The Voice Conversion Challenge 2018:
Promoting Development of Parallel and Nonparallel Methods

Jaime Lorenzo-Trueba, **Junichi Yamagishi** (National Institute of Informatics, Japan)
Tomoki Toda (Nagoya University, Japan),
Daisuke Saito (University of Tokyo, Japan)
Fernando Villavicencio (Oben, USA),
Tomi Kinnunen (University of East Finland, Finland)
Zhenhua Ling (University of Science and Technology of China, China)

VCC organizers

Organizers for the 2018 challenges

- Junichi Yamagishi & Jaime Lorenzo-Trueba (NII)
- Tomoki Toda (Nagoya University)
- Daisuke Saito (Tokyo University)
- Fernando Villavicencio (ObEN)
- Tomi Kinnunen (University of Eastern Finland): Anti-Spoofing side
- Zhenhua Ling (University of Science and Technology of China)



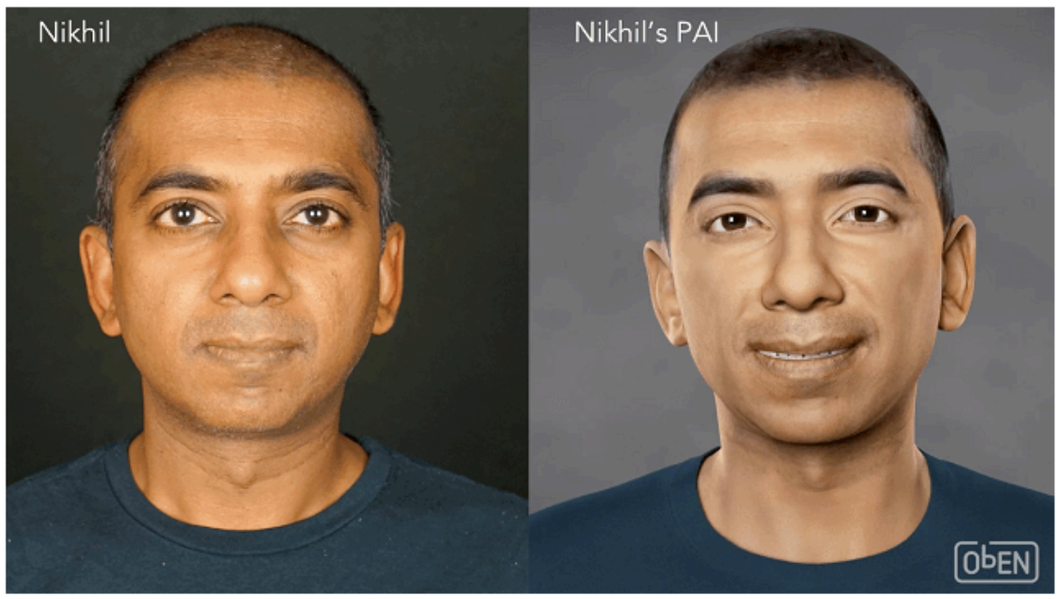




Applications using voice conversions

Who needs Siri or Alexa when you can have a 'Digital Avatar' of yourself? Mind Blown Alert

October 28, 2017



Nikhil Jain, Co-Founder and CEO of ObEN and his PAI. Image courtesy-ObEN

If there's one thing out there Elon Musk is truly afraid of, its Artificial Intelligence

Oben's personal Avatar

Yamaha's singing synthesizer
+ voice conversion

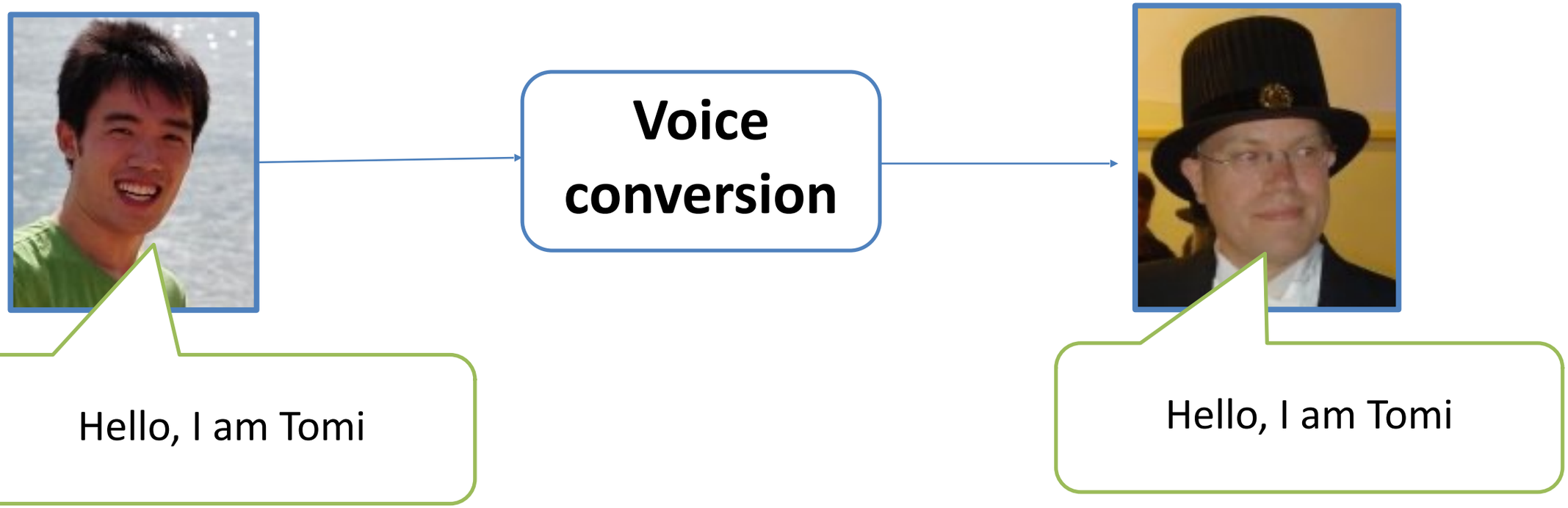
X Japan's hide Releases 'Last Song' With Vocaloid, 16 Years After Passing Away

Musician *hide*, the band *X Japan*'s guitarist who passed away in May 1998, will release a new song in December. The song uses *Yamaha*'s Vocaloid voice synthesis technology to recreate his singing voice, based on lyrics he wrote and a demo he recorded just before he passed away.

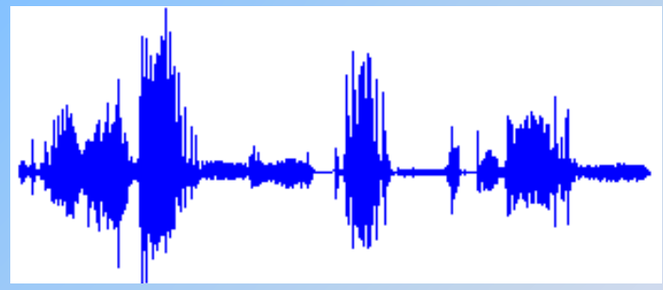
It took about two years to complete this song to commemorate what would have been *hide*'s 50th birthday on December 13.

Voice conversion (VC)

- **<u>Converting para-linguistic information while keeping linguistic information unchanged</u>**
  - Para-linguistic information:
    - speaker identity, speaking styles, etc

