Le Challenge Blizzard 2023

Olivier Perrotin, Brooke Stephenson, Silvain Gerber, Gérard Bailly

29 / 08 / 2023

The 18th Blizzard Challenge Workshop

Program

Morning

- **09:00 - 10:15 | Summary of the Blizzard Challenge 2023**
- **10:15 - 10:30 | Coffee break**
- **10:30 - 11:45 | System presentations: *FastSpeech-based models***
  - LIUM-TTS - *Laboratoire d'Informatique Le Mans Université (LIUM)*
  - GIPSA-lab - *Univ. Grenoble Alpes, CNRS, Grenoble INP, France*
  - IMS - *University of Stuttgart, Institute for Natural Language Processing, Germany*
  - MuLanTTS - *Microsoft*
  - Samsung TTS - *Samsung Electronics HQ and Samsung Research China, Beijing*
- **11:45 - 12:00 | System presentations: *FastSpeech- and Tacotron-based models***
  - SCUT SCSE (remote) - *South China University of Technology*
  - FireRedTTS (remote) - *Xiaohongshu Inc.*
- **12:00 - 13:30 | Lunch at the venue**

Program

Afternoon

- **13:30 - 14:30 | System presentations: *Tacotron-based models***
  - AudioLabs - *International Audio Laboratories Erlangen*
  - TTS-Cube - *Adobe Systems, SCC*
  - La Forge - *Ubisoft*
  - DeepZen - *DeepZen Ltd.*
- **14:30 - 15:00 | Coffee break**
- **15:00 - 16:00 | System presentations: *Stochastic models***
  - Idiap - *Idiap Research Institute, Martigny, Switzerland*
  - BIGAI - *Beijing Institute of General Artificial Intelligence*
  - CASIA Speech (remote) - *Institute of Automation, Chinese Academy of Sciences*
  - Xiaomi-ASLP (remote) - *Xiaomi AI Lab and Audio Speech and Language Processing Group (ASLP@NPU), Northwestern Polytechnical University*
  - Fruit Shell (remote) - *University of Chinese Academy of Sciences*
  - 10AI (remote) - *Beijing Yiling Intelligence Technology Co., Ltd.*
  - IOA-ThinkIT (remote) - *Institute of Acoustics of the Chinese Academy of Sciences*
- **16:00 - 16:30 | Conclusion and discussions about future Challenges**

What is the Blizzard Challenge?

- Goal
  - Better understand and compare techniques in building corpus-based speech synthesisers
- Method
  - Build voices on a common dataset
  - Evaluate them in a single listening test
- The Blizzard Challenge 2023 is the 18th Blizzard Challenge
  - French TTS
  - Data from both publicly available audiobooks and internal recordings

Blizzard Challenge Timeline

	Europe / America	Asia
2005 ...	...	
2013	US English *(Audiobooks)*	Indian Languages (Wikipedia)
2014		Indian Languages (Wikipedia)
2015	UK English *(Children's audiobooks)*	Indian Languages (Wikipedia)
2016	UK English *(Children's audiobooks)*	
2017	UK English *(Children's audiobooks)*	
2018	UK English *(Children's audiobooks)*	
2019		Mandarin *(Spontaneous speech)*
2020		Mandarin / Shanghainese *(Read daily news)*
2021	Spanish *(Dialogue, daily life, etc.)*	
2022		
2023	French *(Audiobooks and parliament transcripts)*	
2024	To be decided	
2025	...	
2026		

Outline of the presentation

An overview of all the aspects of the challenge

- Data
- Tasks
- Participants
- Listening test design
- Analysis methodology
- Results

