Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion

Zhao Yi[1], Wen-Chin Huang[2], Xiaohai Tian[3], Junichi Yamagishi[1],

Rohan Kumar Das[3], Tomi Kinnunen[4], Zhenhua Ling[5], Tomoki Toda[2]

[1]National Institute of Informatics, Japan [2]Nagoya University, Japan
[3]National University of Singapore, Singapore [4]University of Eastern Finland, Finland
[5]University of Science and Technology of China, P.R.China

vcc2020@vc-challenge.org

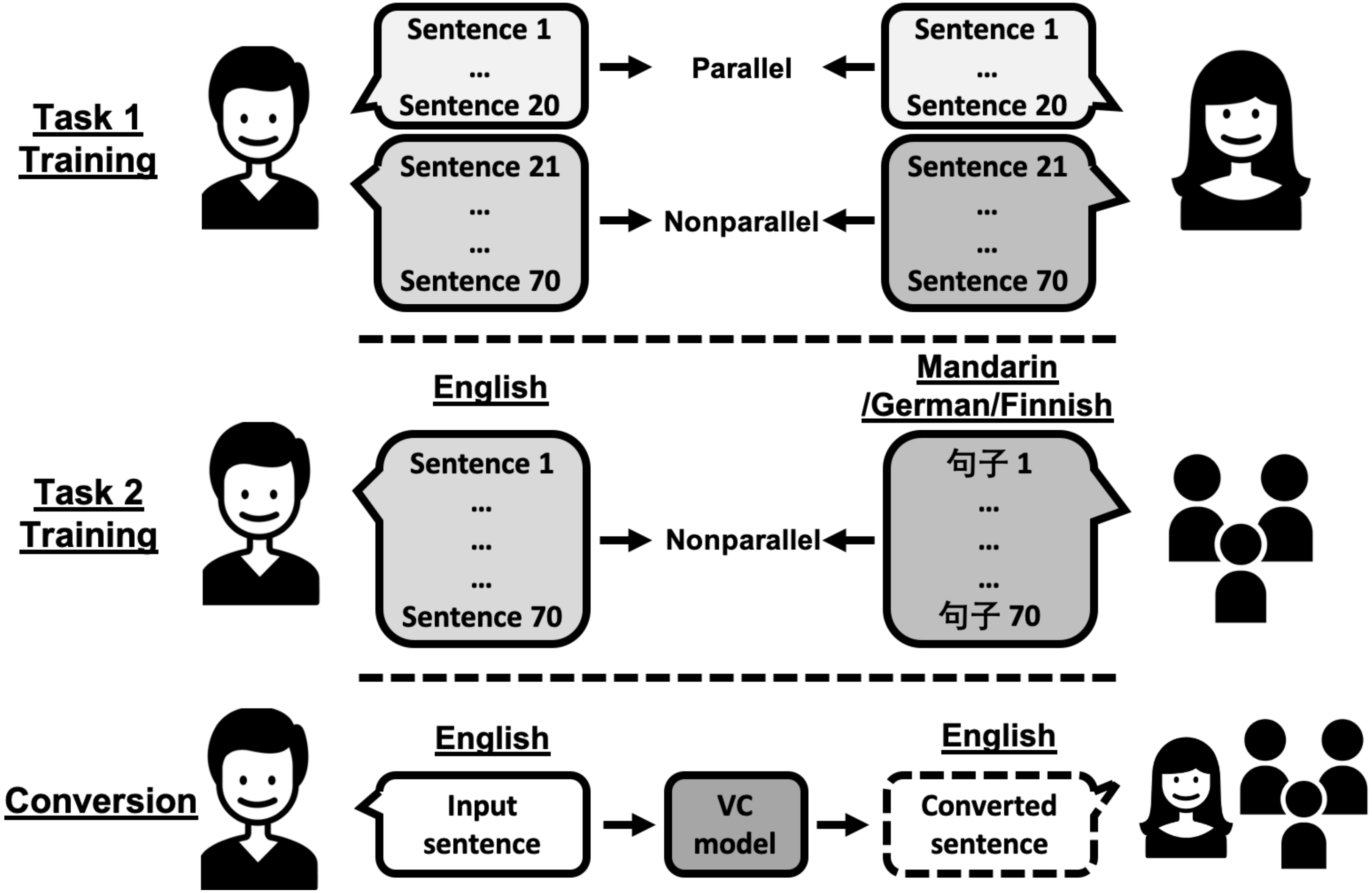
Outline

- Tasks, databases, and timeline for Voice Conversion Challenge 2020
- Participants and submitted systems
- Subjective evaluations and analysis of VCC 2020 results
- Conclusions

Outline

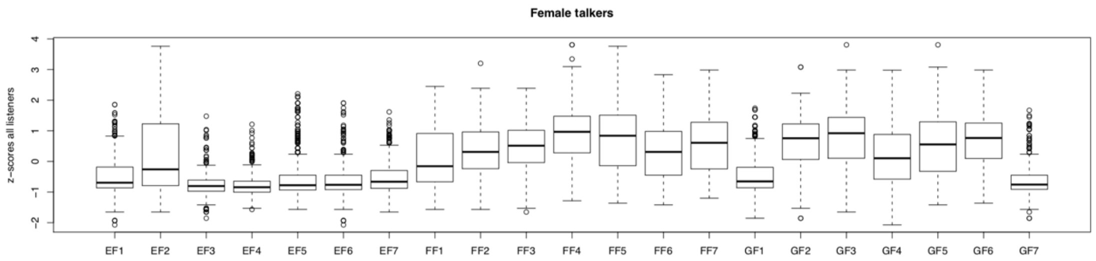
- Tasks, databases, and timeline for Voice Conversion Challenge 2020
- Participants and submitted systems
- Subjective evaluations and analysis of VCC 2020 results
- Conclusions

