Analysis of the Voice Conversion Challenge 2016 Evaluation Results

Mirjam Wester, Zhizheng Wu &
Junichi Yamagishi

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Voice Conversion

Voice converted voices were evaluated in terms of naturalness and similarity. The questions we addressed were:

1. How natural does the voice converted voice sound?

2. How similar does the voice converted voice sound compared to the target speaker and to the source speaker?

Naturalness

- How to make task do-able for listeners?
- How to measure naturalness?

Amount of data…

- 5 target and 5 source speakers -> 25 voices.
- 17 participants + baseline: 20 * 18 = *450 voices* !
- Reduced source-target (ST) pairs from 25 to 16
- 288 voices + 4 source + 4 target = 296 stimuli
  —> 50 minutes
- It would take *too long* for a single listener to judge naturalness and similarity

Amount of data…

- Instead of asking each listener to judge all ST pairs how about just one single ST pair?
- In terms of time this would be an excellent solution.
- However, each listener would then only encounter one gender condition and listeners needed to encounter the full range of gender conditions as ratings are context-sensitive.

Our solution...

- Intermediate solution: each listener hears 8 source-target (ST) pairs
- Two from each gender condition, to make the two sets as comparable as possible.

How to measure?

- Standard MOS like Blizzard for naturalness
- (1) totally unnatural to (5) completely natural
- The subjects were instructed that the score should reflect their opinion of how natural or unnatural the sentence sounded

Listeners

- Each set was rated by 100 subjects
- Duration roughly 25 minutes
- The order of stimuli was random
- Each sentence selected at random with replacement from pool of 30 test sentences
- Sentences > 5 sec or < 2 sec were removed for the listening tests (hence not 54 sentences)

Similarity

- Judging how similar voices are on a scale from 1 to 5 may not be all that meaningful.
- Judging how similar two voices are not part of everyday speech perception.
- However, recognising speakers is something we do all the time.
- —> Same/different paradigm

Similarity: exp set-up

- Listeners were given pairs of stimuli and the instructions:

- *“Do you think these two samples could have been produced by the same speaker? Some of the samples may sound somewhat degraded/distorted. Please try to listen beyond the distortion and concentrate on identifying the voice. Are the two voices the same or different? You have the option to indicate how sure you are of your decision.”*

Similarity: exp set-up

- The scale for judging was:
  - Same: absolutely sure
  - Same: not sure
  - Different: not sure
  - Different: absolutely sure
- VC stimuli compared to target speaker and to source speaker.