Data

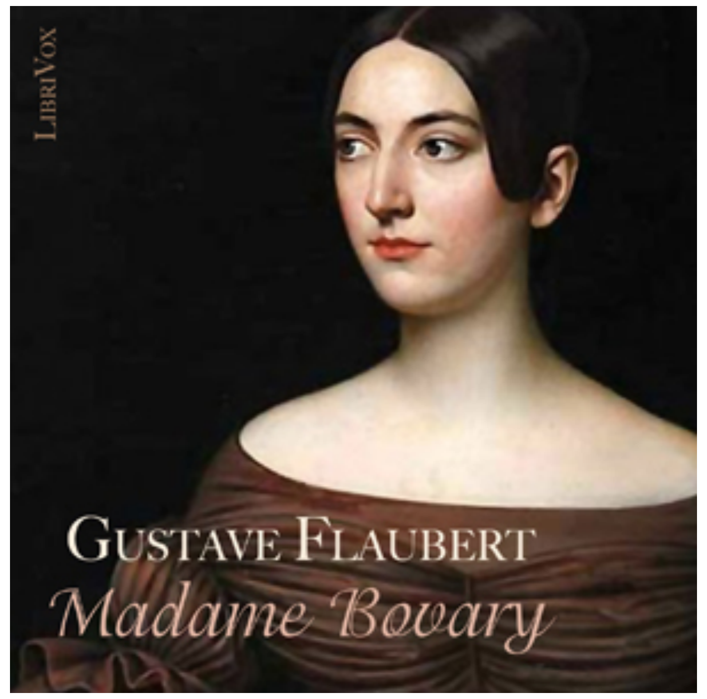
French audiobooks and non-fiction readings

Two datasets for two tasks

Hub dataset

Single French female speaker (large corpus) - Nadine Eckert-Boulet (NEB)

- Audiobooks from LibriVox **J. Kearns (2014), Reference Reviews 28(1)**
  - 51-hour speech material



- Text processing and segmentation
  - Orthographic transcriptions from the Gutenberg project ; all texts spelled out
  - Annotation of paragraphs ; segmentation based on silences of at least 400 ms **M. Lenglet et al. (2021), SSW**

- Annotation
  - 2/3 of the corpus is semi-automatically aligned with phonetic transcription

Two datasets for two tasks

Spoke dataset

Single French female speaker (small corpus) - Aurélie Derbier (AD)

- Text from the SIWIS database **P.-E. Honnet et al. (2017), Idiap Tech. Rep.**
  - French novel / French parliamentary debates transcripts
  - 2-hour speech material
- Text processing and segmentation
  - Orthographic transcriptions ; all texts spelled out
  - Annotation of paragraphs ; segmentation based on silences of at least 400 ms
  - Full audio recording sequences are provided, including in-between utterances
- Annotation
  - The full corpus is semi-automatically aligned with phonetic transcription

Two datasets for two tasks

➡ Both datasets are publicly available on:
https://zenodo.org/record/7560290

(The link is referenced on the Blizzard Challenge website)

Tasks and rules

French TTS and Speaker adaptation

Two tasks

Tasks

- Hub task 2023-FH1 - *French TTS*
  - Build a voice from the provided French data (NEB), using only publicly available data
- Spoke task 2023-FS1 - *Speaker adaptation*
  - Build a voice from the provided French data (AD) that is the closest to AD as possible

Two tasks

Tasks

- Hub task 2023-FH1 - *French TTS*
  - Build a voice from the provided French data (NEB), using only publicly available data
- Spoke task 2023-FS1 - *Speaker adaptation*
  - Build a voice from the provided French data (AD) that is the closest to AD as possible

Reproducibility requirements

- Definitions
  - **"External data"** is defined as data, of any type, that is not part of the provided database.
  - **"External model"** is defined as a model, of any type, that has not been trained by the team (e.g., pre-trained wav2vec, BERT, etc.).
- Reproducibility criteria
  1. Used external models are **publicly-available off-the-shelf pre-trained models**, and references are given
  2. Any audio data used for training models (including for fine-tuning pre-trained models) is **publicly available** and reported
  3. Source code is **provided**