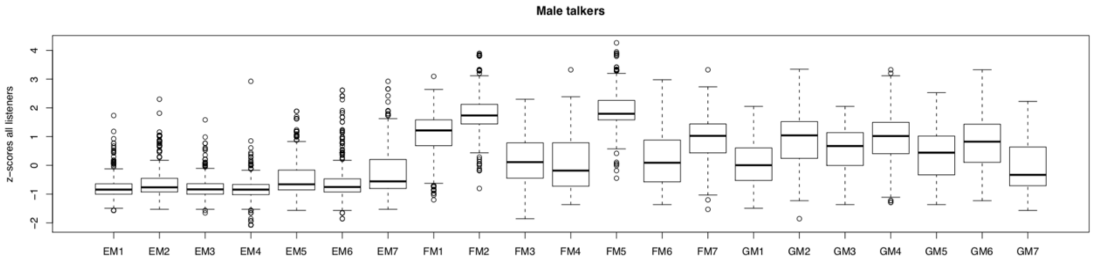
Tasks


Task 1 Training
Sentence 1
...
Sentence 20
Parallel
Sentence 1
...
Sentence 20
Sentence 21
...
...
Sentence 70
Nonparallel
Sentence 21
...
...
Sentence 70
Task 2 Training
English
Mandarin /German/Finnish
Sentence 1
...
...
...
Sentence 70
Nonparallel
句子 1
...
...
...
句子 70
Conversion
English
Input sentence
VC model
English
Converted sentence

Database

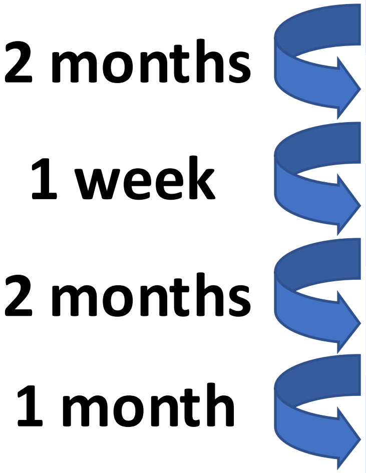
- EMIME: Effective Multilingual Interaction in Mobile Environments
  - Languages: **German**/**English, Finnish**/**English, Mandarin**/**English**
- 4 English source speakers (2M+2F)
- Task 1: 4 English target speakers (2M+2F)
  - Criterion: as perceptually discriminative as possible (chosen manually)
- Task 2: 2 German/Finnish/Mandarin target speakers (1M+1F/language)
  - Criterion: fluency





Timeline

Date	Event
March 9th, 2020	Release of **training** data
May 22nd, 2020	Release of **evaluation** data
May 29th, 2020	Deadline to submit the **converted audio**
July 31st, 2020	**Notification** of the first temporal **results**
Sep. 7th, 2020	Deadline to **submit workshop papers**
Sep. 30th, 2020	Notification of acceptance
Oct 25th-29th 2020	INTERSPEECH 2020
Oct. 30th, 2020	Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020



2 months

1 week

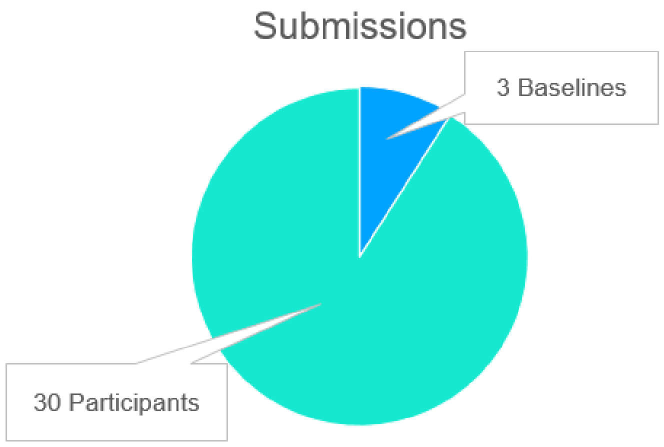
2 months

1 month

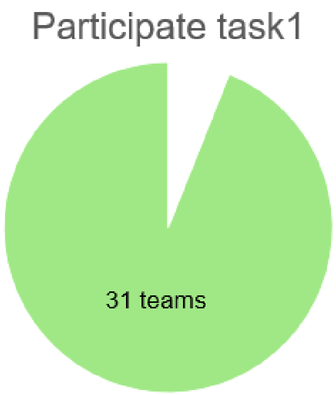
Outline

- Tasks, databases, and timeline for Voice Conversion Challenge 2020
- Participants and submitted systems
- Subjective evaluations and analysis of VCC 2020 results
- Conclusions

Participants and submitted systems



- Total submissions 33
  - 3 baselines
  - 30 participants



Participate task2
29 teams

Participate both
26 teams

- For different tasks
  - 31 teams for task1
  - 29 teams for task2
  - 26 teams for both task2

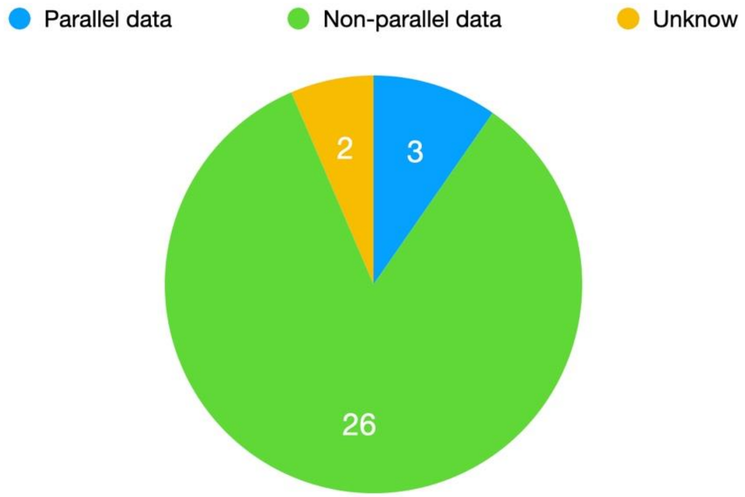
Feature conversion models

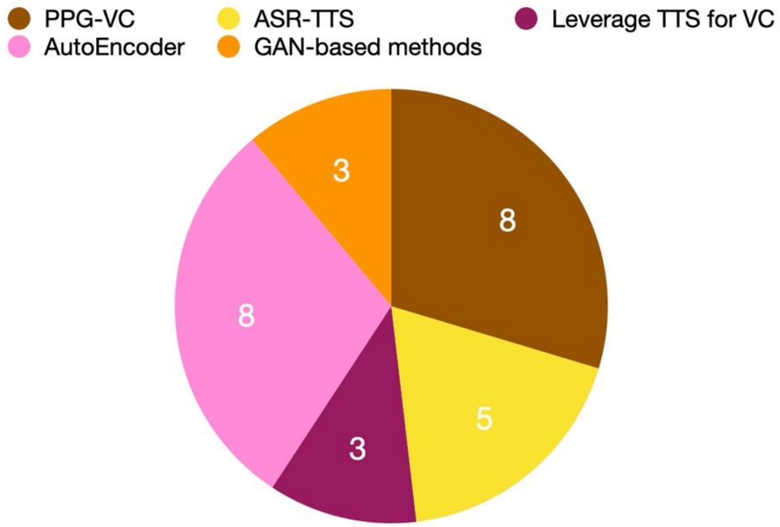
- Summary of feature conversion models used in submitted systems