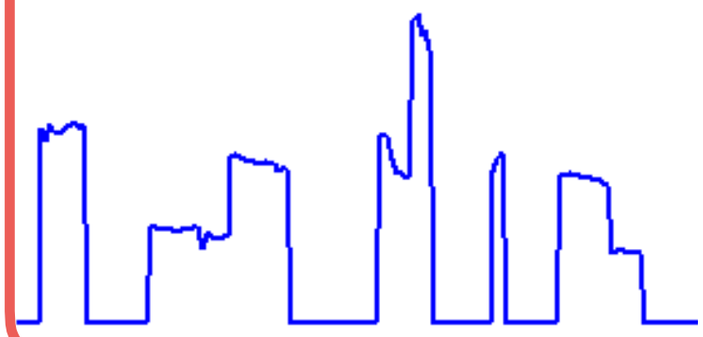


Content

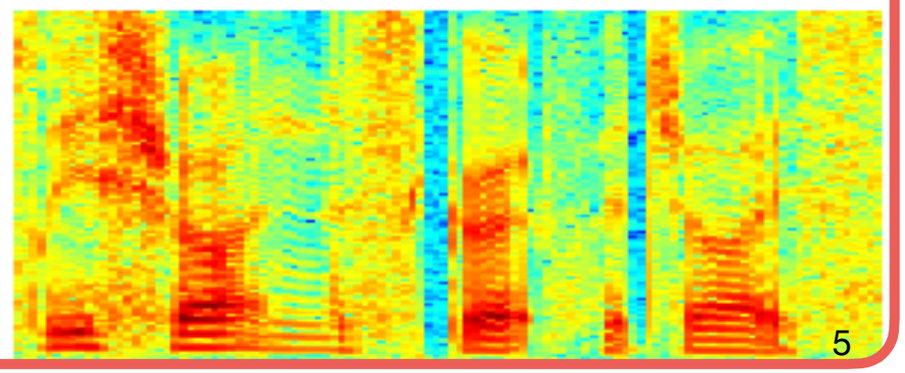


Speech

Prosody

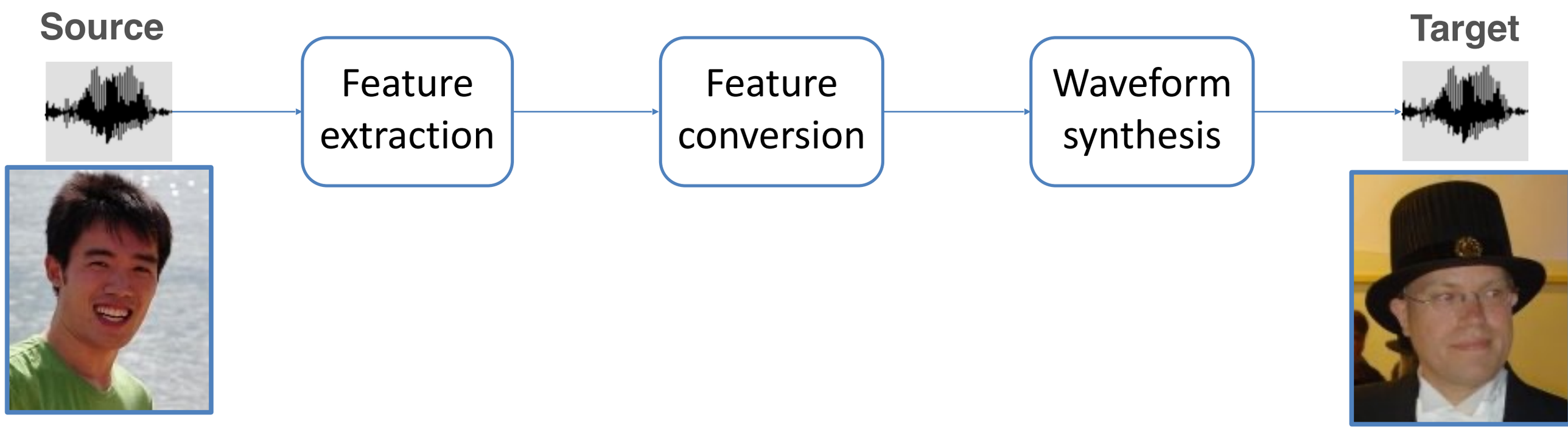


Timbre

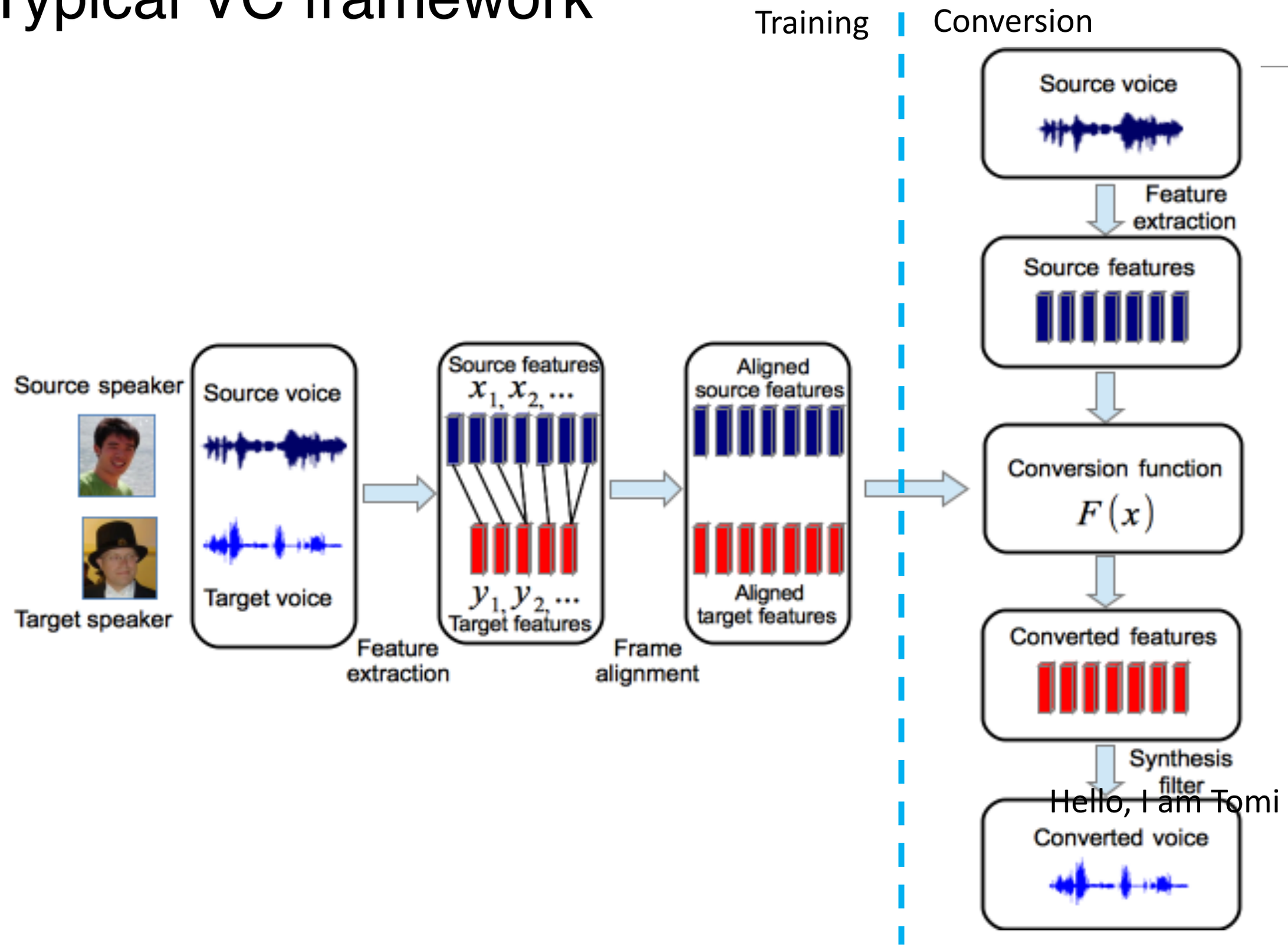


How to convert a voice?