Two tasks

Tasks

- Hub task 2023-FH1 - *French TTS*
  - Build a voice from the provided French data (NEB), using only publicly available data
  - **Reproducibility criteria 1 and 2**
- Spoke task 2023-FS1 - *Speaker adaptation*
  - Build a voice from the provided French data (AD) that is the closest to AD as possible
  - **No reproducibility criteria**
- **Reproducibility criterion 3 encouraged for all tasks**

Reproducibility requirements

- Definitions
  - "External data" is defined as data, of any type, that is not part of the provided database.
  - "External model" is defined as a model, of any type, that has not been trained by the team (e.g., pre-trained wav2vec, BERT, etc.).
- Reproducibility criteria
  1. Used external models are **publicly-available off-the-shelf pre-trained models**, and references are given
  2. Any audio data used for training models (including for fine-tuning pre-trained models) is **publicly available** and reported
  3. Source code is **provided**

Additional rules

- Use of external data and external models is entirely optional and is not compulsory
- You must use the provided audio files
- You must not use any additional speech data from the same speakers
- You may exclude any parts of the provided databases if you wish
- There is no limitation on the amount of external non-audio data you may use (e.g., text, dictionaries)
- Use of any provided transcriptions is optional.
- If you are in any doubt about how to apply these rules, please contact the organisers for clarification.

Participants

2 benchmarks and 18 teams

Benchmark systems

BT: *Tacotron2 baseline*

- Acoustic model
  - Tacotron2 - NVIDIA implementation **J. Shen et al. (2018), ICASSP**
  - Trained from scratch on the full Hub dataset for FH1 (158.5k training steps)
  - Fine-tuned on the Spoke dataset for FS1 (57.5k steps from the 100k checkpoint)
  - Hyper parameters from the implementation
- Vocoder
  - HiFi-GAN **J. Kong et al. (2020), NIPS**
  - Pre-trained UNIVERSAL model provided
- Text input
  - Orthographic characters
  - Preprocessed with the transliteration cleaner provided with the implementation

Benchmark systems

BT: *Tacotron2 baseline*

- Acoustic model
  - Tacotron2 - NVIDIA implementation J. Shen et al. (2018), ICASSP
  - Trained from scratch on the full Hub dataset for FH1 (158.5k training steps)
  - Fine-tuned on the Spoke dataset for FS1 (57.5k steps from the 100k checkpoint)
  - Hyper parameters from the implementation
- Vocoder
  - HiFi-GAN J. Kong et al. (2020), NIPS
  - Pre-trained UNIVERSAL model provided
- Text input
  - Orthographic characters
  - Preprocessed with the transliteration cleaner provided with the implementation

BF: *FastSpeech2 baseline*

- Acoustic model
  - FastSpeech2 - FairSeq implementation Y. Shen et al. (2021), ICLR
  - Trained from scratch on the **annotated** Hub dataset (333.9k training steps)
  - Fine-tuned on the Spoke dataset for FS1 (7.25k steps from the last checkpoint)
  - Hyper parameters from the implementation
- Vocoder
  - HiFi-GAN
  - Pre-trained UNIVERSAL model provided
- Text input
  - Phonetic characters
  - L2S with eSpeak while keeping punctuations

