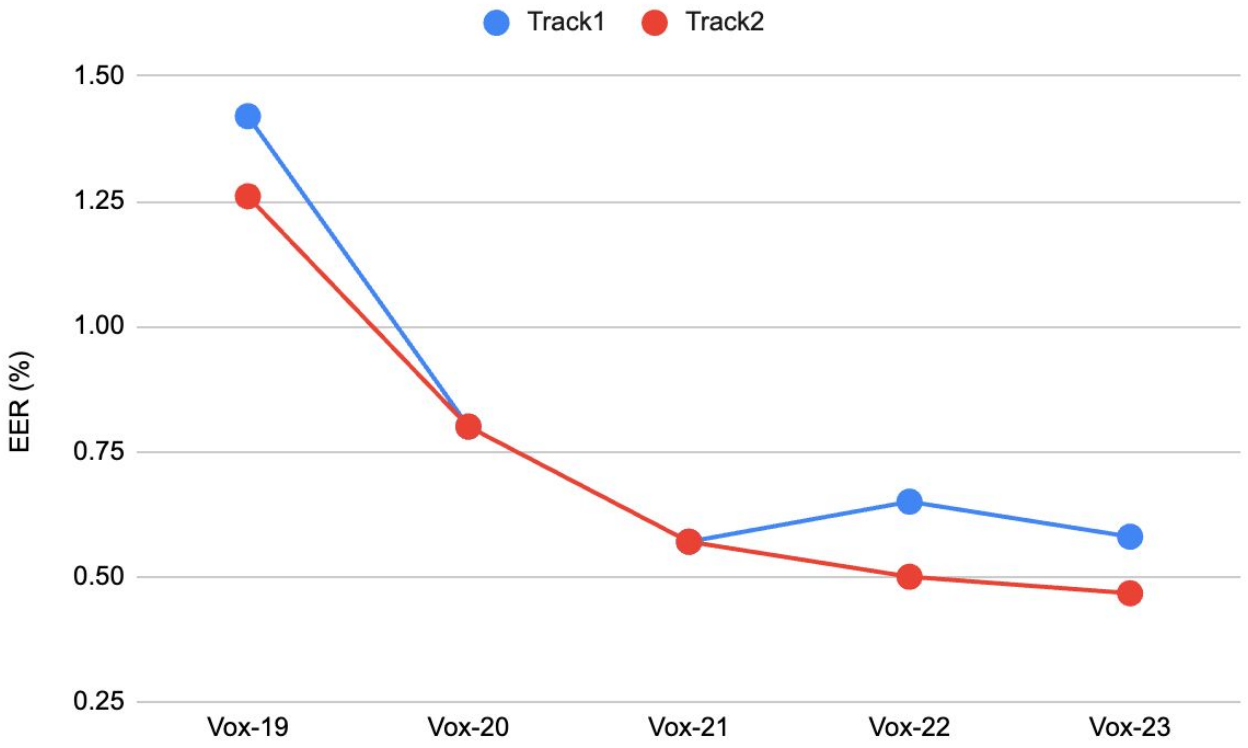
Multilingual focus

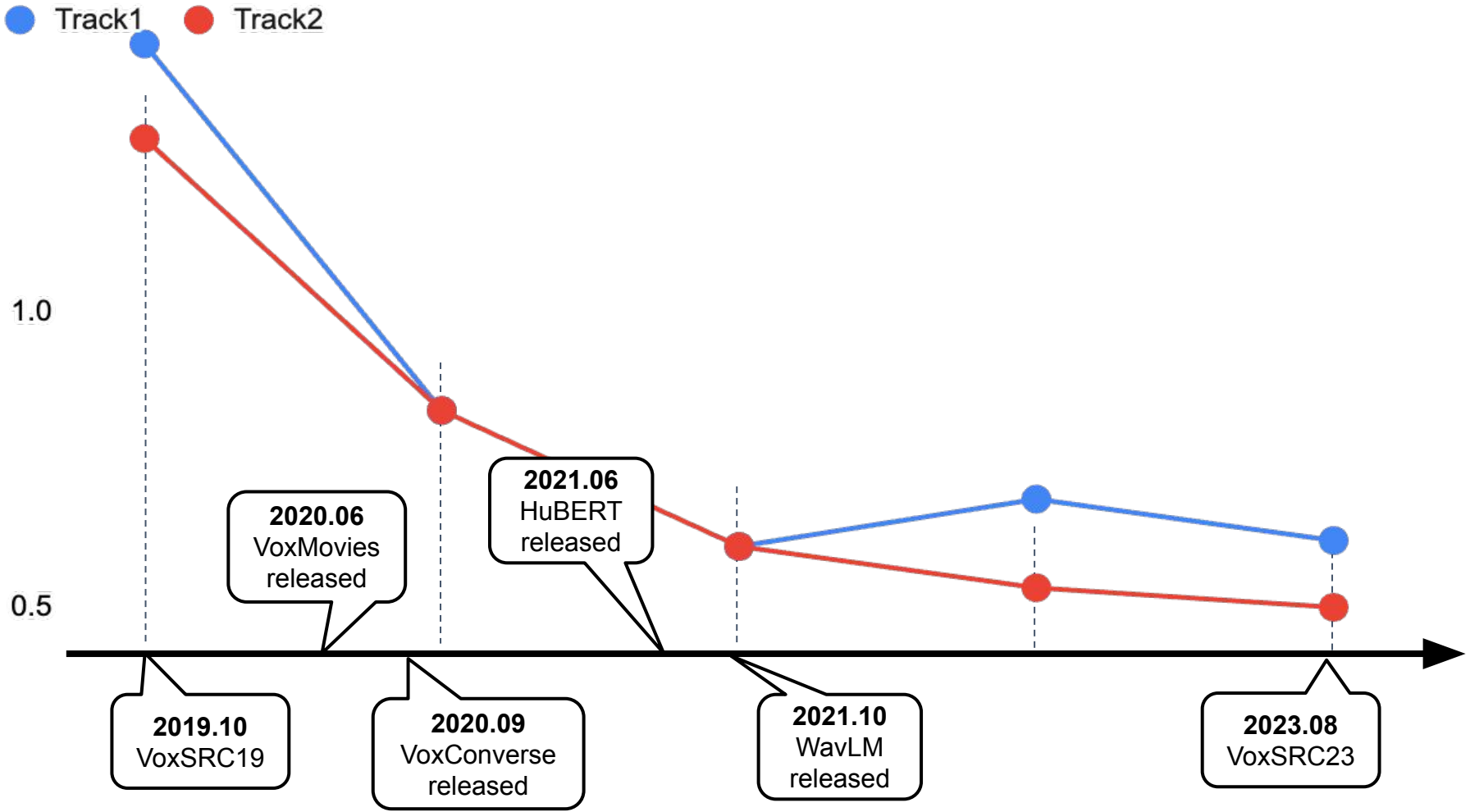
Positive pairs from different age

Negative pairs from same environment

Progress over the five years

- **Top winner's performance on VoxSRC 2019 test set.**




Track1
Track2
1.0
0.5
2019.10
VoxSRC19
2020.06
VoxMovies
released
2020.09
VoxConverse
released
2021.06
HuBERT
released
2021.10
WavLM
released
2023.08
VoxSRC23

Blueprint of a strong verification system

- A successful recipe for and individual system in VoxSRC challenges
  - Data augmentation
    - Additive noise/music, reverb simulation, speed/tempo perturbation
  - Backbone: modified ResNet-{34,...,518}
  - Loss: variant of Angular Margin Softmax
  - Multi-stage training:
    - Short-duration pre-training (multi-class)