Category	Feature conversion model
Non-parallel data solutions	PPG-VC
	ASR-TTS
	Leverage TTS for VC
	AutoEncoder VC
	GAN-based VC
Parallel data solution	Tacotron
	VTLN + spectral differential

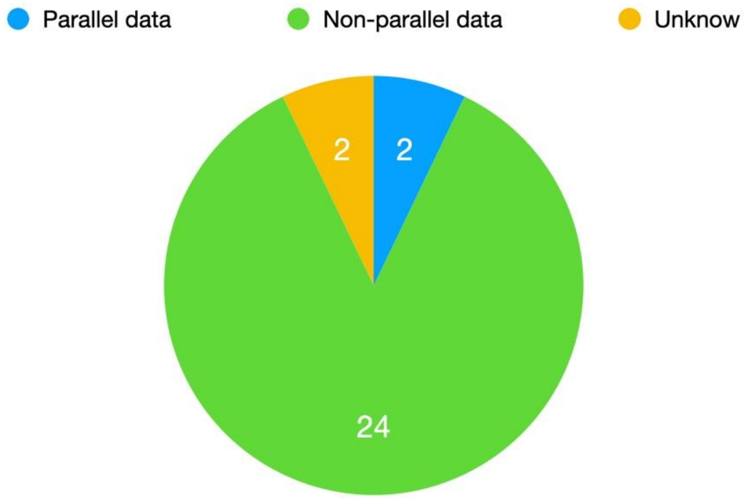
Feature conversion module

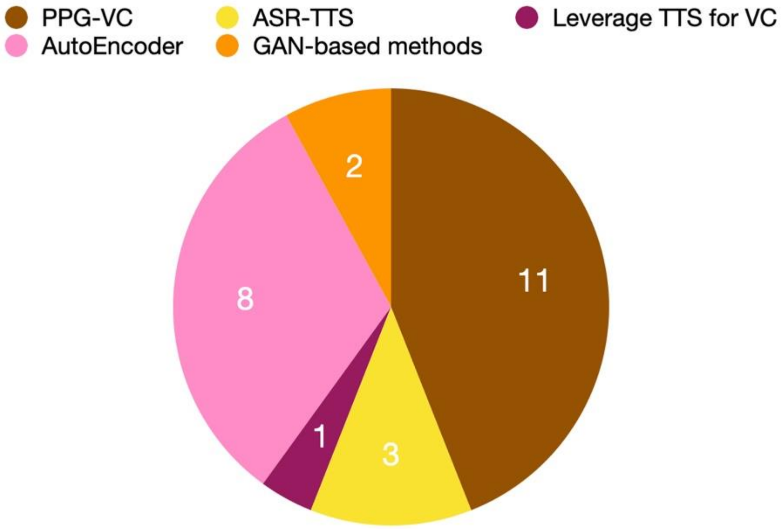
Task 1: Monolingual VC





Task 2: Cross-lingual VC





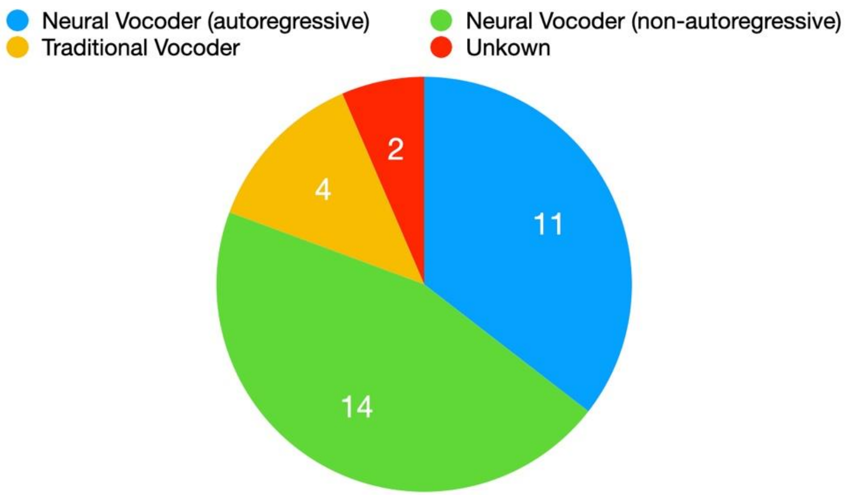
Vocoders

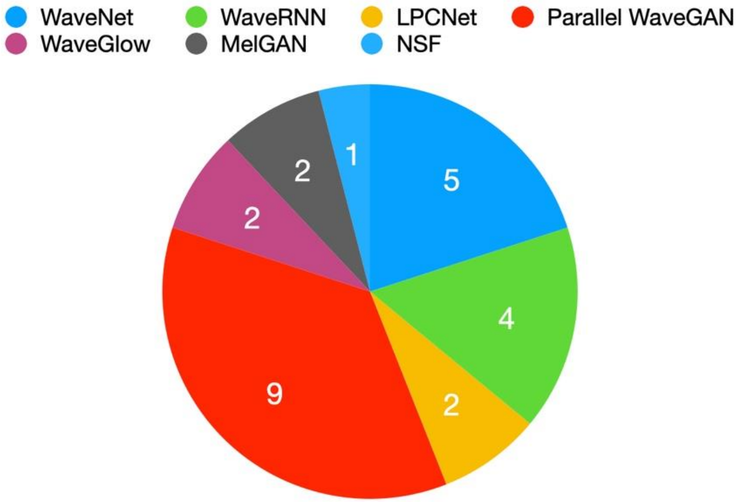
- Summary of vocoders used in submitted systems

Category	Vocoder
Neural Vocoder (Autoregressive)	WaveNet
	WaveRNN
	LPCNet
Neural Vocoder (Non-autoregressive)	Parallel WaveGAN
	WaveGlow
	MelGAN
	NSF
Traditional Vocoder	WORLD
	AHOCoder
	Griffin-Lim