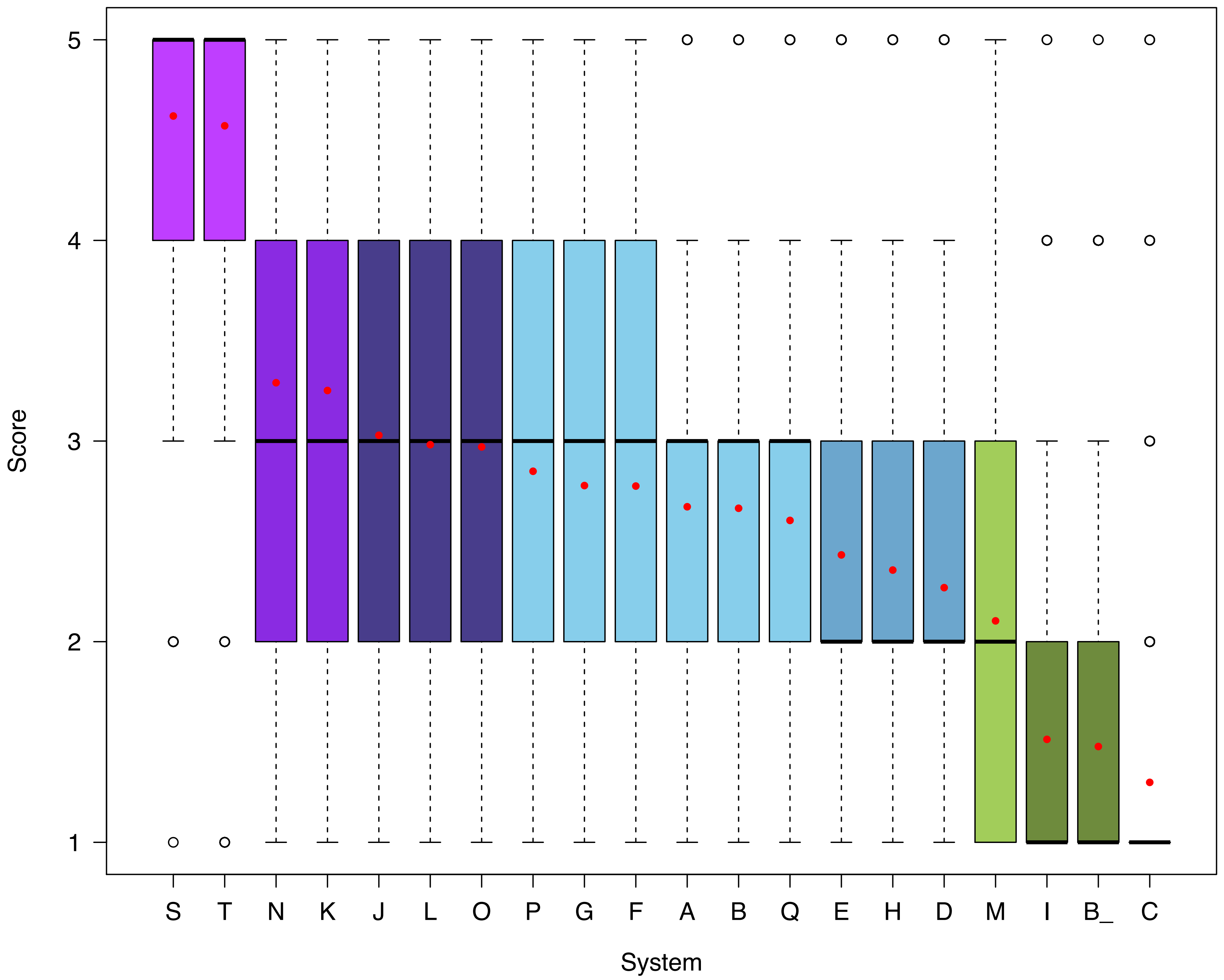
Similarity: exp set-up

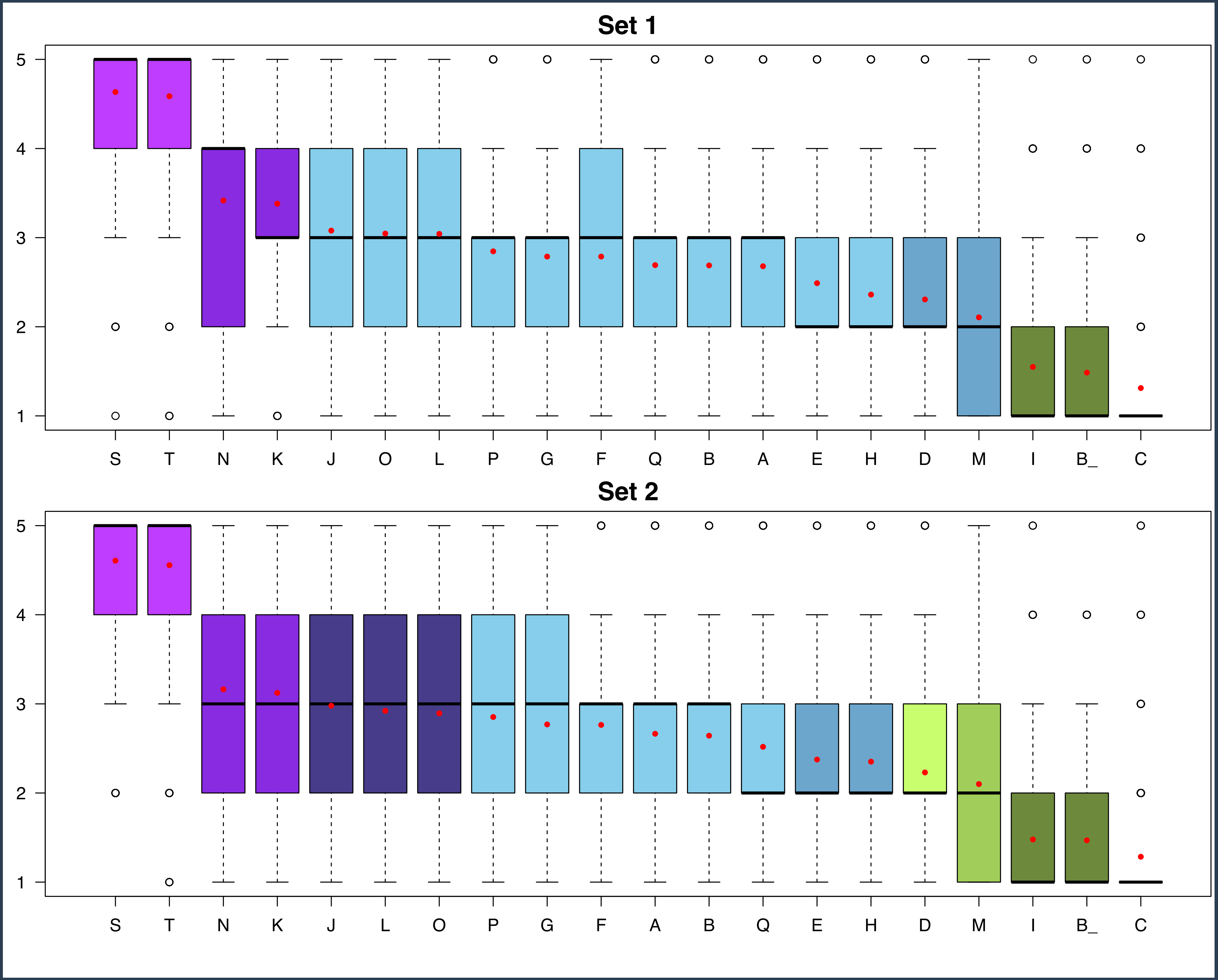
- Each listener was given three ST pairs to judge, one within-gender, one cross-gender and one at random ensuing all ST pairs were covered across listeners.