    - Large-margin longer-duration fine-tuning (multi-class)
    - Side-information calibration (binary cross-entropy loss)
  - Cosine scoring + adaptive score normalization

Open vs Closed

- **Last 2 years have validated the Deep Learning Mantra:**
  - **MORE DATA + MORE PARAMETERS -> Good things happen!**

Are we stuck?

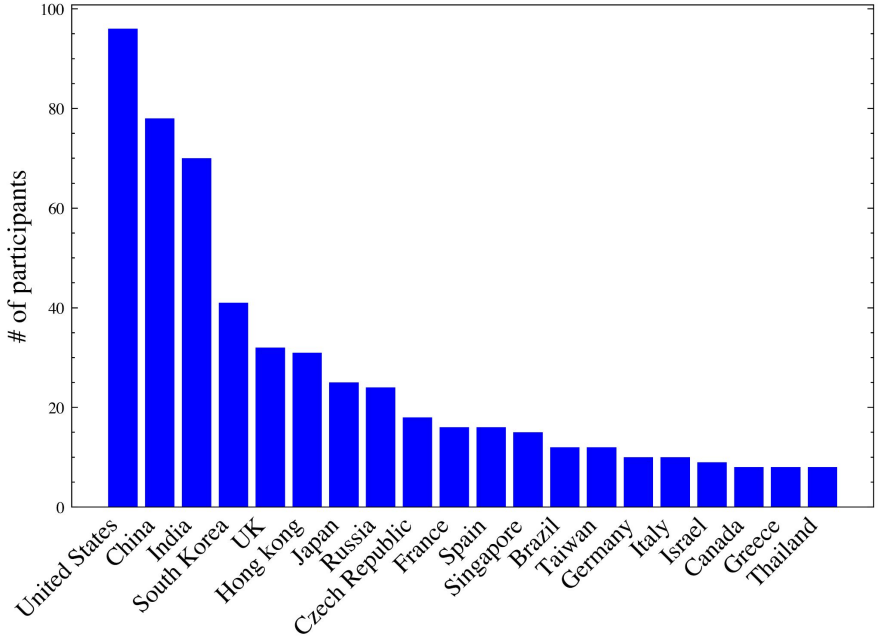
- The ingredients in the recipe were already there in 2019 systems
  - Refined/improved over the last 5 years
- We have found alternatives that match/get close:
  - Pre-trained SSL front-end + supervised layers
  - Different backbones: TDNNs, Conformers
- Do SSL front-ends and Transformer/Conformer deliver gains like in other fields?
  - Good discussion for the coffee break