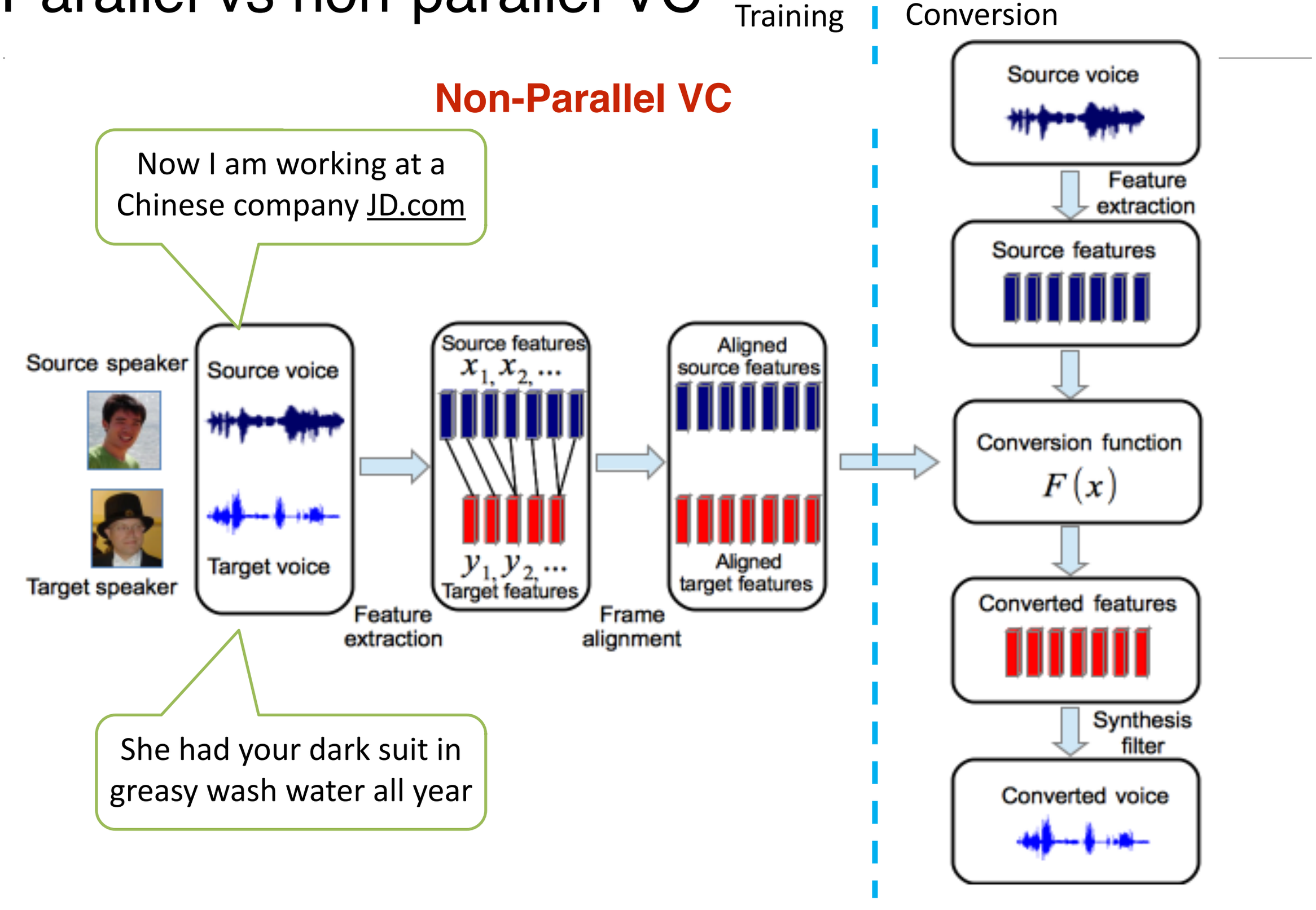
- Waveform to waveform conversion



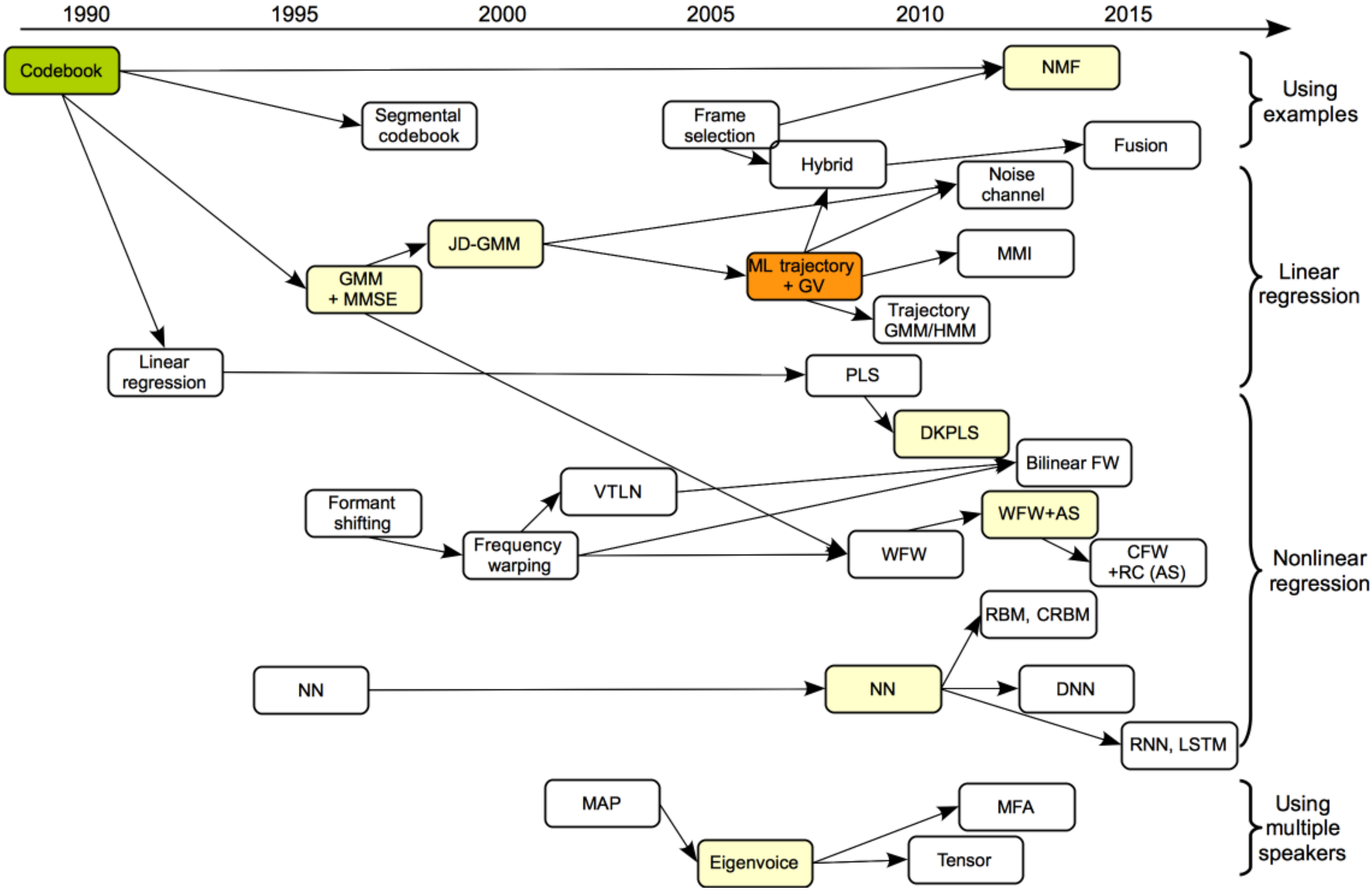
Typical VC framework



Parallel vs non-parallel VC



Progress of voice conversion approaches



Voice conversion challenges

- There are many voice conversion techniques!!
- Need to understand pros and cons of the methods
- **But it was not possible to directly compare results shown in papers with results reported in other papers**
  - Different databases, different training/evaluation lists, different evaluation methodologies

- **Voice conversion challenge (VCC) launched in 2016**
- **Motivations**
  - Understand the state of the art of Voice conversion techniques
  - Standard database
  - Common protocol
  - Common evaluation metric

Schedule of VCC 2018

Timeline

- **October 1st, 2017**: release of training data
- **December 1st, 2017**: release of evaluation data
- **December 8th, 2017**: deadline to submit the converted audio.
- **January 26th, 2018**: notification of results
- **February 25th, 2018** paper submission
- **June 26-29th, 2018**: special session at the 2018 Odyssey workshop

Participants were asked

1. to **build their VC system based on the common database and protocols released from the organizers for 2 months**, and
2. to submit converted speech to the organizers

VCC 2018 database

- DAPS (Data And Production Speech) [Mysore, 2015]
  - 20 professional US English speakers
  - Clean reading speech recorded in a professional studio
  - Freely available [https://archive.org/details/daps_dataset]

- Design of VCC 2018 datasets
  - Manually segment DAPS audio files into individual sentences
  - Select 12 speakers (6 female and 6 male speakers)
  - Down-sampled to 22.05 kHz

Main **Hub** task

- The VCC 2018 database contains two tasks, *Hub and Spoke*

- Hub task: Parallel voice conversion
  - Participants build their VC systems using speech debases where source and target speaks read out the **SAME** sets of sentences
    - 4 source speakers (2M, 2F), 4 target speakers (2M, 2F)
    - 81 sentences for each speaker
    - Build all 16 combinations of speaker pairs
    - Generate 35 converted utterances for each pair

- All participants have to do this task

Optional **spoke** task

- Spoke task: Non-parallel voice conversion
  - Participants build their VC systems using speech debases where source and target speaks read out the **DIFFERENT** sets of sentences
  - 4 source speakers (2M, 2F), 4 target speakers (same as Hub task)
  - 81 sentences for each speaker
    - Build all 16 combinations of speaker pairs
    - Generate 35 converted utterances for each pair

- This is an optional task for participant
- Systems for hub and spoke tasks were evaluated at the same time

- Total 32 unique speaker pairs
  - 16 speaker pair for the hub task + 16 speaker for the spoke task

Baseline systems

- **Distributed two baseline systems as open source program:**
-
    - **Sprocket**:
        - Open-source implementation of a Winner system in the VCC 2016 challenge
        - GMM-based voice conversion system that directly modify speech waveforms
        - https://github.com/k2kobayashi/sprocket
        - This is named ***B01***
- **Merlin**:
    - Open-source implementation of a DNN based VC system
    - https://github.com/CSTR-Edinburgh/merlin/
    - This is named **D05**

Participants

- Registered organizations：75
- Organizations who submitted converted speech：23

tively.

Team name	Institution name	Tasks
AhoLab	University of the Basque Country	H
AS	STMS-IRCAM/Sorbonne University/CNRS/IntelligentVoice	H,S
AST	Academia Sinica	H,S
Azurite	Indian Institute of Technology Bombay	H,S
CMU	Carnegie Mellon University	H
CPqD	CPqD	H
CSLU	Oregon Health & Science University	H
CSTR	University of Edinburgh	H
CUHK	The Chinese University of Hong Kong	H,S
DA-IICT	Dhirubhai Ambani Institute of Information and Communication Technology	H,S
DSP-AGH	AGH University of Sciencie and Technology	H
Hulk2	Shanghai Jiao Tong University	H
NWPU-I2R-NUS	Northwestern Polytechnical University/Institute for Infocomm Research/National University of Singapore	H
NTT-CSlab	Nippon Telegraph and Telephone Corporation	H,S
NTU	Nanyang Technological University	H,S
NTUT	National Taipei University of Technology	H
NU	Nagoya University	H,S
PDL	Pindrop	H
RBM	University of Electro-Communications	H,S
TEXAGS	Texas A&M University	H,S
USTC	University of Science and Technology of China	H,S
UTokyo	The University of Tokyo	H
xmuspeech	Xiamen University	H

Evaluation methodology

- Subjective evaluation (listening tests) using crowdsourcing
- Listeners judge the following aspects of converted speech
  - **Quality of converted speech:**
    - 5-point scale
    - 1: very unnatural, 5: very natural
  - **Similarity of converted speech to a target speaker:**
    - Compared with natural speech of the target speaker
    - Same/different judgement using 4-point scale
      - *Same speaker, sure*
      - *Same speaker, not sure*
      - *Different speaker, not sure*
      - *Different speaker, sure*