- 200 listeners

Results

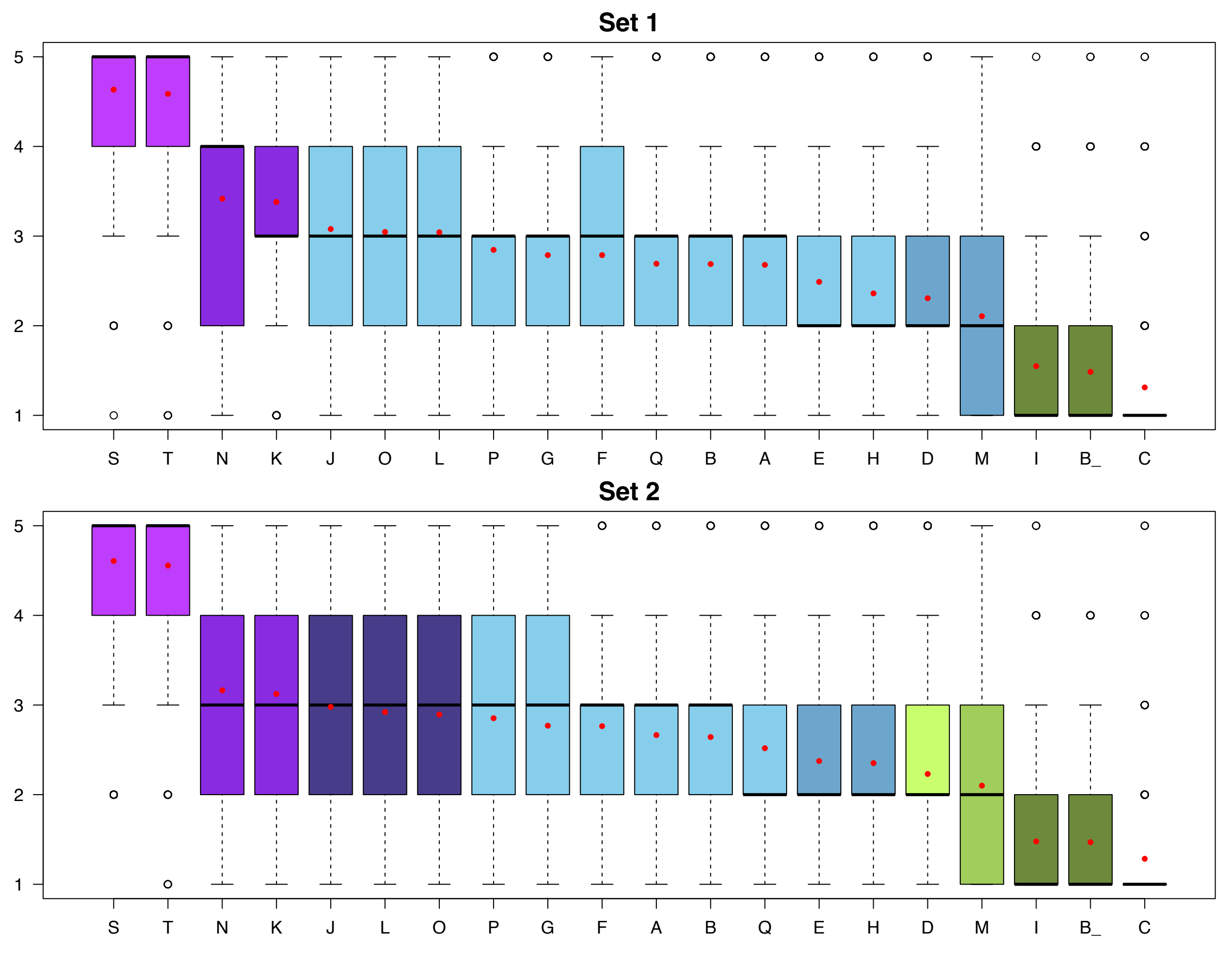
- Naturalness -MOS

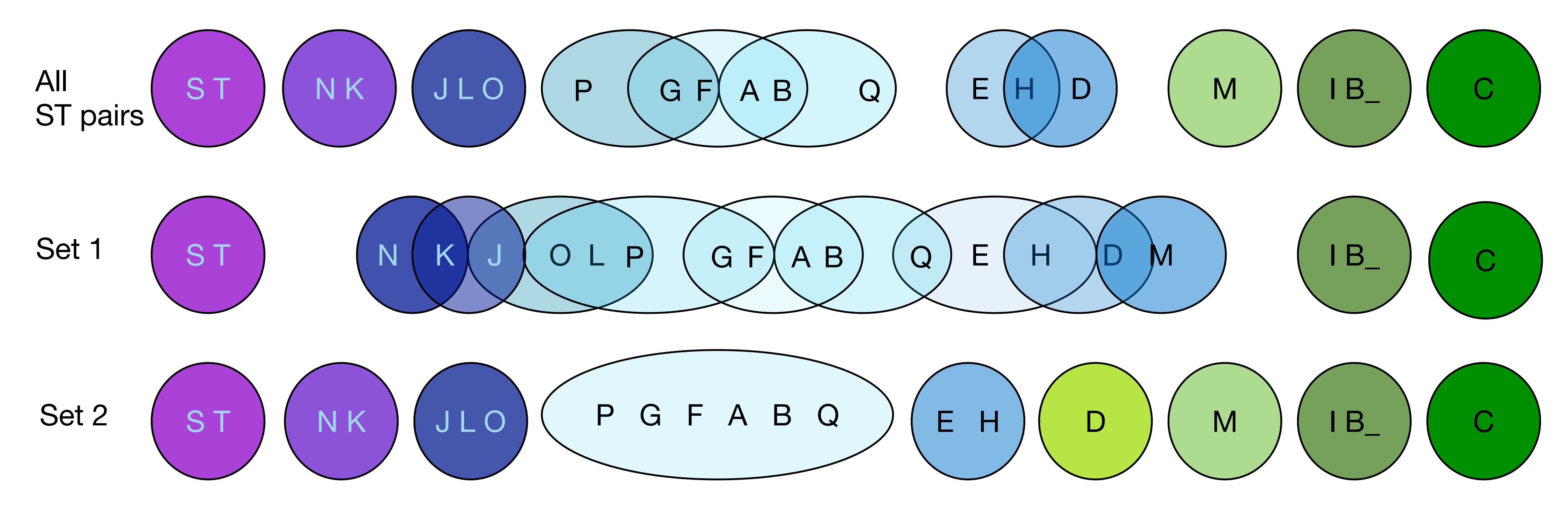

Score
5
4
3
2
1
S
T
N
K
J
L
O
P
G
F
A
B
Q
E
H
D
M
I
B_
C
System