Top Organisations that have won the challenges

Rank	Organisation	# of times that made to top-3
1	• Duke Kunshan University	7
2	• Microsoft Corporation • ID R&D Lab	4
3	• Brno University of Technology • Johns Hopkins University • Ghent University • Kriston AI Lab	3

Participant trends

Map of where virtual workshop participants come from in VoxSRC 2022



Top 20 countries attended in VoxSRC workshops virtually

Excellent keynote speakers

Mitchell McLaren (2019)

Daniel Garcia-Romero (2020)

Shinji Watanabe (2020)

Andreas Stolcke (2021)

Junichi Yamagishi (2022)

Wei-Ning Hsu (2023)

Workshop Programme

VoxSRC

17:30	Introduction: “VoxCeleb, VoxConverse & VoxSRC”
17:40	Reflecting our last five years
18:00	**Keynote speech : Wei-Ning Hsu “Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations”**
18:45	Coffee break
19:00	Announcement of Winners (Track 1,2 and 3)
19:05	Invited Talks from Track 1 and 2
19:30	Announcement of Winners (Track 3)
19:32	Invited Talks from Track 3
19:50	Announcement of Winners (Track 4)
19:55	Invited Talks from Track 4
20:20	Wrap up discussion and conclusion

Keynote:
Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations

Wei-Ning Hsu

Facebook AI Research (FAIR)

VoxSRC

Q & A

Workshop Programme

VoxSRC

17:30	Introduction: “VoxCeleb, VoxConverse & VoxSRC”
17:40	Reflecting our last five years
18:00	Keynote speech : Wei-Ning Hsu “Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations”
18:45	**Coffee break**
19:00	Announcement of Winners (Track 1,2 and 3)
19:05	Invited Talks from Track 1 and 2
19:30	Announcement of Winners (Track 3)
19:32	Invited Talks from Track 3
19:50	Announcement of Winners (Track 4)
19:55	Invited Talks from Track 4
20:20	Wrap up discussion and conclusion

VoxSRC

Coffee Break

Restarting at 19:00 Dublin time

Workshop Programme

VoxSRC

17:30	Introduction: “VoxCeleb, VoxConverse & VoxSRC”
17:40	Reflecting our last five years
18:00	Keynote speech : Wei-Ning Hsu “Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations”
18:45	Coffee break
19:00	**Announcement of Winners (Track 1,2 and 3)**
19:05	Invited Talks from Track 1 and 2
19:30	Announcement of Winners (Track 3)
19:32	Invited Talks from Track 3
19:50	Announcement of Winners (Track 4)
19:55	Invited Talks from Track 4
20:20	Wrap up discussion and conclusion

Track 1,2: Speaker Verification

- Determine whether two samples of speech are from the same person.
- Track 1 : Closed track (only use VoxCeleb2 dev as train set)
- Track 2 : Open track (can use any data for training)
- Primary metric - minDCF

VoxSRC Winners - Track1 Speaker verification [closed training set]

VoxSRC

- 1st Team UniVox (Unisound AI Technology, Ltd.)

  Yu Zheng , Yajun Zhang , Chuanying Niu , Yibin Zhan , Yanhua Long, Dongxing Xu

- 2nd Team XG0721 (BiliBili)

  Xingui Zeng, Zhuo Yang , Shiyi Wan, Wei Deng, XiangCao

- 3rd Team teamxxx (ChinaTelecom)

  Mengjie Du, Xiang Fang, Jie Li

60 participants, 37 teams submitted, 150 total submissions

VoxSRC Winners - Track2 Speaker verification [open training set]

VoxSRC

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

  Nikita Torgashov, Rostislav Makarov, Ivan Yakovlev, Pavel Malov, Andrei Balykin, Anton Okhotnikov

- 2nd Team UniVox (Unisound AI Technology, Ltd.)

  Yu Zheng , Yajun Zhang , Chuanying Niu , Yibin Zhan , Yanhua Long, Dongxing Xu

- 3rd Team Intema (Intema SARL)