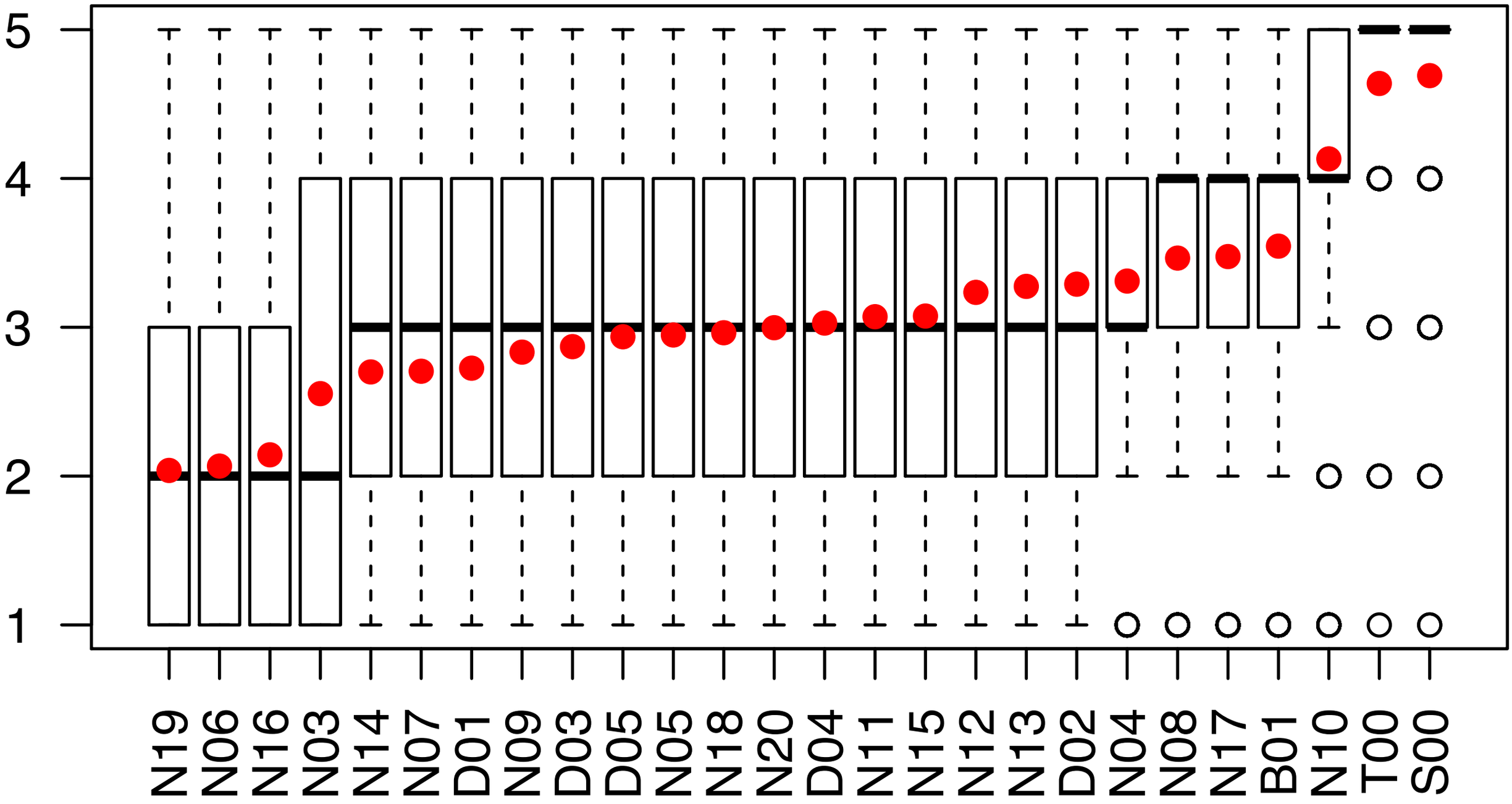
Subjects

- **Large scale evaluation with 267 paid subjects**
  - Majorities are native speakers of English (American or British)
  - 146 male, 121 female

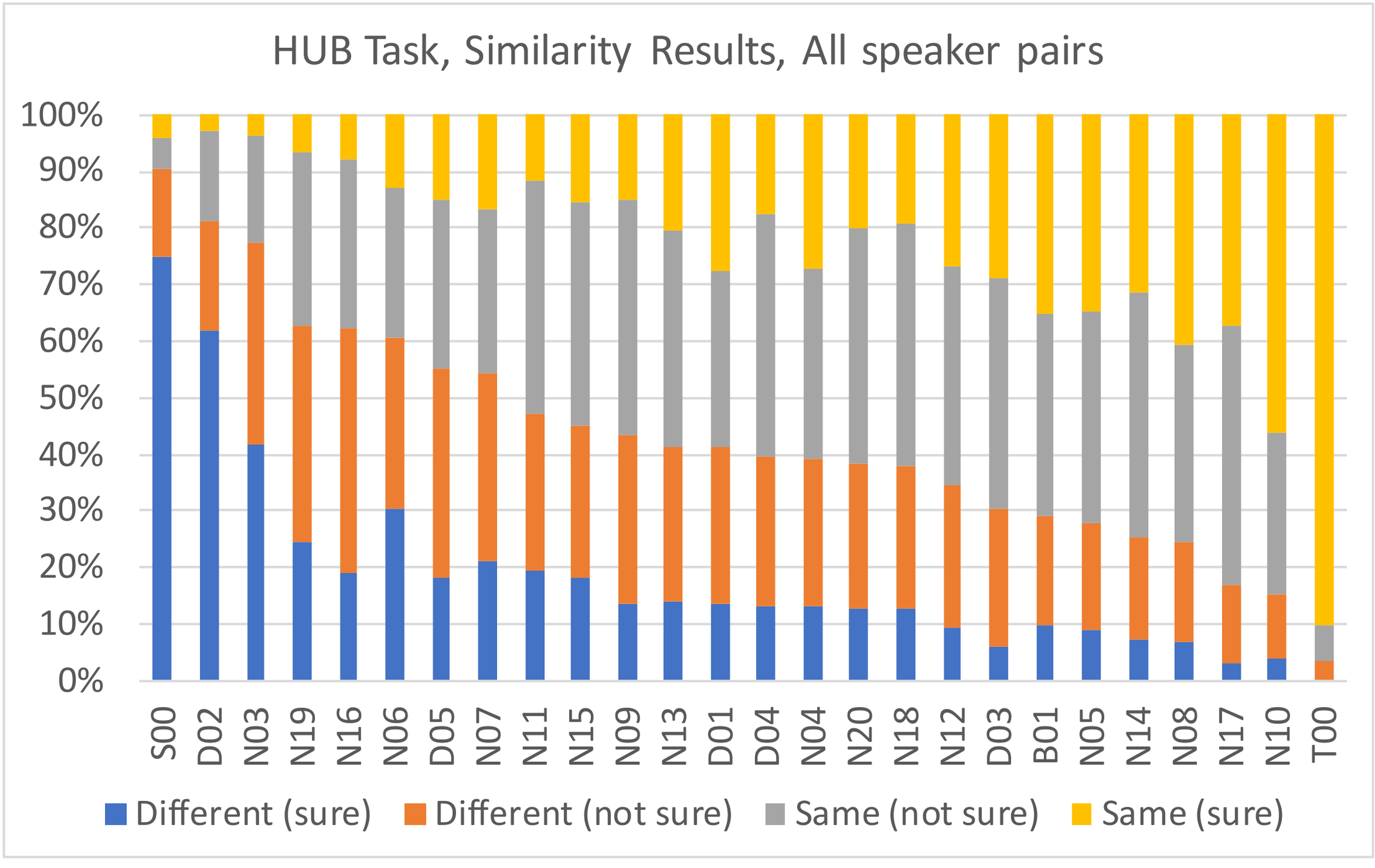
Age	#	Accent	#
18-30	116	North American	141
31-40	94	British	58
41-50	45	Other	22
51+	12	Non-native	46

- **A largest listening test that we organized ever!**
  - 16 speaker pairs x 38 systems (23 hub systems + 11 spoke systems + baseline for each + 4 human speech) x 35 utterances x 4 coverages

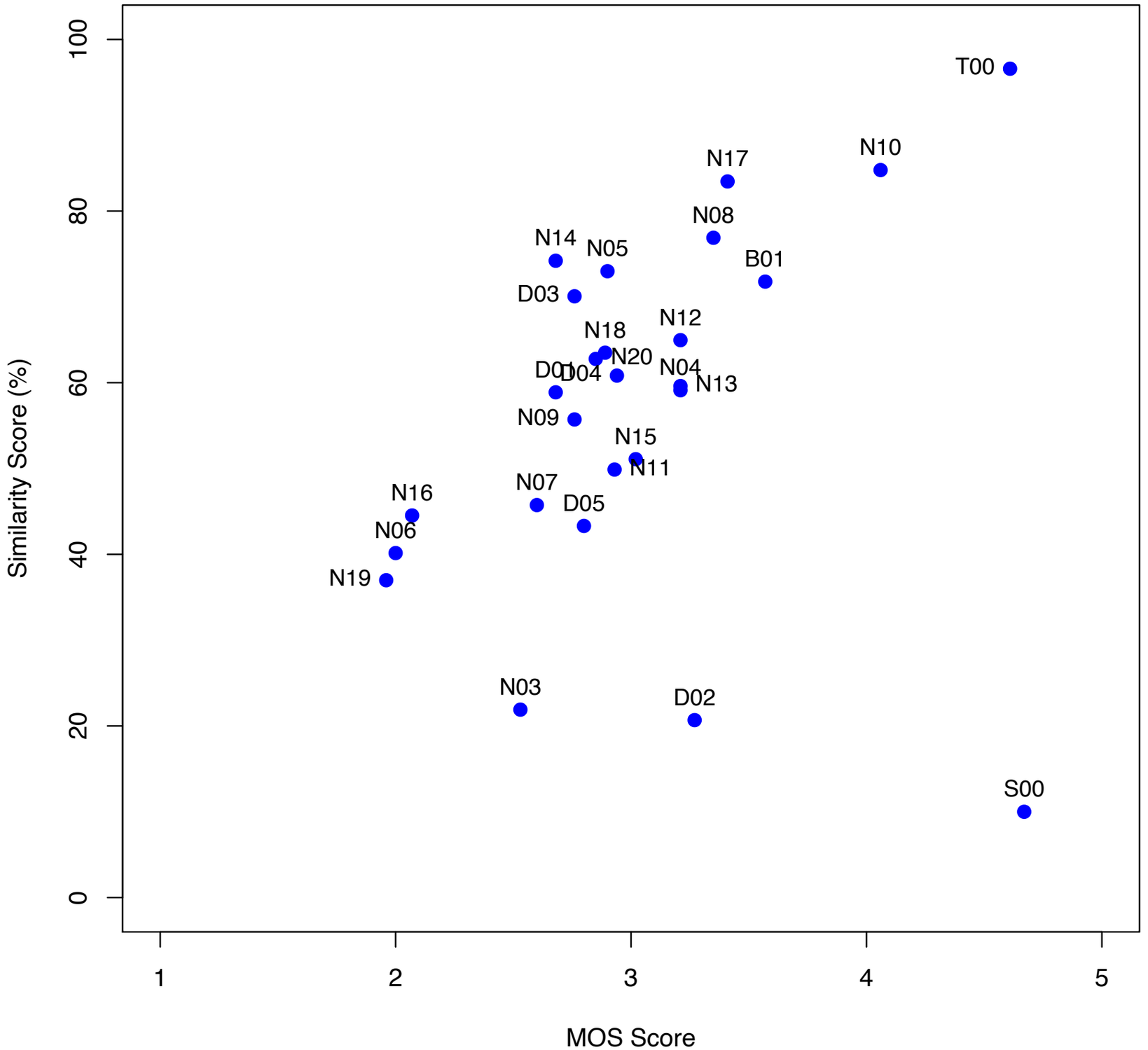
Quality evaluation (Hub task)