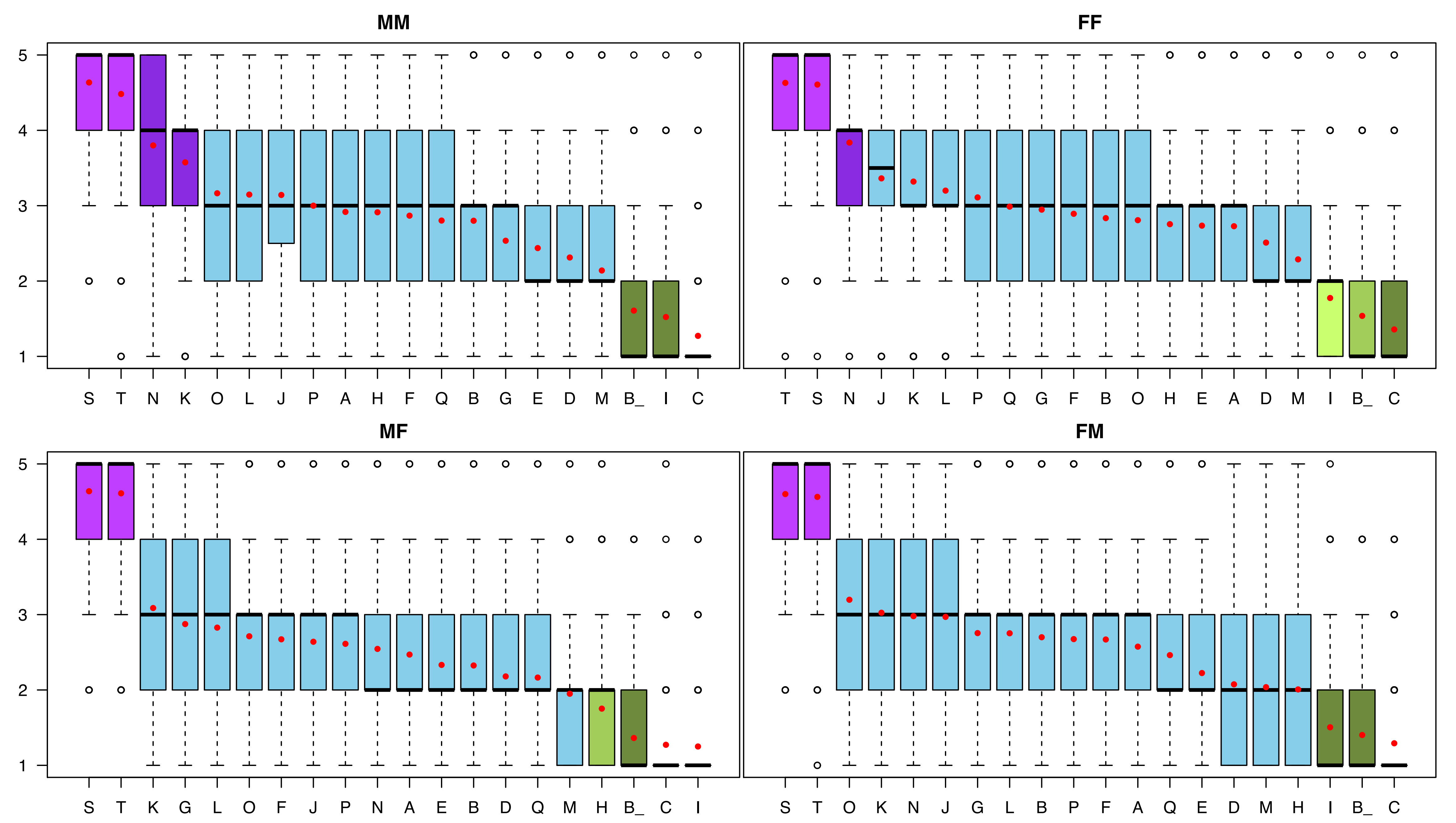


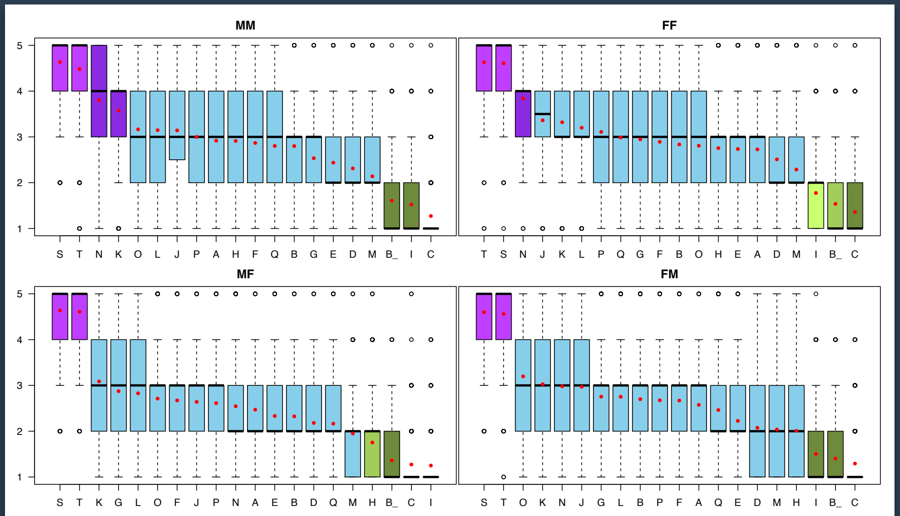
Set 1
5
4
3
2
1
S
T
N
K
J
O
L
P
G
F
Q
B
A
E
H
D
M
I
B_
C
Set 2
5
4
3
2
1
S
T
N
K
J
L
O
P
G
F
A
B
Q
E
H
D
M
I
B_
C

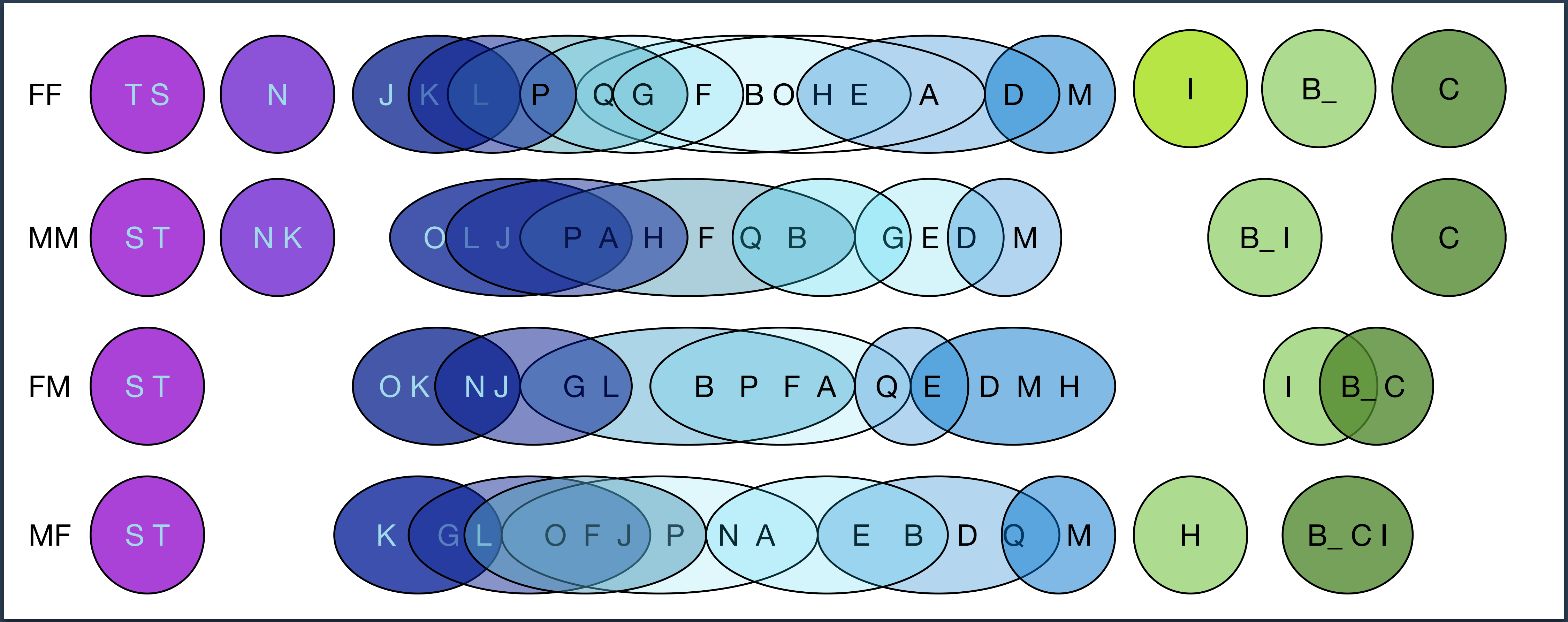
Significance






MM
FF
MF
FM
5
4
3
2
1
S T N K O L J P A H F Q B G E D M B_ I C
T S N J K L P Q G F B O H E A D M I B_ C
S T K G L O F J P N A E B D Q M H B_ C I
S T O K N J G L B P F A Q E D M H I B_ C