  Ali Aliyev

46 participants, 25 teams submitted, 125 total submissions

Workshop Programme

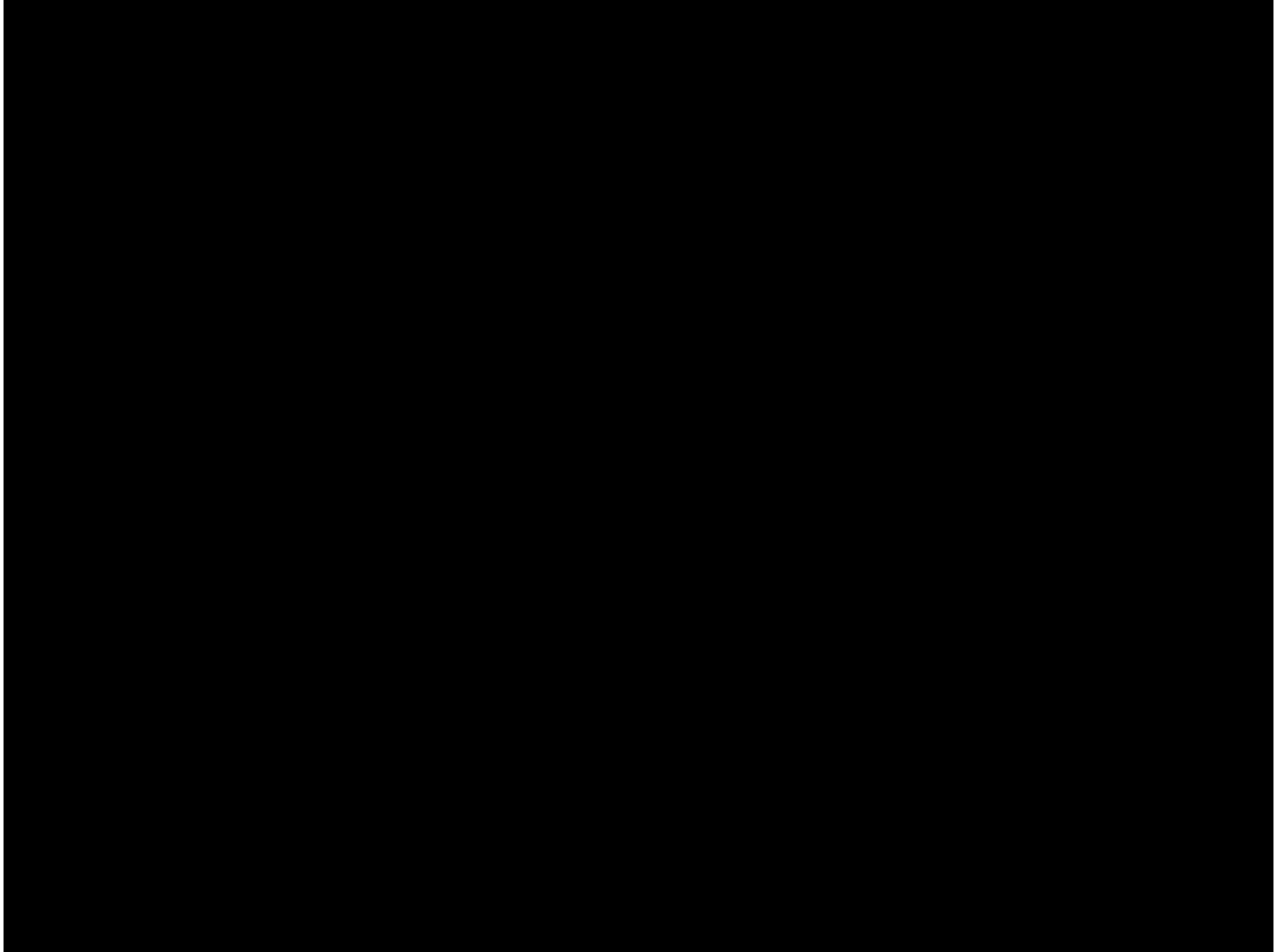
VoxSRC

17:30	Introduction: “VoxCeleb, VoxConverse & VoxSRC”
17:40	Reflecting our last five years
18:00	Keynote speech : Wei-Ning Hsu “Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations”
18:45	Coffee break
19:00	Announcement of Winners (Track 1,2 and 3)
19:05	**Invited Talks from Track 1 and 2**
19:30	Announcement of Winners (Track 3)
19:32	Invited Talks from Track 3
19:50	Announcement of Winners (Track 4)
19:55	Invited Talks from Track 4
20:20	Wrap up discussion and conclusion