Participants

	Team	Affiliation	Country	L2S	Prosody control (inference)	Acoustic model	Vocoder	LLM
BF	FastSpeech benchmark			eSpeak	Variance predictors from text	FastSpeech2	HiFi-GAN	
BT	Tacotron benchmark			/		Tacotron2	HiFi-GAN	
	LIUM-TTS	Laboratoire d'Informatique Le Mans Université	FR	Data-driven L2S	Variance predictors from text	FastSpeech2 (TTS) + WavLM-Tacotron2 (VC)	WaveGlow	
	GIPSA-lab	Univ. Grenoble Alpes, CNRS, Grenoble INP	FR	Phonetic prediction task in encoder	Variance predictors from text	FastSpeech2-based	WaveGlow	
	SCUT SCSE	South China University of Technology	CN	eSpeak	Prosody predictor (VQ-VAE) from FlauBERT Variance predictors from text	FastSpeech2-based	HiFi-GAN	✔
	IMS (Toucan)	University of Stuttgart, Institute for Natural Language Processing	DE	eSpeak + CamemBERT (POS)	Prosody predictor (GST) from input Variance predictors from text + GST	FastSpeech2-based (conformers)	BigVGAN	✔
	MuLanTTS	Microsoft	CN	Own L2S + BERT (liaisons and homographs)	Prosody predictor (GST) from text Variance predictors from text	FastSpeech2-based (conformers)	HiFi-GAN	✔
	Samsung TTS	Samsung Electronics HQ and Samsung Research China, Beijing	KR	CART + CamemBERT (breaks, liaisons, POS) + ChatGPT (some homographs)	Prosody predictor (GST/VAE) from text + CamemBERT + Speech type	FastSpeech2-based (conformers)	HiFi-GAN	✔
	AudioLabs	International Audio Laboratories Erlangen	DE	Lexicons + eSpeak	Variance predictors from text	Forward Tacotron / FastTacotron	StyleMelGAN	
	TTS-Cube	Adobe Systems, SCC	RO	Data-driven L2S	Variance predictors from text + CamemBERT	RNN-based	(HiFi-GAN)	✔
	La Forge	Ubisoft	CA	eSpeak + CamemBERT (POS)	Prosody predictor (VAE) from text	VAE-Tacotron	HiFi-GAN	✔
	FireRedTTS	Xiaohongshu Inc.	CN	Lexicon + CamemBERT (POS, DEP)	Prosody predictor (RNN) from text Rhythmic rules predictor from POS, NER, DEP	Non-attentive Tacotron	HiFi++	✔
	DeepZen	DeepZen Ltd.	GB	Lexicons + FlauBERT (POS)	Prosody predictor (GST/LST) from FlauBERT	Non-attentive Tacotron	HiFi-GAN-based	✔
	CASIA Speech (VIBVG)	Institute of Automation, Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(BigVGAN)	
	Fruit shell 2023	University of Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	BIGAI	Beijing Institute of General Artificial Intelligence	CN	eSpeak + pBART	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	Xiaomi-ASLP	Xiaomi AI Lab and Audio Speech and Language Processing Group (ASLP@NPU), Northwestern Polytechnical University	CN	eSpeak	Prosody predictor (Flow) from text + GPT-3	VITS	(HiFi-GAN)	✔
	10AI (Xpress)	Beijing Yiling Intelligence Technology Co., Ltd.	CN	/	Prosody predictor (Flow) from text	Flow-VAE	BigVGAN	
	IOA-ThinkIT	Institute of Acoustics of the Chinese Academy of Sciences	CN	Own L2S + BERT (word embeding)	Prosody predictor (H-VAE) from text	Hierarchical VAE	/	✔
	Idiap	Idiap Research Institute, Martigny	CH	eSpeak + CamemBERT (POS)	Variance predictors from text	Diffusion transformer	FastDiff	✔