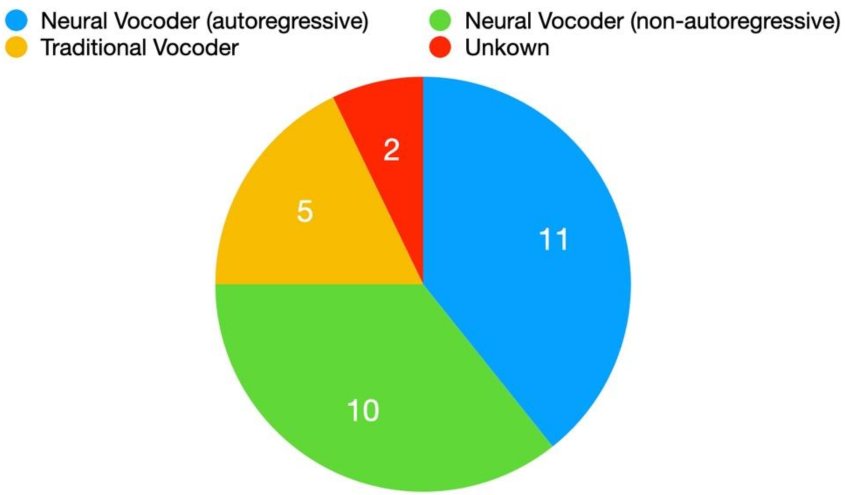
Vocoder

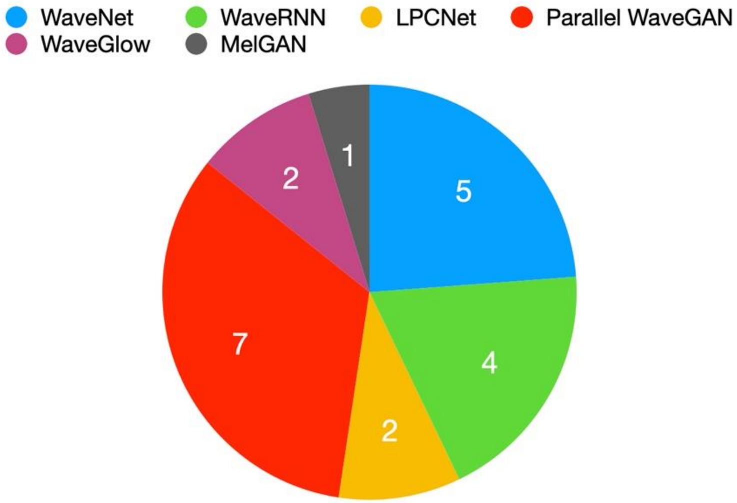
Task 1: Monolingual VC





Task 2: Cross-lingual VC





Outline

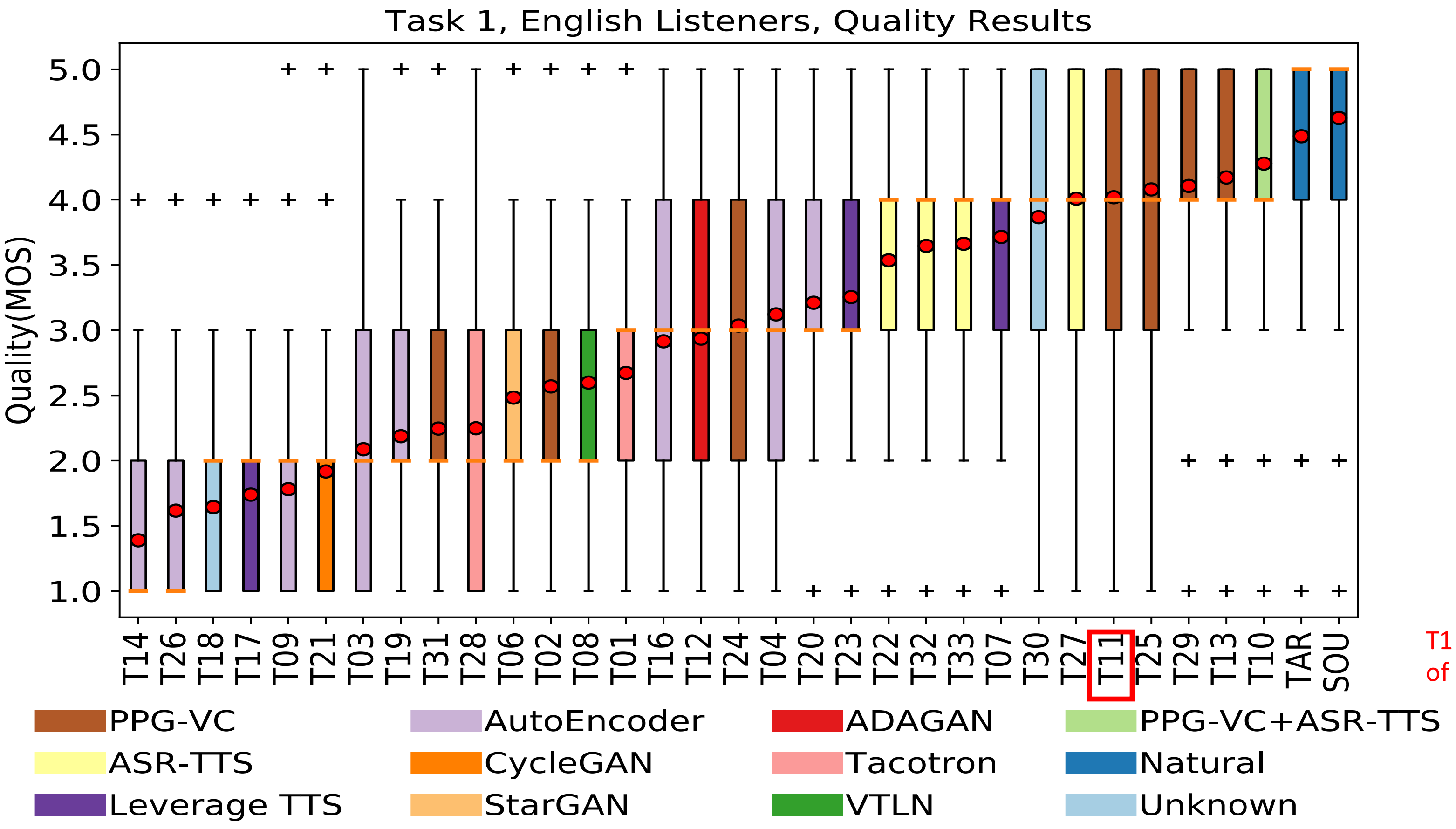
- Tasks, databases, and timeline for Voice Conversion Challenge 2020
- Participants and submitted systems
- Subjective evaluations and analysis of VCC 2020 results
- Conclusions