Talks by the winners

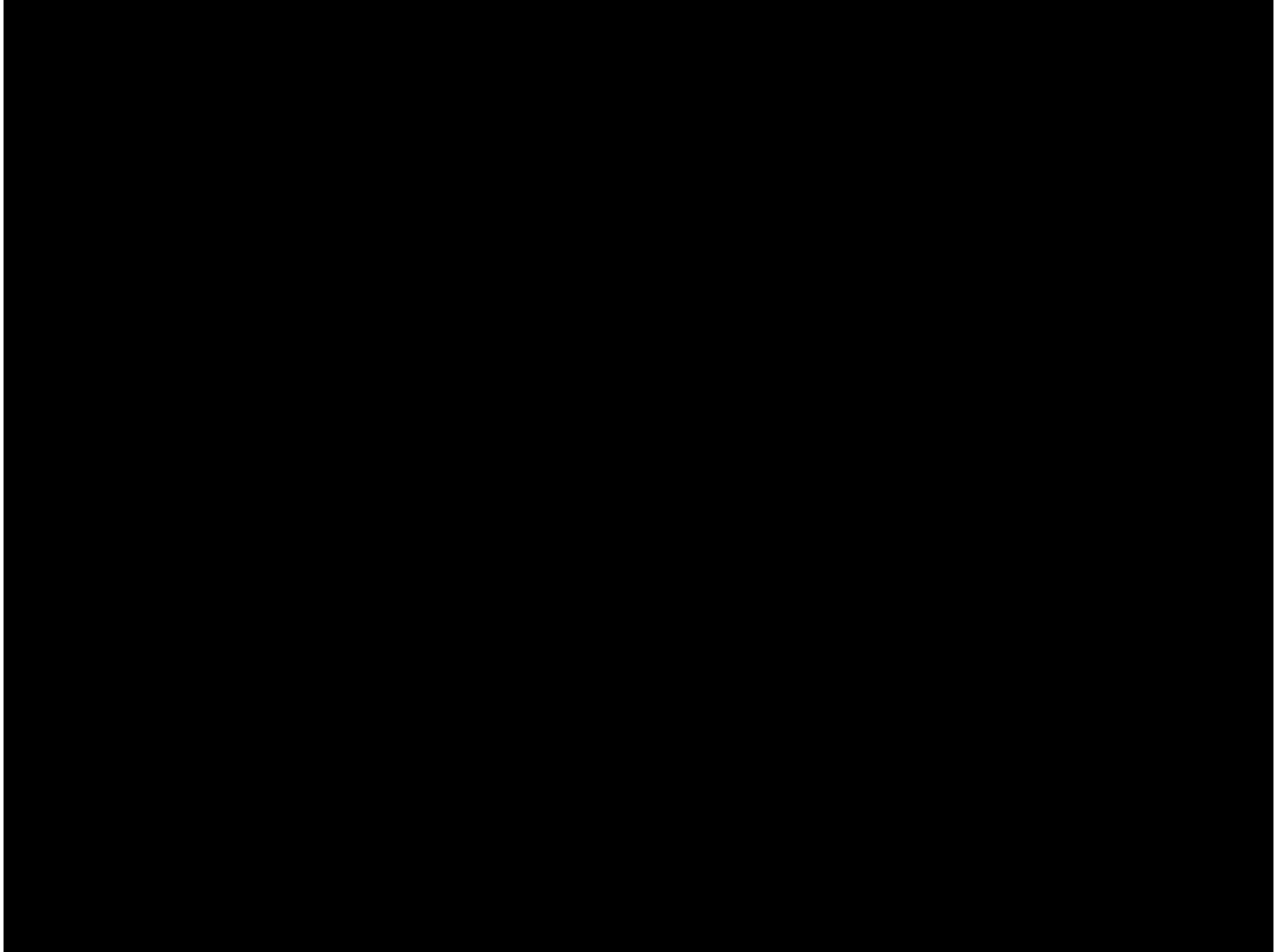
1. Team UniVox (Unisound AI Tech Co. Ltd.) - tracks 1,2
2. Team XG0721 (BiliBili) - track 1
3. Team carpincho (ID R&D lab) - track 2

Univox



VoxSRC

BiliBili



VoxSRC

VoxSRC

Team carpincho (ID R&D lab)

In-person talk

Workshop Programme

VoxSRC

17:30	Introduction: “VoxCeleb, VoxConverse & VoxSRC”
17:40	Reflecting our last five years
18:00	Keynote speech : Wei-Ning Hsu “Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations”
18:45	Coffee break
19:00	Announcement of Winners (Track 1,2 and 3)
19:05	Invited Talks from Track 1 and 2
19:30	**Announcement of Winners (Track 3)**
19:32	Invited Talks from Track 3
19:50	Announcement of Winners (Track 4)
19:55	Invited Talks from Track 4
20:20	Wrap up discussion and conclusion

Track 3: Semi-supervised domain adaptation

VoxSRC

- 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 DKU-MSXF (Duke Kunshan University,)

  Ze Li ,Yuke Lin ,Xiaoyi Qin, Ning Jiang, Guoqing Zhao, Ming Li

- 2nd Team xx205 (Shanghai Jiao Tong University)

  Xu Xiang

- 3rd Team wespeaker_baseline (Shanghai Jiao Tong University, Tencent Ethereal Audio Lab)

  Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, Yanmin Qian

16 participants, 4 teams submitted, 11 total submissions

Workshop Programme

VoxSRC

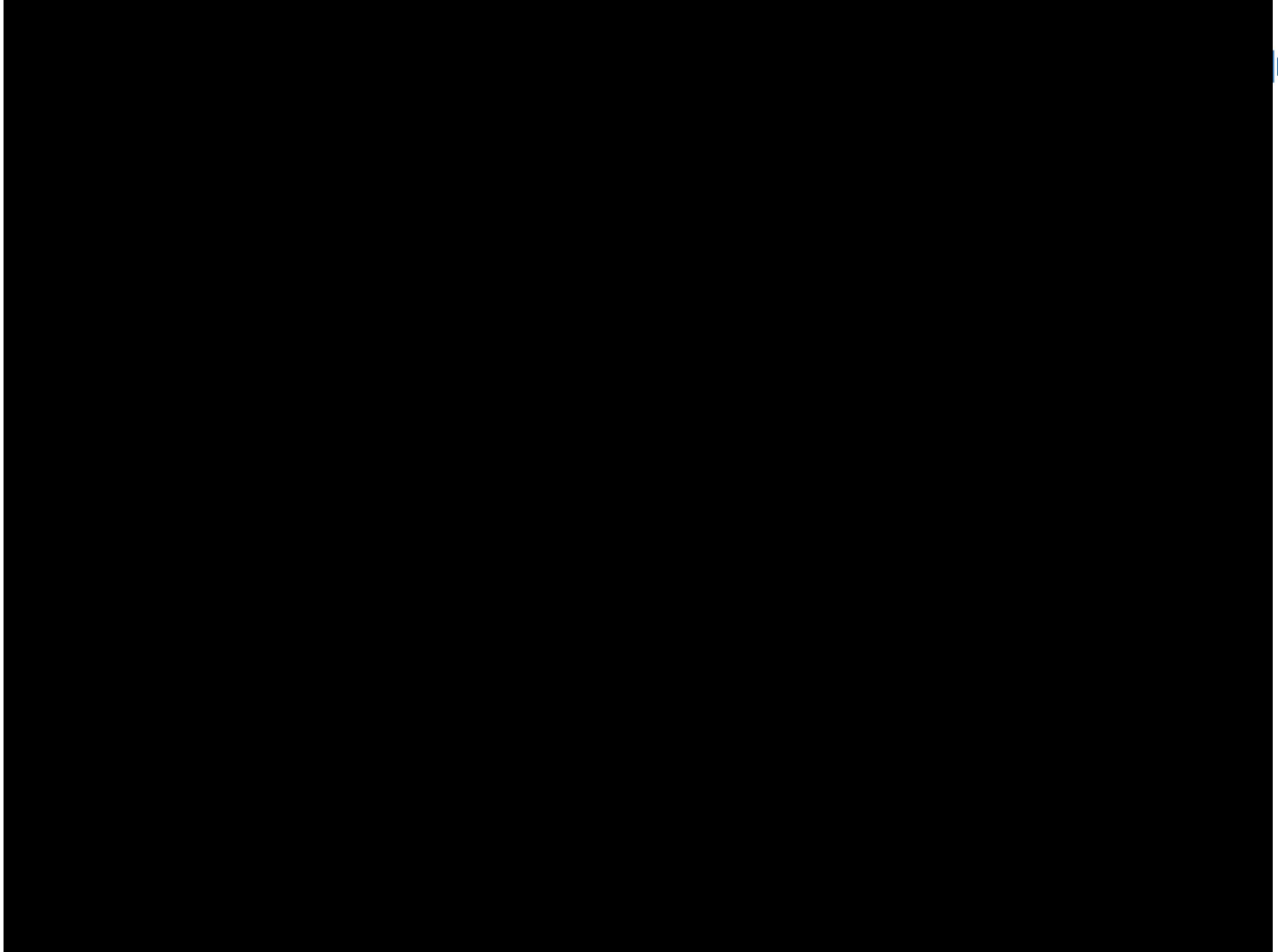
17:30	Introduction: “VoxCeleb, VoxConverse & VoxSRC”
17:40	Reflecting our last five years
18:00	Keynote speech : Wei-Ning Hsu “Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations”
18:45	Coffee break
19:00	Announcement of Winners (Track 1,2 and 3)
19:05	Invited Talks from Track 1 and 2
19:30	Announcement of Winners (Track 3)
19:32	**Invited Talks from Track 3**
19:50	Announcement of Winners (Track 4)
19:55	Invited Talks from Track 4
20:20	Wrap up discussion and conclusion

VoxSRC

Talks by the winners

1. Team DKU-MSXF (Duke Kunshan University)

DKU-MSXF



xSRC

Workshop Programme

VoxSRC

17:30	Introduction: “VoxCeleb, VoxConverse & VoxSRC”
17:40	Reflecting our last five years
18:00	Keynote speech : Wei-Ning Hsu “Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations”
18:45	Coffee break
19:00	Announcement of Winners (Track 1,2 and 3)
19:05	Invited Talks from Track 1 and 2
19:30	Announcement of Winners (Track 3)
19:32	Invited Talks from Track 3
19:50	**Announcement of Winners (Track 4)**
19:55	Invited Talks from Track 4
20:20	Wrap up discussion and conclusion

VoxSRC

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 DKU-MSXF (Duke Kunshan Univ., Mashang Consumer Finance Co. Ltd)

  Ming Cheng, Weiqing Wang, Xiaoyi Qin, Yuke Lin, Ning Jiang, Guoqing Zhao, Ming Li

- 2nd Team KrispAI (Krisp.AI)

  Davit Karamyan, Grigor Kirakosyan

- 3rd Team pyannote (IRIT, Universite de Toulouse)