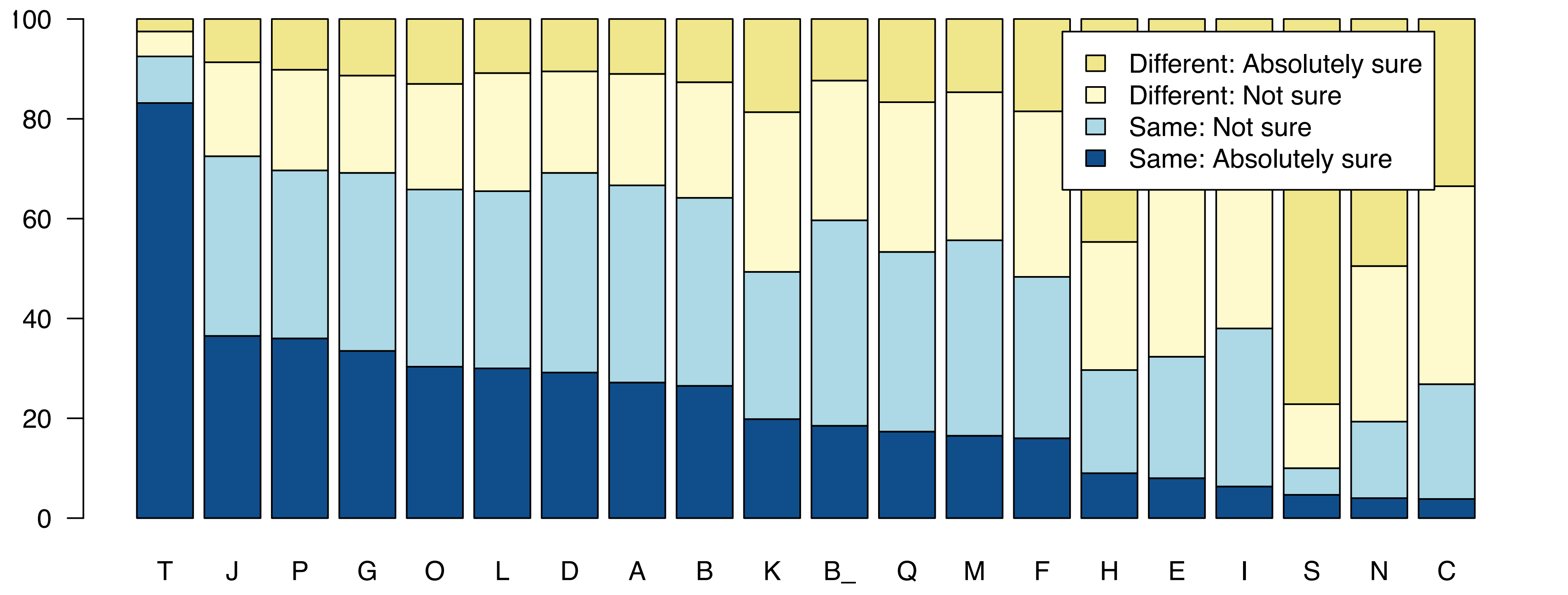

MM
5
4
3
2
1
S T N K O L J P A H F Q B G E D M B_ I C
FF
T S N J K L P Q G F B O H E A D M I B_ C
MF
S T K G L O F J P N A E B D Q M H B_ C I
FM
S T O K N J G L B P F A Q E D M H I B_ C


FF
TS
N
J K L P Q G F B O H E A D M
I
B_
C
MM
ST
NK
O L J P A H F Q B G E D M
B_ I
C
FM
ST
O K N J G L B P F A Q E D M H
I B_ C
MF
ST
K G L O F J P N A E B D Q M
H
B_ C I