Participants

FastSpeech-style
Tacotron-style
Stochastic models

	Team	Affiliation	Country	L2S	Prosody control (inference)	Acoustic model	Vocoder	LLM
BF	FastSpeech benchmark			eSpeak	Variance predictors from text	FastSpeech2	HiFi-GAN	
BT	Tacotron benchmark			/		Tacotron2	HiFi-GAN	
	LIUM-TTS	Laboratoire d'Informatique Le Mans Université	FR	Data-driven L2S	Variance predictors from text	FastSpeech2 (TTS) + WavLM-Tacotron2 (VC)	WaveGlow	
	GIPSA-lab	Univ. Grenoble Alpes, CNRS, Grenoble INP	FR	Phonetic prediction task in encoder	Variance predictors from text	FastSpeech2-based	WaveGlow	
	SCUT SCSE	South China University of Technology	CN	eSpeak	Prosody predictor (VQ-VAE) from FlauBERT Variance predictors from text	FastSpeech2-based	HiFi-GAN	✔
	IMS (Toucan)	University of Stuttgart, Institute for Natural Language Processing	DE	eSpeak + CamemBERT (POS)	Prosody predictor (GST) from input Variance predictors from text + GST	FastSpeech2-based (conformers)	BigVGAN	✔
	MuLanTTS	Microsoft	CN	Own L2S + BERT (liaisons and homographs)	Prosody predictor (GST) from text Variance predictors from text	FastSpeech2-based (conformers)	HiFi-GAN	✔
	Samsung TTS	Samsung Electronics HQ and Samsung Research China, Beijing	KR	CART + CamemBERT (breaks, liaisons, POS) + ChatGPT (some homographs)	Prosody predictor (GST/VAE) from text + CamemBERT + Speech type	FastSpeech2-based (conformers)	HiFi-GAN	✔
	AudioLabs	International Audio Laboratories Erlangen	DE	Lexicons + eSpeak	Variance predictors from text	Forward Tacotron / FastTacotron	StyleMelGAN	
	TTS-Cube	Adobe Systems, SCC	RO	Data-driven L2S	Variance predictors from text + CamemBERT	RNN-based	(HiFi-GAN)	✔
	La Forge	Ubisoft	CA	eSpeak + CamemBERT (POS)	Prosody predictor (VAE) from text	VAE-Tacotron	HiFi-GAN	✔
	FireRedTTS	Xiaohongshu Inc.	CN	Lexicon + CamemBERT (POS, DEP)	Prosody predictor (RNN) from text Rhythmic rules predictor from POS, NER, DEP	Non-attentive Tacotron	HiFi++	✔
	DeepZen	DeepZen Ltd.	GB	Lexicons + FlauBERT (POS)	Prosody predictor (GST/LST) from FlauBERT	Non-attentive Tacotron	HiFi-GAN-based	✔
	CASIA Speech (VIBVG)	Institute of Automation, Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(BigVGAN)	
	Fruit shell 2023	University of Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	BIGAI	Beijing Institute of General Artificial Intelligence	CN	eSpeak + pBART	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	Xiaomi-ASLP	Xiaomi AI Lab and Audio Speech and Language Processing Group (ASLP@NPU), Northwestern Polytechnical University	CN	eSpeak	Prosody predictor (Flow) from text + GPT-3	VITS	(HiFi-GAN)	✔
	10AI (Xpress)	Beijing Yiling Intelligence Technology Co., Ltd.	CN	/	Prosody predictor (Flow) from text	Flow-VAE	BigVGAN	
	IOA-ThinkIT	Institute of Acoustics of the Chinese Academy of Sciences	CN	Own L2S + BERT (word embeding)	Prosody predictor (H-VAE) from text	Hierarchical VAE	/	✔
	Idiap	Idiap Research Institute, Martigny	CH	eSpeak + CamemBERT (POS)	Variance predictors from text	Diffusion transformer	FastDiff	✔