Speaker similarity evaluation (Hub task)

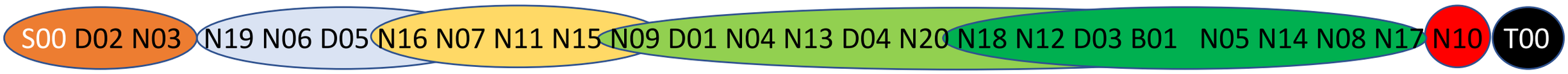


Quality and similarity visualization (Hub task)



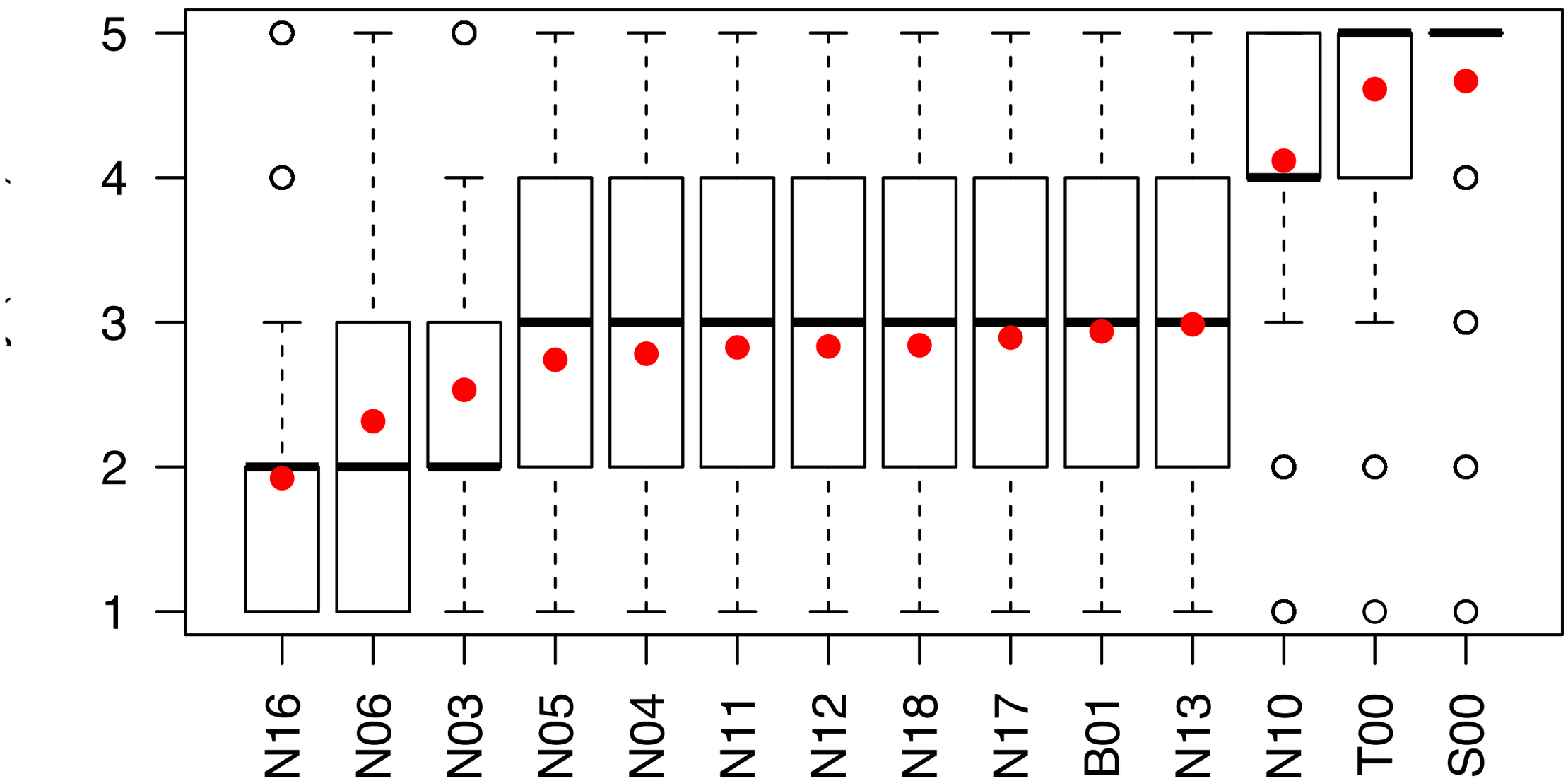
Significant differences (Hub task)

(a) Groupings of systems that do not differ significantly from each other in terms of naturalness (Hub task, all speaker pairs)

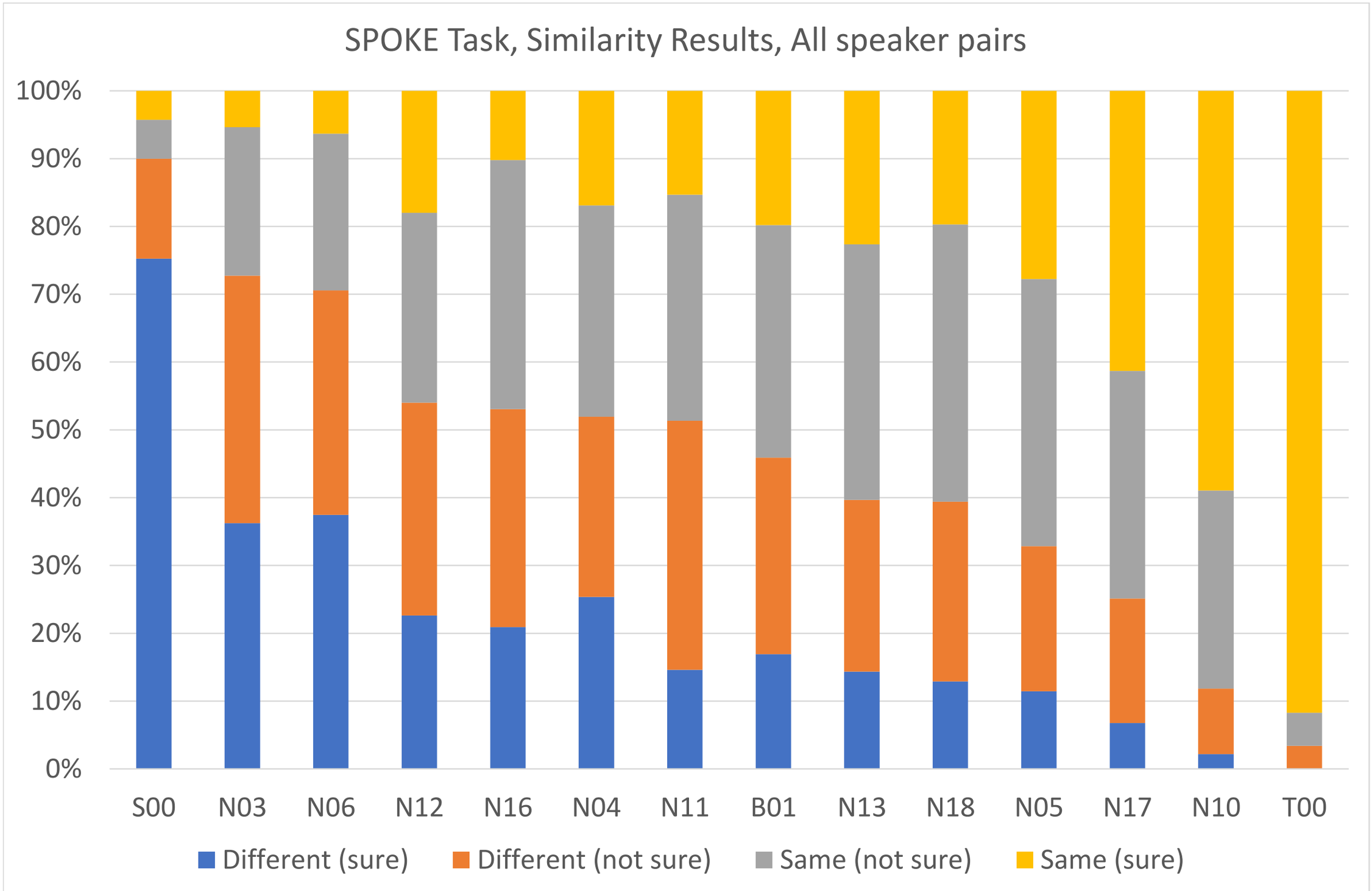


(b) Groupings of systems that do not differ significantly from each other in terms of similarity to target speaker (Hub task, all speaker pairs)