  Severin Baroudi, Herve Bredin, Alexis Plaquet, Thomas Pellegrini

42 participants, 24 teams submitted, 92 total submissions

Workshop Programme

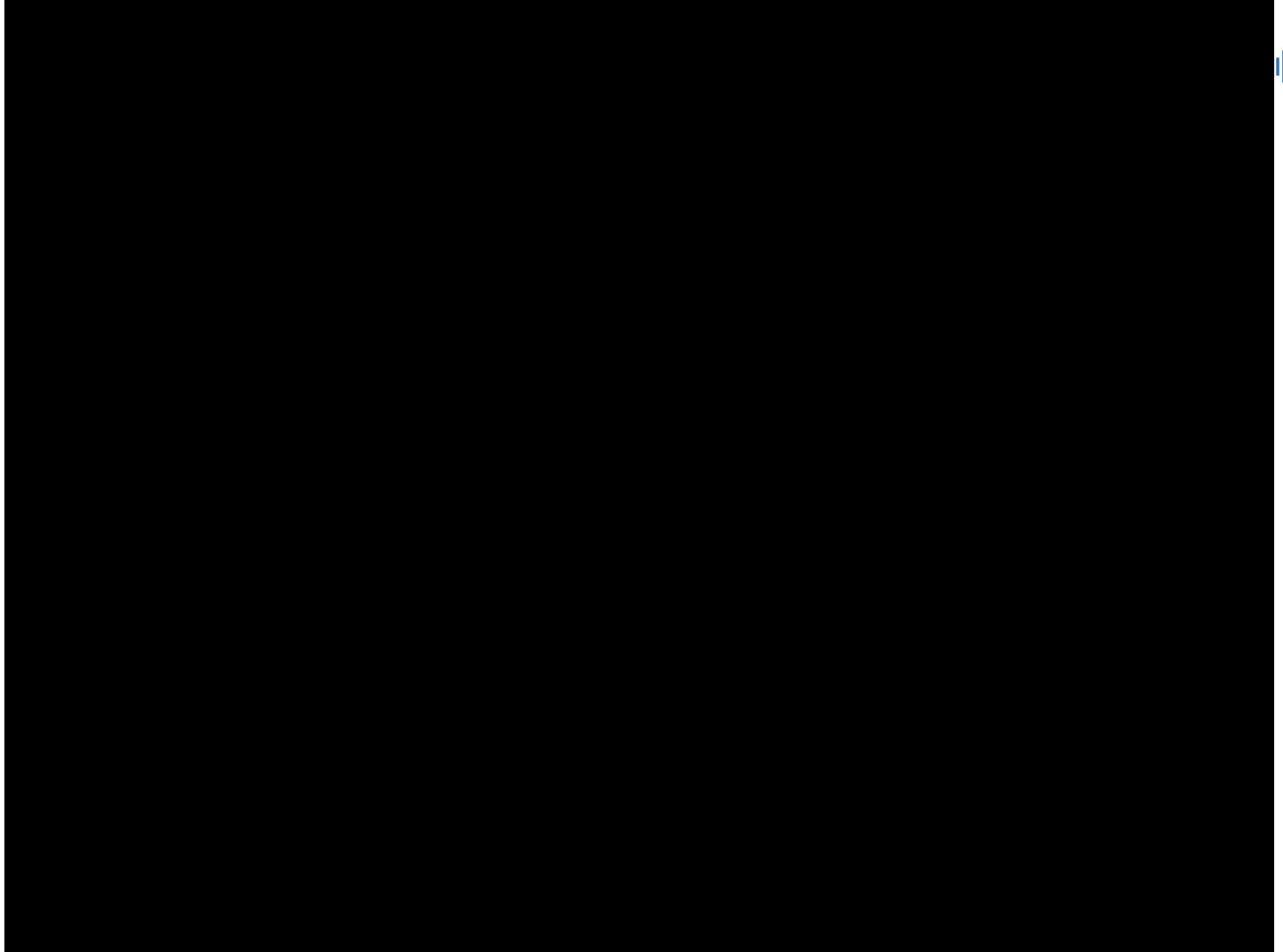
VoxSRC

17:30	Introduction: “VoxCeleb, VoxConverse & VoxSRC”
17:40	Reflecting our last five years
18:00	Keynote speech : Wei-Ning Hsu “Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations”
18:45	Coffee break
19:00	Announcement of Winners (Track 1,2 and 3)
19:05	Invited Talks from Track 1 and 2
19:30	Announcement of Winners (Track 3)
19:32	Invited Talks from Track 3
19:50	Announcement of Winners (Track 4)
19:55	**Invited Talks from Track 4**
20:20	Wrap up discussion and conclusion