Participants

Flow
GAN-based

	Team	Affiliation	Country	L2S	Prosody control (inference)	Acoustic model	Vocoder	LLM
BF	FastSpeech benchmark			eSpeak	Variance predictors from text	FastSpeech2	HiFi-GAN	
BT	Tacotron benchmark			/		Tacotron2	HiFi-GAN	
	LIUM-TTS	Laboratoire d'Informatique Le Mans Université	FR	Data-driven L2S	Variance predictors from text	FastSpeech2 (TTS) + WavLM-Tacotron2 (VC)	WaveGlow	
	GIPSA-lab	Univ. Grenoble Alpes, CNRS, Grenoble INP	FR	Phonetic prediction task in encoder	Variance predictors from text	FastSpeech2-based	WaveGlow	
	SCUT SCSE	South China University of Technology	CN	eSpeak	Prosody predictor (VQ-VAE) from FlauBERT Variance predictors from text	FastSpeech2-based	HiFi-GAN	✔
	IMS (Toucan)	University of Stuttgart, Institute for Natural Language Processing	DE	eSpeak + CamemBERT (POS)	Prosody predictor (GST) from input Variance predictors from text + GST	FastSpeech2-based (conformers)	BigVGAN	✔
	MuLanTTS	Microsoft	CN	Own L2S + BERT (liaisons and homographs)	Prosody predictor (GST) from text Variance predictors from text	FastSpeech2-based (conformers)	HiFi-GAN	✔
	Samsung TTS	Samsung Electronics HQ and Samsung Research China, Beijing	KR	CART + CamemBERT (breaks, liaisons, POS) + ChatGPT (some homographs)	Prosody predictor (GST/VAE) from text + CamemBERT + Speech type	FastSpeech2-based (conformers)	HiFi-GAN	✔
	AudioLabs	International Audio Laboratories Erlangen	DE	Lexicons + eSpeak	Variance predictors from text	Forward Tacotron / FastTacotron	StyleMelGAN	
	TTS-Cube	Adobe Systems, SCC	RO	Data-driven L2S	Variance predictors from text + CamemBERT	RNN-based	(HiFi-GAN)	✔
	La Forge	Ubisoft	CA	eSpeak + CamemBERT (POS)	Prosody predictor (VAE) from text	VAE-Tacotron	HiFi-GAN	✔
	FireRedTTS	Xiaohongshu Inc.	CN	Lexicon + CamemBERT (POS, DEP)	Prosody predictor (RNN) from text Rhythmic rules predictor from POS, NER, DEP	Non-attentive Tacotron	HiFi++	✔
	DeepZen	DeepZen Ltd.	GB	Lexicons + FlauBERT (POS)	Prosody predictor (GST/LST) from FlauBERT	Non-attentive Tacotron	HiFi-GAN-based	✔
	CASIA Speech (VIBVG)	Institute of Automation, Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(BigVGAN)	
	Fruit shell 2023	University of Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	BIGAI	Beijing Institute of General Artificial Intelligence	CN	eSpeak + pBART	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	Xiaomi-ASLP	Xiaomi AI Lab and Audio Speech and Language Processing Group (ASLP@NPU), Northwestern Polytechnical University	CN	eSpeak	Prosody predictor (Flow) from text + GPT-3	VITS	(HiFi-GAN)	✔
	10AI (Xpress)	Beijing Yiling Intelligence Technology Co., Ltd.	CN	/	Prosody predictor (Flow) from text	Flow-VAE	BigVGAN	
	IOA-ThinkIT	Institute of Acoustics of the Chinese Academy of Sciences	CN	Own L2S + BERT (word embeding)	Prosody predictor (H-VAE) from text	Hierarchical VAE	/	✔
	Idiap	Idiap Research Institute, Martigny	CH	eSpeak + CamemBERT (POS)	Variance predictors from text	Diffusion transformer	FastDiff	✔

gipsa-lab cnrs INP UGA Université Grenoble Alpes

Participants

Variance predictor from text / LLM
Prosody predictor (Flow, VAE, GST) from text / LLM
Both