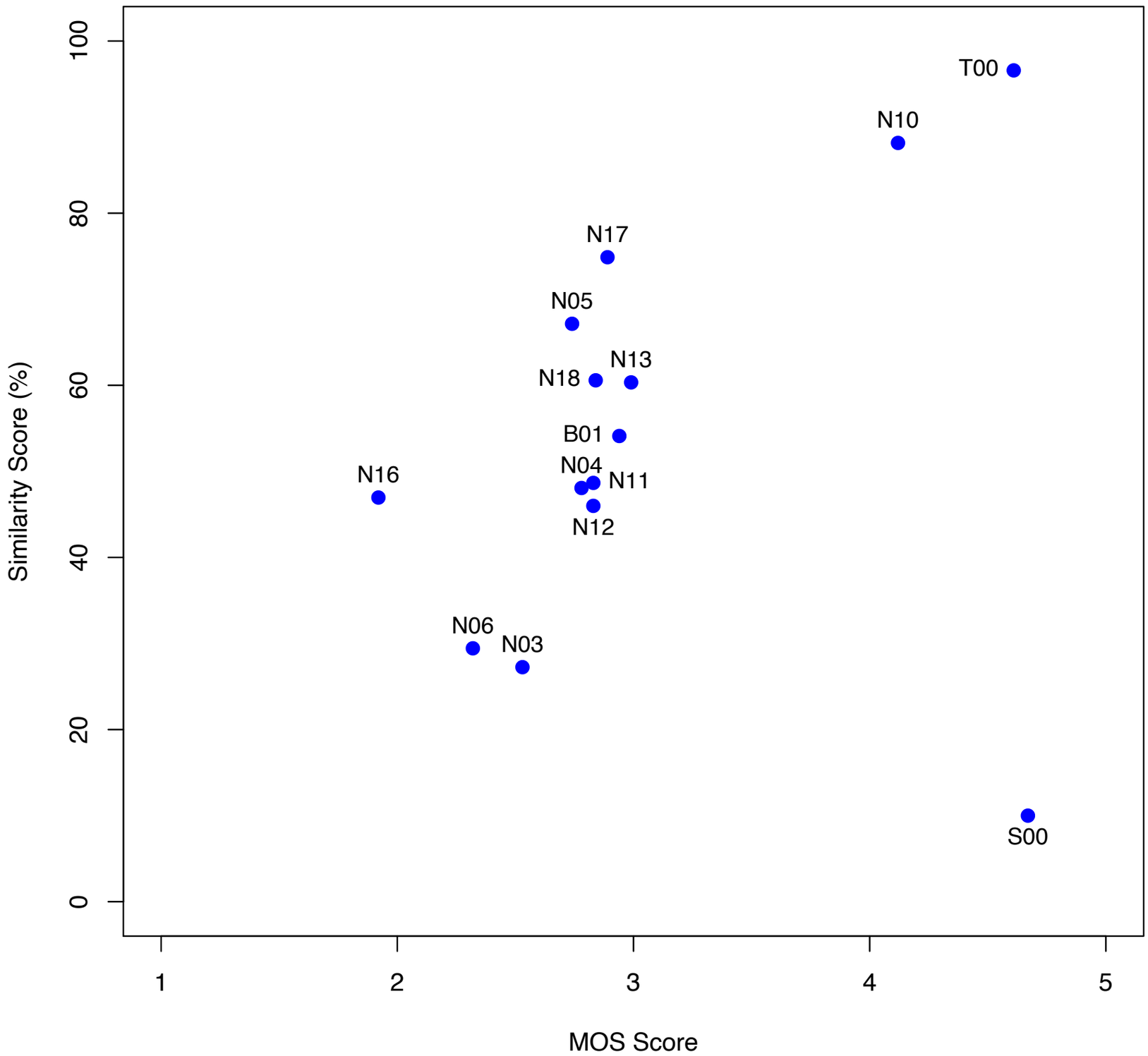
Quality evaluation (Spoke task)



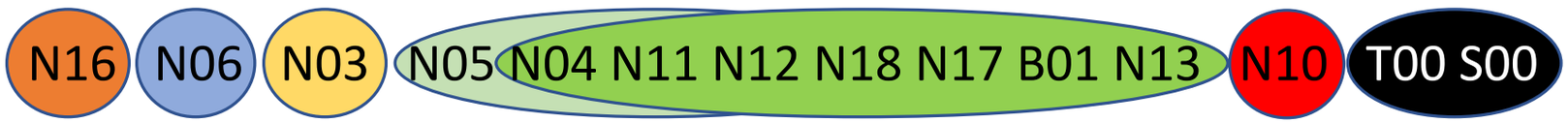
Speaker similarity evaluation (Spoke task)



Quality and similarity visualization (Spoke task)



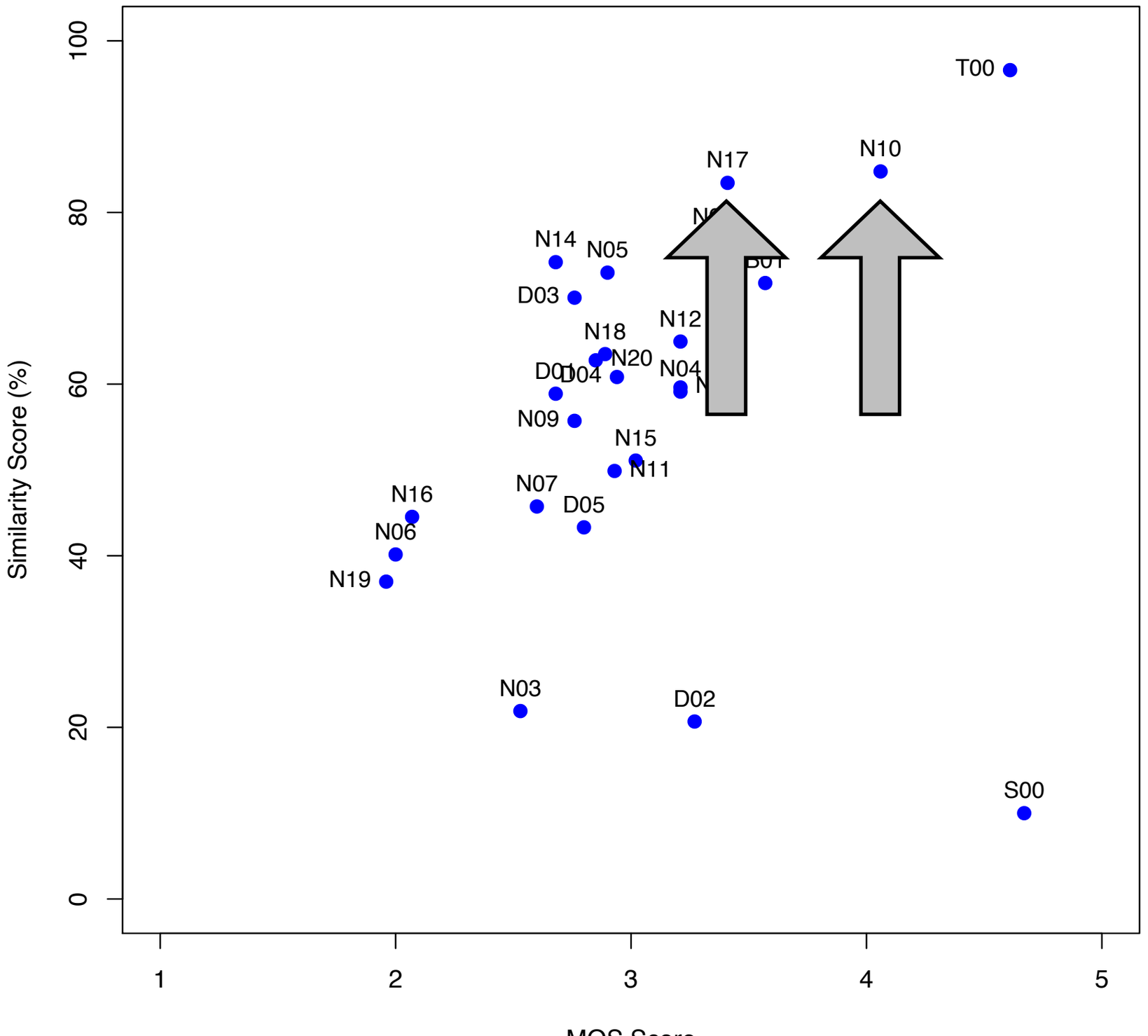
Significant differences (Spoke task)



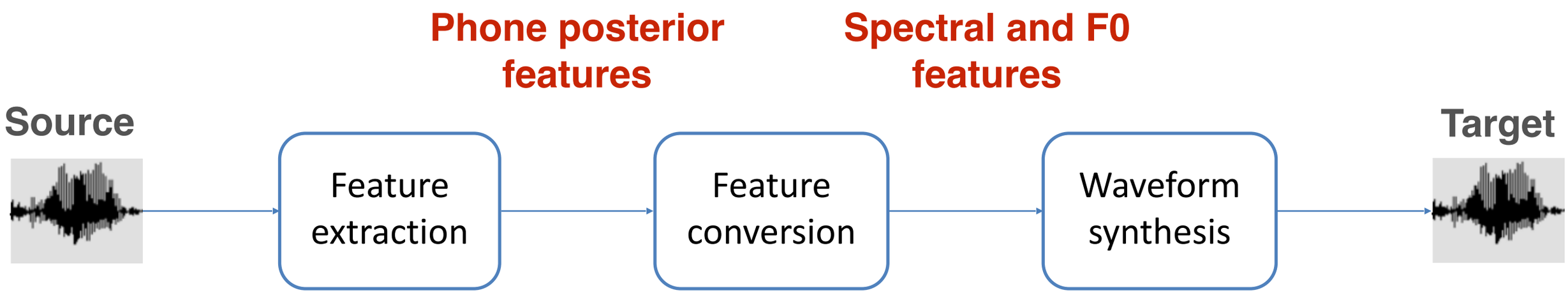
(c) Groupings of systems that do not differ significantly from each other in terms of naturalness (Spoke task, all speaker pairs)