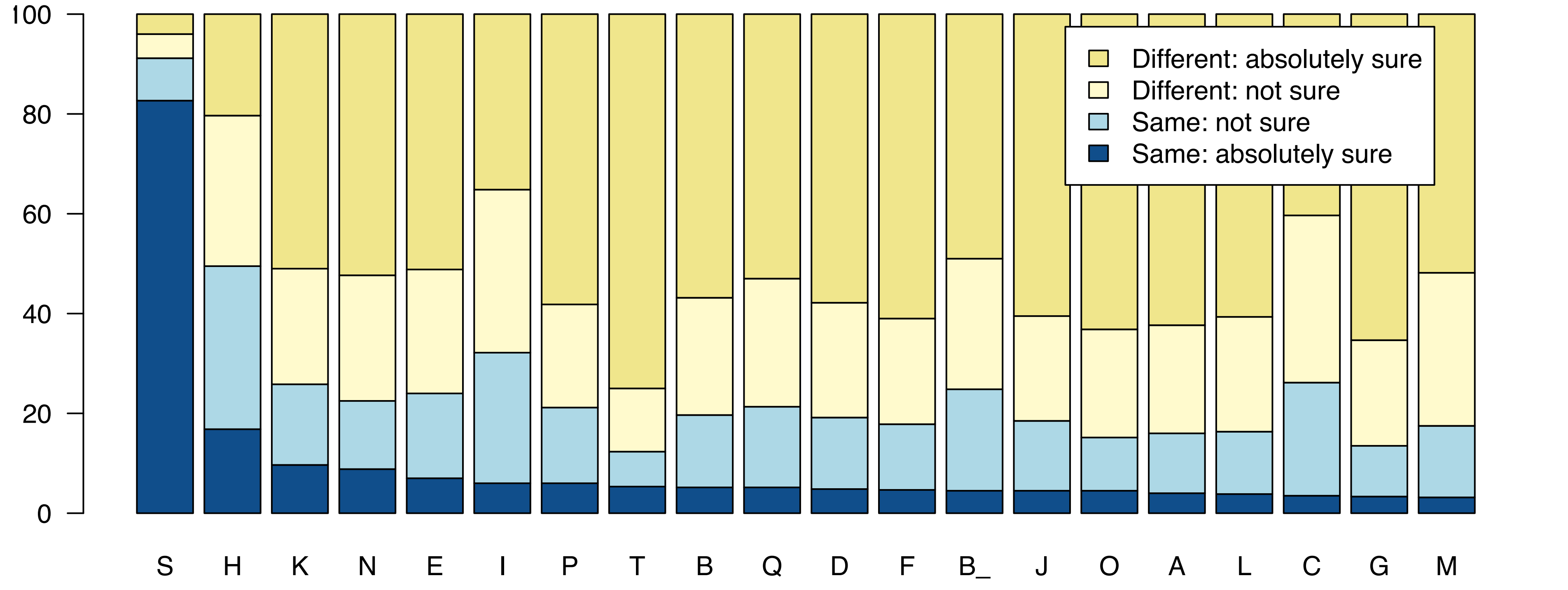
Results

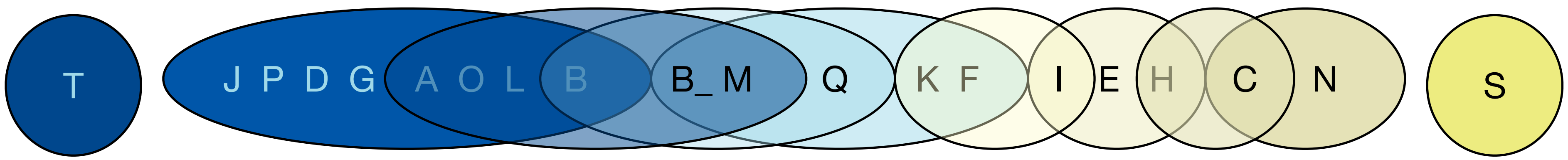
- Similarity: Same-Different

Target



Source





VCC - evaluation

- Such a large evaluation complex, compromises inevitable.
- Two sets of source-target pairs for naturalness ratings not ideal.
- Including comparisons to source as well as target was informative.

VCC data set

- Database (training and test samples)
- Participants’ submissions
- Listening test materials
- Available at:

http://dx.doi.org/10.7488/ds/1430