	Team	Affiliation	Country	L2S	Prosody control (inference)	Acoustic model	Vocoder	LLM
BF	FastSpeech benchmark			eSpeak	Variance predictors from text	FastSpeech2	HiFi-GAN	
BT	Tacotron benchmark			/		Tacotron2	HiFi-GAN	
	LIUM-TTS	Laboratoire d'Informatique Le Mans Université	FR	Data-driven L2S	Variance predictors from text	FastSpeech2 (TTS) + WavLM-Tacotron2 (VC)	WaveGlow	
	GIPSA-lab	Univ. Grenoble Alpes, CNRS, Grenoble INP	FR	Phonetic prediction task in encoder	Variance predictors from text	FastSpeech2-based	WaveGlow	
	SCUT SCSE	South China University of Technology	CN	eSpeak	Prosody predictor (VQ-VAE) from FlauBERT Variance predictors from text	FastSpeech2-based	HiFi-GAN	✔
	IMS (Toucan)	University of Stuttgart, Institute for Natural Language Processing	DE	eSpeak + CamemBERT (POS)	Prosody predictor (GST) from input Variance predictors from text + GST	FastSpeech2-based (conformers)	BigVGAN	✔
	MuLanTTS	Microsoft	CN	Own L2S + BERT (liaisons and homographs)	Prosody predictor (GST) from text Variance predictors from text	FastSpeech2-based (conformers)	HiFi-GAN	✔
	Samsung TTS	Samsung Electronics HQ and Samsung Research China, Beijing	KR	CART + CamemBERT (breaks, liaisons, POS) + ChatGPT (some homographs)	Prosody predictor (GST/VAE) from text + CamemBERT + Speech type	FastSpeech2-based (conformers)	HiFi-GAN	✔
	AudioLabs	International Audio Laboratories Erlangen	DE	Lexicons + eSpeak	Variance predictors from text	Forward Tacotron / FastTacotron	StyleMelGAN	
	TTS-Cube	Adobe Systems, SCC	RO	Data-driven L2S	Variance predictors from text + CamemBERT	RNN-based	(HiFi-GAN)	✔
	La Forge	Ubisoft	CA	eSpeak + CamemBERT (POS)	Prosody predictor (VAE) from text	VAE-Tacotron	HiFi-GAN	✔
	FireRedTTS	Xiaohongshu Inc.	CN	Lexicon + CamemBERT (POS, DEP)	Prosody predictor (RNN) from text Rhythmic rules predictor from POS, NER, DEP	Non-attentive Tacotron	HiFi++	✔
	DeepZen	DeepZen Ltd.	GB	Lexicons + FlauBERT (POS)	Prosody predictor (GST/LST) from FlauBERT	Non-attentive Tacotron	HiFi-GAN-based	✔
	CASIA Speech (VIBVG)	Institute of Automation, Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(BigVGAN)	
	Fruit shell 2023	University of Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	BIGAI	Beijing Institute of General Artificial Intelligence	CN	eSpeak + pBART	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	Xiaomi-ASLP	Xiaomi AI Lab and Audio Speech and Language Processing Group (ASLP@NPU), Northwestern Polytechnical University	CN	eSpeak	Prosody predictor (Flow) from text + GPT-3	VITS	(HiFi-GAN)	✔
	10AI (Xpress)	Beijing Yiling Intelligence Technology Co., Ltd.	CN	/	Prosody predictor (Flow) from text	Flow-VAE	BigVGAN	
	IOA-ThinkIT	Institute of Acoustics of the Chinese Academy of Sciences	CN	Own L2S + BERT (word embeding)	Prosody predictor (H-VAE) from text	Hierarchical VAE	/	✔
	Idiap	Idiap Research Institute, Martigny	CH	eSpeak + CamemBERT (POS)	Variance predictors from text	Diffusion transformer	FastDiff	✔