(d) Groupings of systems that do not differ significantly from each other in terms of similarity to target speaker (Spoke task, all speaker pairs)

Quality and similarity visualization (Hub task)



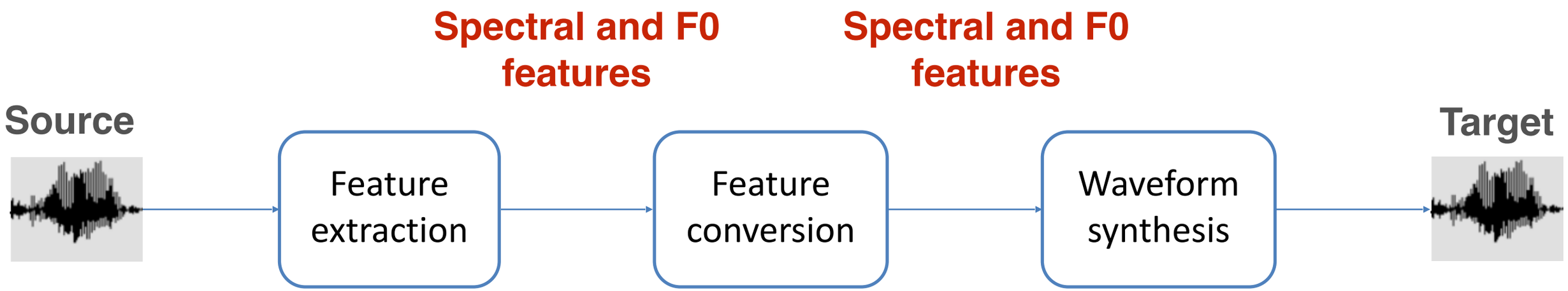
Best system N10



Context posterior features extracted by a speaker independent phone recognizer trained using hundreds of hours of external speech data	Speaker-dependent LSTM that maps context features to vocoder features (mel-cepstrum, F0, UV/V); no need of parallel training samples	Speaker dependent Wavenet (CNN-AR) that maps the vocoder parameters to waveform samples: Pre-trained using an external multi-speaker dataset including 10 male and 10 female (80 hrs)

More details about this system have been described in a paper titled “WaveNet Vocoder with Limited Training Data for Voice Conversion" which has been accepted to Interspeech 2018.

2nd best (?) system N17



Signal processing: WORLD analysis to extract acoustic features (mel-cepstra, Two-band aperiodicity, F0, UV/V)

Speaker-pair dependent mixture density net (MDN) that maps mel-cepstrum to target ones: Global variance enhancement and differential waveform modification are also used as post processing: F0 is linearly converted

Speaker-dependent Wavenet (CNN-AR) that maps target vocoder parameters to waveform samples: Pre-trained using all target/source speakers and two external speakers included in ARCTIC DB: Output layer was fine-tuned

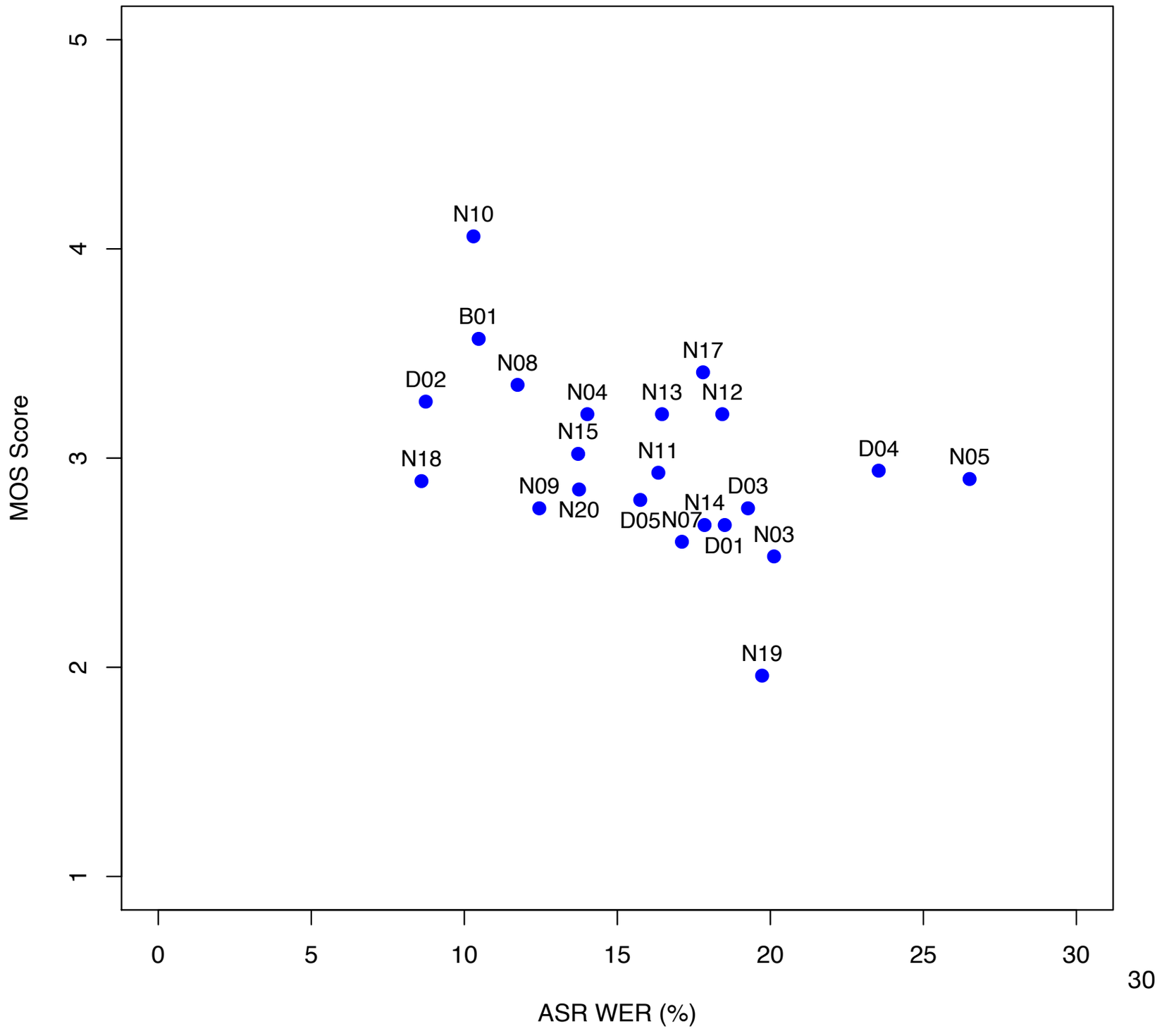
More details about this system will be presented at a poster titled “NU Voice Conversion System for the Voice Conversion Challenge 2018 in a poster session.