Talks by the winners

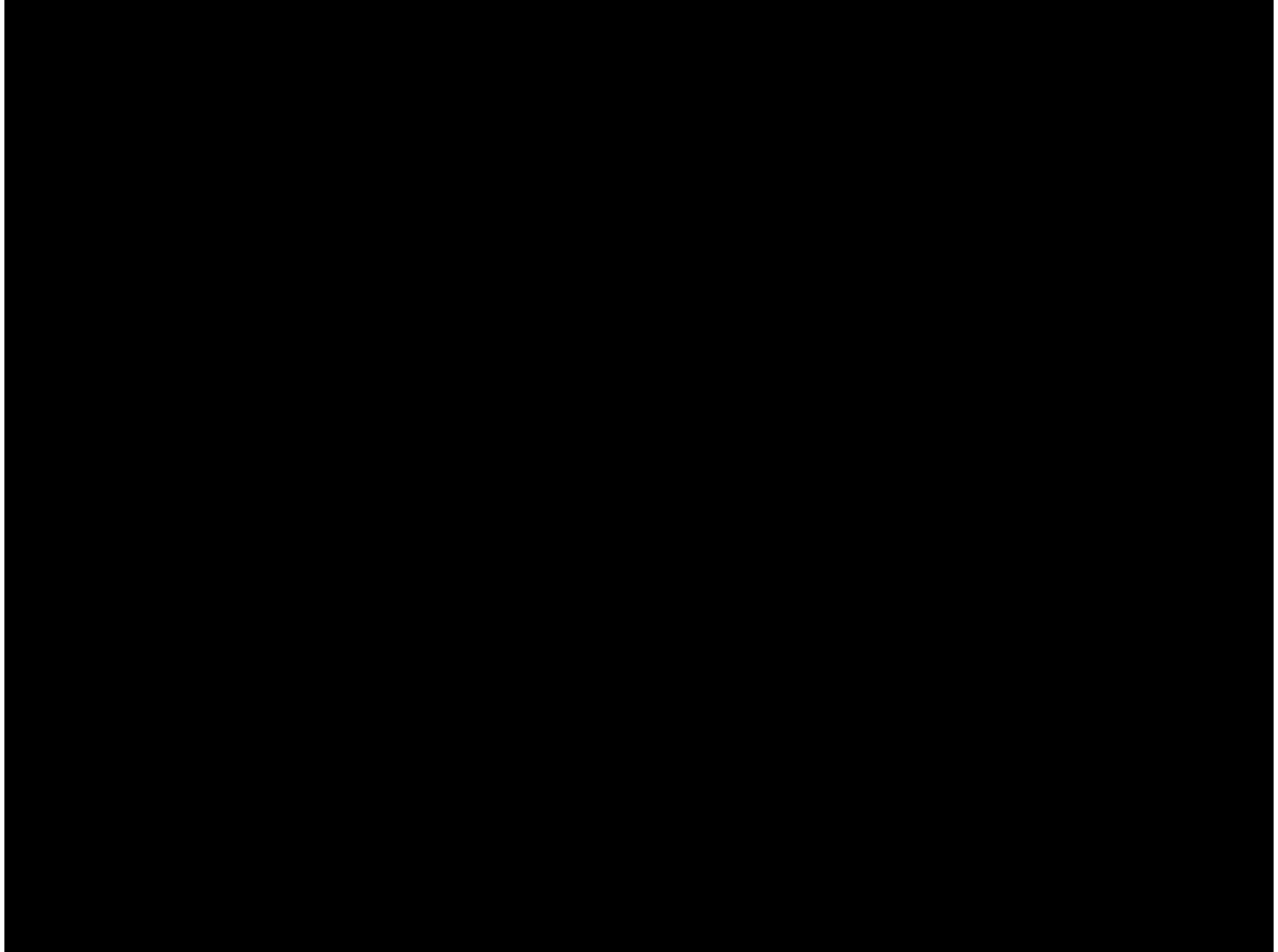
1. Team DKU-MSXF (Duke Kunshan University)
2. Team KrispAI (KrispAI)
3. Team pyannote (IRIT, Universite de Toulouse, CNRS)

Team
DKU-MSXF



xSRC

Team
KrispAI



VoxSRC

Team pyannote

VoxSRC

In-person talk

Workshop Programme

VoxSRC

17:30	Introduction: “VoxCeleb, VoxConverse & VoxSRC”
17:40	Reflecting our last five years
18:00	Keynote speech : Wei-Ning Hsu “Scalable Controllable Speech Generation via Explicitly and Implicitly Disentangled Speech Representations”
18:45	Coffee break
19:00	Announcement of Winners (Track 1,2 and 3)
19:05	Invited Talks from Track 1 and 2
19:30	Announcement of Winners (Track 3)
19:32	Invited Talks from Track 3
19:50	Announcement of Winners (Track 4)
19:55	Invited Talks from Track 4
20:20	**Wrap up discussion and conclusion**

Wrap-up

- Officially, the last VoxSRC workshop is over.
- We are planning to write a journal paper to summarise our last five years.
- Evaluation server will be running, test set will be kept private.

We would like to thank participants over the last five years for their excellent performance and contribution to the community.

Acknowledgements

VoxSRC

- Organisers
  - Joon Son Chung
  - Arsha Nagrani
  - Andrew Brown
  - Jaesung Huh
  - Jee-weon Jung
  - Daniel Garcia-Romero
  - Andrew Zisserman
- Advisors
  - Mitch Mclaren
  - Doug Reynolds
- Sponsors
  - Naver / Line Corporation
  - AWS AI
  - UK EPSRC
- Annotators
  - Rajan and Elancer Ltd.
  - VGG members
- CN-Celeb authors
- Engineers
  - Abhishek Dutta
  - David Pinto
  - Ashish Thandavan

Thank you!

KAIST

Visual AI

VGG
University of Oxford

UNIVERSITY OF
OXFORD