Design of crowd sourcing test

- Motivations: evaluate naturalness and speaker similarity
- Evaluation methodology
  - Naturalness: five-point scale MOS
  - Similarity: four-point scale score
  - In task2, in addition to reference speech in English, reference speech in either German, Finnish, and Mandarin were also presented to subjects for judging speaker similarity across languages.
- English & Japanese listeners
  - 68 unique valid English listeners (32 female and 33male, and 3 unknown)
  - 206 unique valid Japanese listeners, 96 male and 110 female)

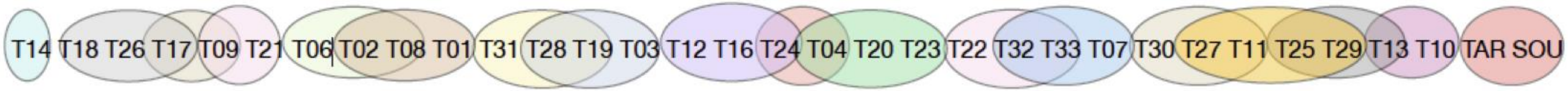
Naturalness results for Task 1

- Bar plot for MOS score:



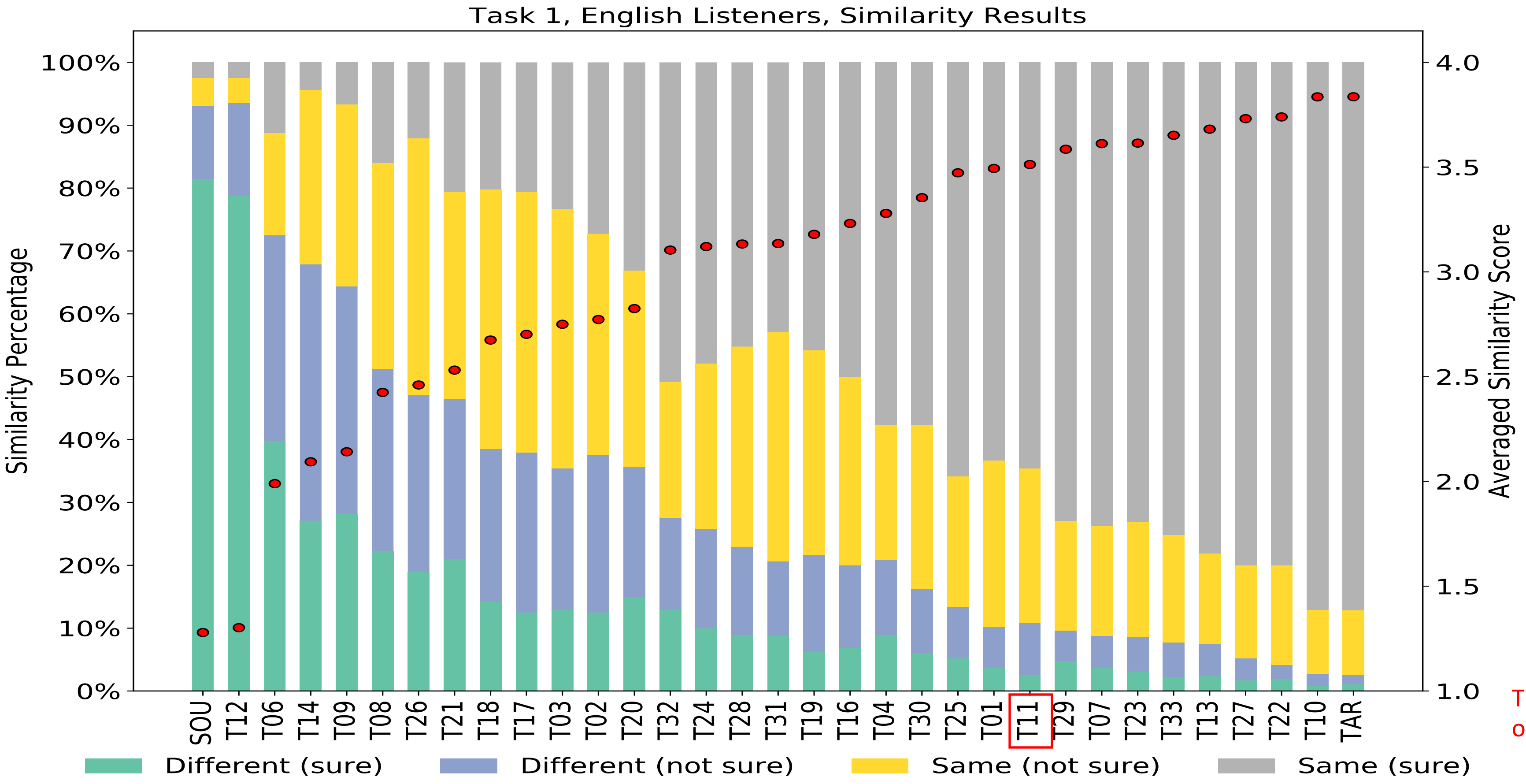
T11 is the top system of last VCC

- Groupings of systems that did not differ significantly from each other in terms of naturalness for Task 1:


T14 T18 T26 T17 T09 T21 T06 T02 T08 T01 T31 T28 T19 T03 T12 T16 T24 T04 T20 T23 T22 T32 T33 T07 T30 T27 T11 T25 T29 T13 T10 TAR SOU

Similarity results for Task 1

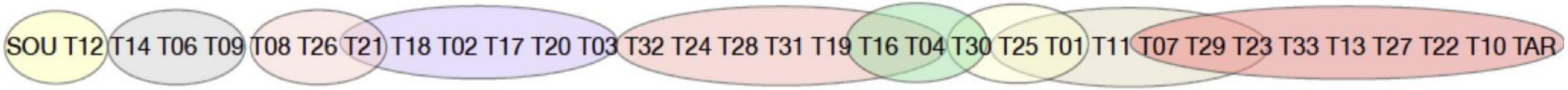
- Plot for similarity score:



T11 is the top system of last VCC

- Groupings of systems that did not differ significantly from each other in terms of similarity for Task 1:

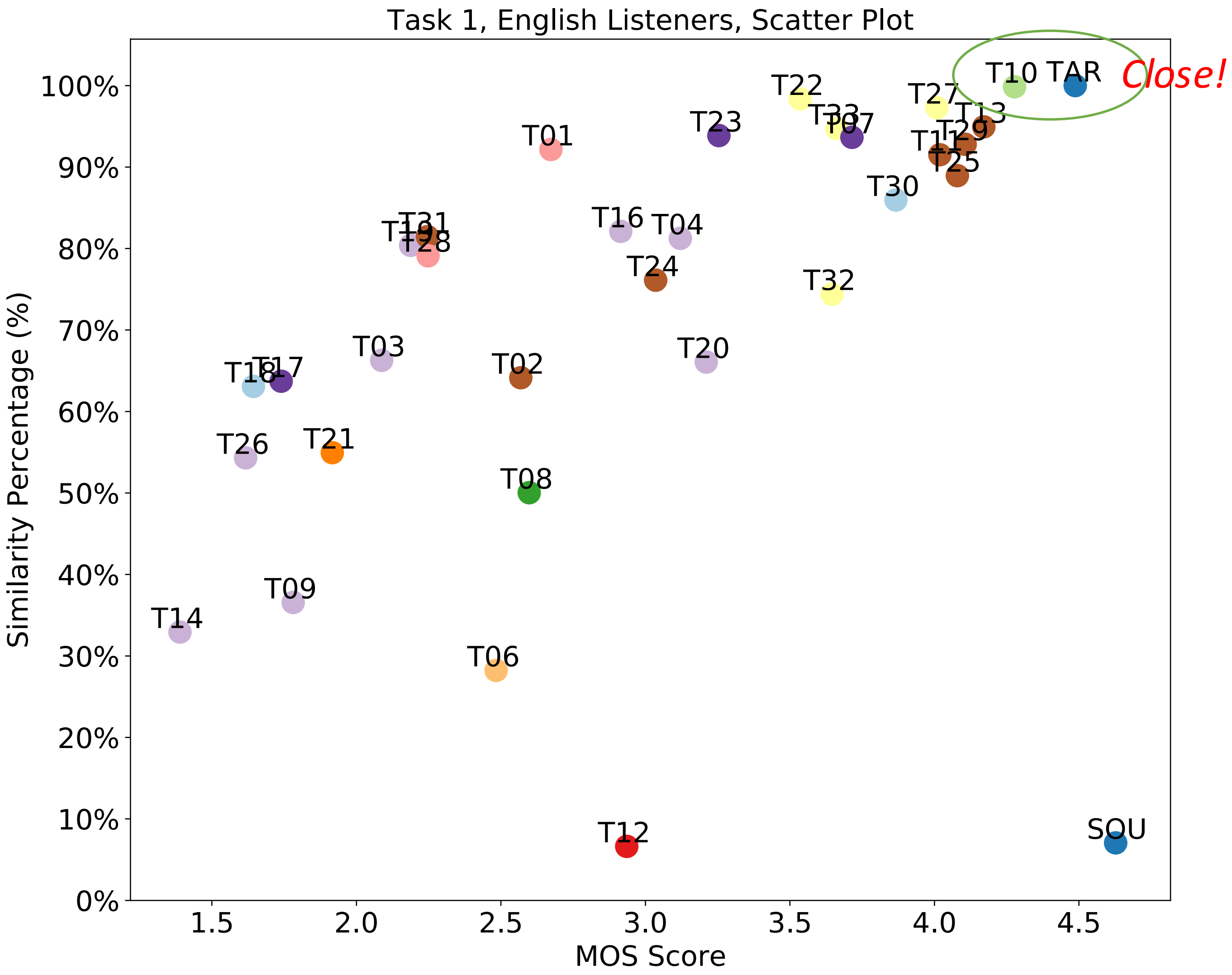
Milestone!



SOU T12 T14 T06 T09 T08 T26 T21 T18 T02 T17 T20 T03 T32 T24 T28 T31 T19 T16 T04 T30 T25 T01 T11 T07 T29 T23 T33 T13 T27 T22 T10 TAR

Scatter plot for Task 1

- Scatter plot between MOS score and speaker similarity percentage:



Close!

Speaker: SEF1 TEM1

Source: 🔊 Target: 🔊

Converted samples of several top teams:

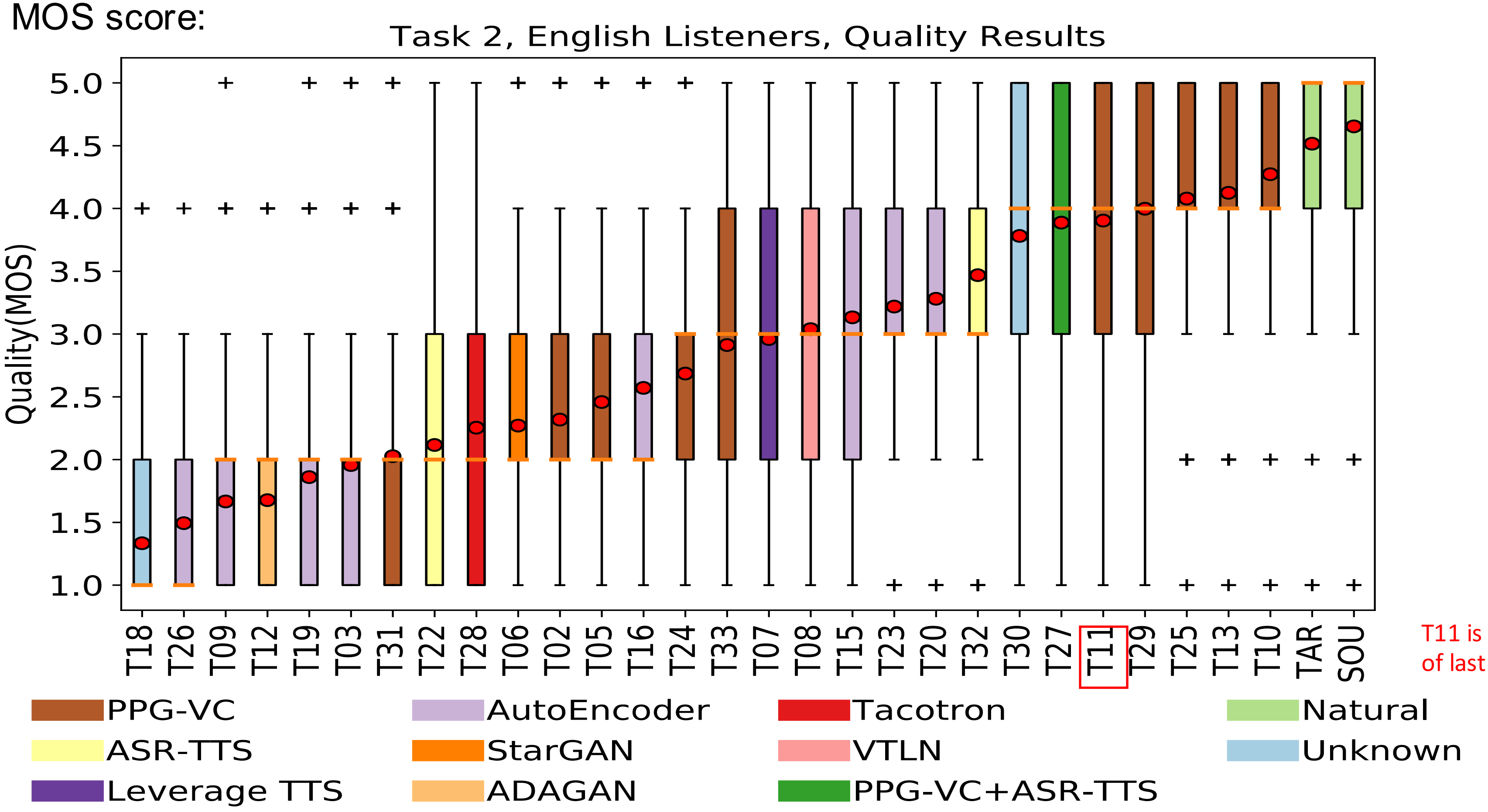
T11 🔊 T29 🔊

T13 🔊 T27 🔊

T10 🔊

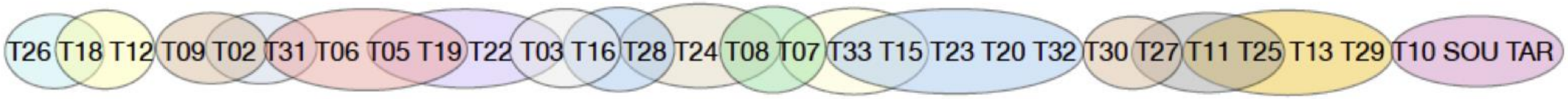
Naturalness results for Task 2

- Bar plot for MOS score:



T11 is the top system of last VCC

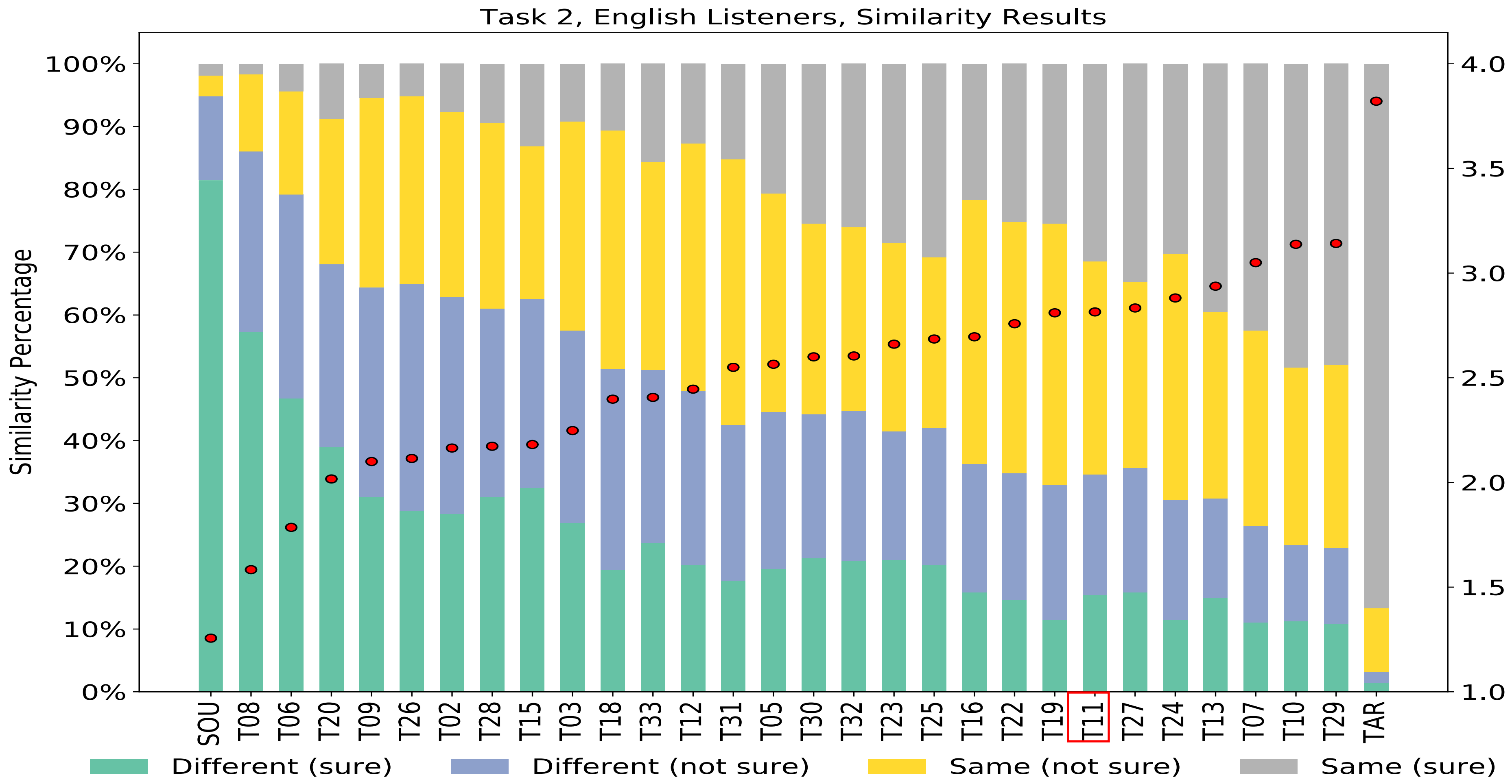
- Groupings of systems that did not differ significantly from each other in terms of naturalness for Task 2:



T26 T18 T12 T09 T02 T31 T06 T05 T19 T22 T03 T16 T28 T24 T08 T07 T33 T15 T23 T20 T32 T30 T27 T11 T25 T13 T29 T10 SOU TAR

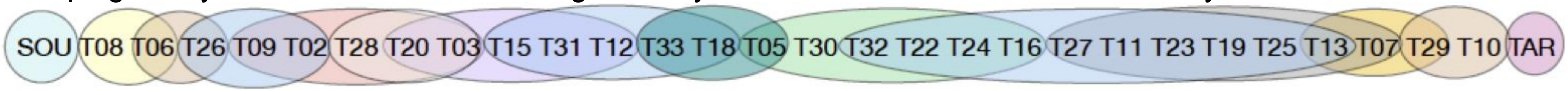
Similarity results for Task 2

- Plot for similarity score:



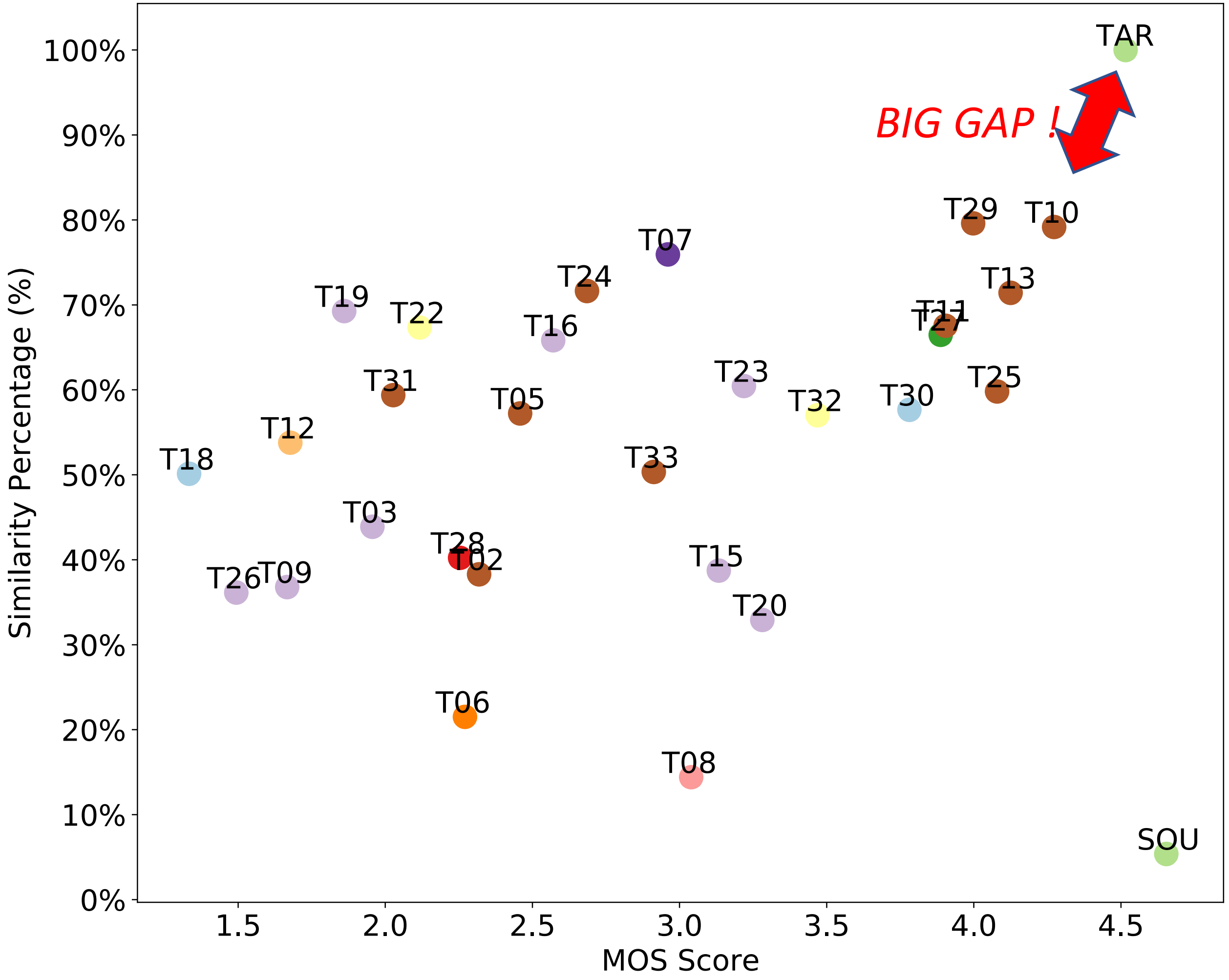
T11 is the top system of last VCC

- Groupings of systems that did not differ significantly from each other in terms of similarity for Task 2:



Scatter plot for Task 2

- Scatter plot between MOS score and speaker similarity percentage:


Task 2, English Listeners, Scatter Plot

Speaker: SEF1 TMM1

Source: Target:

Converted samples of several top teams:

T11 T25

T13 T29

T10

Further analysis

- Listeners: English vs. Japanese
  - ✓ *It is acceptable to use non-native listeners to assess the performance of VC systems to some extent.*
- Reference audio in cross-lingual task: English vs. L2 languages
  - ✓ *Subjects generally gave lower speaker similarity scores in the case of the L2 language reference.*
- Language of target speakers: Finnish vs. German vs. Mandarin
  - ✓ The language of the target speakers affected both the speaker similarity and naturalness of the VC systems. (e.g. the VC systems had the highest MOS and similarity scores for German target speakers and lowest similarity scores for Mandarin speakers)

Please visit https://arxiv.org/abs/2008.12527 for more information!

Outline

- Tasks, databases, and timeline for Voice Conversion Challenge 2020
- Participants and submitted systems
- Subjective evaluations and analysis of VCC 2020 results
- Conclusions

Conclusions

- VCC 2020 *Great progress in techniques!*
  - Intra-lingual conversion :
    - Semi-Parallel dataset: a small parallel dataset + a large non-parallel dataset
    - The best system:
      - Average naturalness MOS: 4.27/5.0 (4.1/5.0 in VCC2018)
      - Over 95% converted speech samples were to be the same as the target speakers (80% in VCC2018).
  - Cross-lingual conversion:
    - Non-parallel, different languages
    - The best system:
      - Average naturalness MOS: 4.27/5.0
      - 75 % converted speech samples were to be the same as the target speakers.

Conclusions

Milestone! The speaker similarity scores of several systems turned out to be as high as target speakers for intra-lingual VC .

None of the system could have achieved human-level naturalness.

The overall naturalness and similarity scores of cross-lingual task were lower than intra-lingual task.

Thank you!

Please visit https://arxiv.org/abs/2008.12527 for more information!