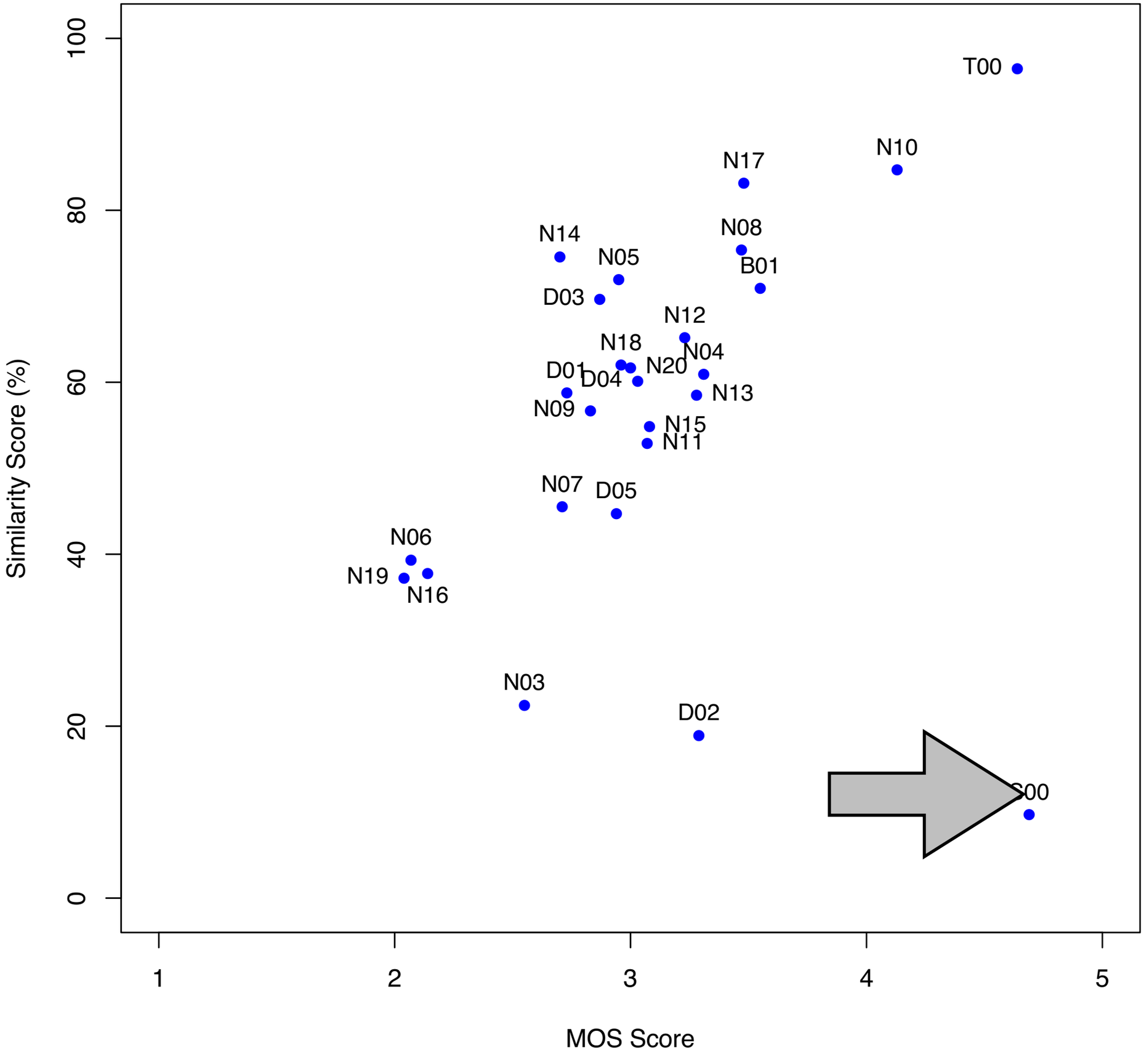
Correlation between MOS vs WER

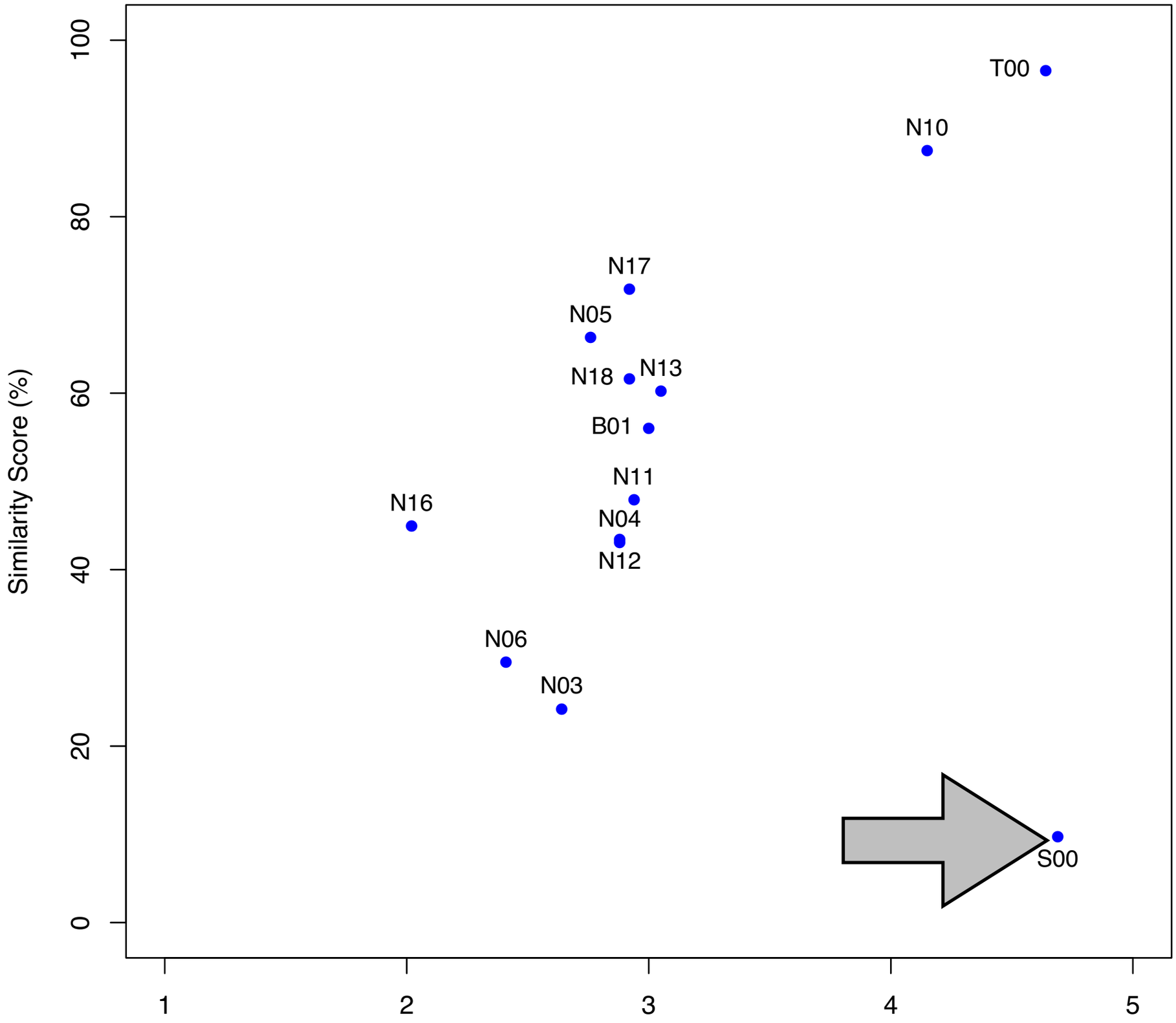
- ASR was used to guess intelligibility of converted speech

- MOS scores on the quality are found to be correlated with ASR's WER scores
  - r = –0.6587

- **Lower intelligibility tends to be judged as lower quality**
- N10 vs N17




Similarity Score (%)
0
20
40
60
80
100
1
2
3
4
5
MOS Score
T00
N10
N17
N14
N08
N05
B01
D03
N12
N18
N04
D01
N20
D04
N13
N09
N15
N11
N07
D05
N06
N19
N16
N03
D02
C00


Similarity Score (%)
0
20
40
60
80
100
1
2
3
4
5
MOS Score
T00
N10
N17
N05
N18
N13
B01
N11
N16
N04
N12
N06
N03
S00

Comparison with VCC 2016

- VCC 2016
  - Parallel voice conversion
  - 162 sentences for training
  - Participants: 17 teams
  - Results: Best MOS score= 3.5, Best speaker similarity score= 75%

- VCC 2018
  - Parallel and non-parallel voice conversion
  - **81 sentences for training** (a half of vcc2016 case)
  - **Participants: 23 teams**
  - Results: **Best MOS score= 4.1, Best speaker similarity score= 80%**

Summary

- Overview of VCC2018
  - Common evaluation of advanced voice conversion techniques using a standard database and common protocol
  - Very large-scale listening test of many VC systems built by 23 teams

- N10 has shown the incredible progress:
  - Its MOS score was 4.1 out of 5 and 80% of its converted speech was judged as the same speaker as a target speaker
  - Similar performance in both parallel and non-parallel tasks
- ***This could be a breakthrough VC technology***
- This also implies that the best VC technology can fool human perception.

- How about their spoofing capability? How about machine perception?

VCC 2018 database freely available

Edinburgh DataShare

THE UNIVERSITY of EDINBURGH

INFORMATION SERVICES

Contact us

The Voice Conversion Challenge 2018: database and results

No Thumbnail

Date Available
2018-04-10

Type
sound

Data Creator
Lorenzo-Trueba, Jaime
Yamagishi, Junichi
Toda, Tomoki
Saito, Daisuke
Villavicencio, Fernando
Kinnunen, Tomi

Citation
Lorenzo-Trueba, Jaime; Yamagishi, Junichi; Toda, Tomoki; Saito, Daisuke; Villavicencio, Fernando; Kinnunen, Tomi; Ling, Zhenhua. (2018). The Voice Conversion Challenge 2018: database and results, [sound]. The Centre for Speech Technology Research, The University of Edinburgh, UK. http://dx.doi.org/10.7488/ds/2337.

Description
Voice conversion (VC) is a technique to transform a speaker identity included in a source speech waveform into a different one while preserving linguistic information of the source speech waveform. In 2016, we have launched the Voice Conversion Challenge (VCC) 2016 at Interspeech 2016. The objective of the 2016 challenge was to better understand different VC techniques built on a freely-available common dataset to look at a common goal, and to share views about unsolved problems and challenges faced by the current VC techniques. The VCC 2016 focused on the most basic VC task, that is, the construction of VC models that automatically transform the voice identity of a source speaker into that of a target speaker using a parallel clean training database where source and target speakers read out the same set of utterances in a professional recording studio. 17 research groups had participated in the 2016 challenge. The challenge

Our paper available at ArXiv

Electrical Engineering and Systems Science > Audio and Speech Processing

The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods

Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, Zhenhua Ling

(Submitted on 12 Apr 2018)

We present the Voice Conversion Challenge 2018, designed as a follow up to the 2016 edition with the aim of providing a common framework for evaluating and comparing different state-of-the-art voice conversion (VC) systems. The objective of the challenge was to perform speaker conversion (i.e. transform the vocal identity) of a source speaker to a target speaker while maintaining linguistic information. As an update to the previous challenge, we considered both parallel and non-parallel data to form the Hub and Spoke tasks, respectively. A total of 23 teams from around the world submitted their systems, 11 of them additionally participated in the optional Spoke task. A large-scale crowdsourced perceptual evaluation was then carried out to rate the submitted converted speech in terms of naturalness and similarity to the target speaker identity. In this paper, we present a brief summary of the state-of-the-art techniques for VC, followed by a detailed explanation of the challenge tasks and the results that were obtained.

Comments: Accepted for Speaker Odyssey 2018
Subjects: **Audio and Speech Processing (eess.AS)**; Computation and Language (cs.CL); Sound (cs.SD); Machine Learning (stat.ML)
Cite as: **arXiv:1804.04262 [eess.AS]**
(or **arXiv:1804.04262v1 [eess.AS]** for this version)

Submission history

From: Junichi Yamagishi Dr. [view email]
[v1] Thu, 12 Apr 2018 00:14:10 GMT (388kb,D)

Sponsors

This work was supported in part by

- iFLYTEK (http://www.iflytek.com/en/)
- MEXT KAKENHI Grant Number JP17H06101
- MEXT KAKENHI Grant Numbers (15H01686, 16H06302, 17H04687)
- Academy of Finland (grant no. 309629)