Participants

- eSpeak
- Re-training
- Use of a large language model

	Team	Affiliation	Country	L2S	Prosody control (inference)	Acoustic model	Vocoder	LLM
BF	FastSpeech benchmark			eSpeak	Variance predictors from text	FastSpeech2	HiFi-GAN	
BT	Tacotron benchmark			/		Tacotron2	HiFi-GAN	
	LIUM-TTS	Laboratoire d'Informatique Le Mans Université	FR	Data-driven L2S	Variance predictors from text	FastSpeech2 (TTS) + WavLM-Tacotron2 (VC)	WaveGlow	
	GIPSA-lab	Univ. Grenoble Alpes, CNRS, Grenoble INP	FR	Phonetic prediction task in encoder	Variance predictors from text	FastSpeech2-based	WaveGlow	
	SCUT SCSE	South China University of Technology	CN	eSpeak	Prosody predictor (VQ-VAE) from FlauBERT Variance predictors from text	FastSpeech2-based	HiFi-GAN	✔
	IMS (Toucan)	University of Stuttgart, Institute for Natural Language Processing	DE	eSpeak + CamemBERT (POS)	Prosody predictor (GST) from input Variance predictors from text + GST	FastSpeech2-based (conformers)	BigVGAN	✔
	MuLanTTS	Microsoft	CN	Own L2S + BERT (liaisons and homographs)	Prosody predictor (GST) from text Variance predictors from text	FastSpeech2-based (conformers)	HiFi-GAN	✔
	Samsung TTS	Samsung Electronics HQ and Samsung Research China, Beijing	KR	CART + CamemBERT (breaks, liaisons, POS) + ChatGPT (some homographs)	Prosody predictor (GST/VAE) from text + CamemBERT + Speech type	FastSpeech2-based (conformers)	HiFi-GAN	✔
	AudioLabs	International Audio Laboratories Erlangen	DE	Lexicons + eSpeak	Variance predictors from text	Forward Tacotron / FastTacotron	StyleMelGAN	
	TTS-Cube	Adobe Systems, SCC	RO	Data-driven L2S	Variance predictors from text + CamemBERT	RNN-based	(HiFi-GAN)	✔
	La Forge	Ubisoft	CA	eSpeak + CamemBERT (POS)	Prosody predictor (VAE) from text	VAE-Tacotron	HiFi-GAN	✔
	FireRedTTS	Xiaohongshu Inc.	CN	Lexicon + CamemBERT (POS, DEP)	Prosody predictor (RNN) from text Rhythmic rules predictor from POS, NER, DEP	Non-attentive Tacotron	HiFi++	✔
	DeepZen	DeepZen Ltd.	GB	Lexicons + FlauBERT (POS)	Prosody predictor (GST/LST) from FlauBERT	Non-attentive Tacotron	HiFi-GAN-based	✔
	CASIA Speech (VIBVG)	Institute of Automation, Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(BigVGAN)	
	Fruit shell 2023	University of Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	BIGAI	Beijing Institute of General Artificial Intelligence	CN	eSpeak + pBART	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	Xiaomi-ASLP	Xiaomi AI Lab and Audio Speech and Language Processing Group (ASLP@NPU), Northwestern Polytechnical University	CN	eSpeak	Prosody predictor (Flow) from text + GPT-3	VITS	(HiFi-GAN)	✔
	10AI (Xpress)	Beijing Yiling Intelligence Technology Co., Ltd.	CN	/	Prosody predictor (Flow) from text	Flow-VAE	BigVGAN	
	IOA-ThinkIT	Institute of Acoustics of the Chinese Academy of Sciences	CN	Own L2S + BERT (word embeding)	Prosody predictor (H-VAE) from text	Hierarchical VAE	/	✔
	Idiap	Idiap Research Institute, Martigny	CH	eSpeak + CamemBERT (POS)	Variance predictors from text	Diffusion transformer	FastDiff	✔

Participants

Text input
Style input
Both

	Team	Affiliation	Country	L2S	Prosody control (inference)	Acoustic model	Vocoder	LLM
BF	FastSpeech benchmark			eSpeak	Variance predictors from text	FastSpeech2	HiFi-GAN	
BT	Tacotron benchmark			/		Tacotron2	HiFi-GAN	
	LIUM-TTS	Laboratoire d'Informatique Le Mans Université	FR	Data-driven L2S	Variance predictors from text	FastSpeech2 (TTS) + WavLM-Tacotron2 (VC)	WaveGlow	
	GIPSA-lab	Univ. Grenoble Alpes, CNRS, Grenoble INP	FR	Phonetic prediction task in encoder	Variance predictors from text	FastSpeech2-based	WaveGlow	
	SCUT SCSE	South China University of Technology	CN	eSpeak	Prosody predictor (VQ-VAE) from FlauBERT Variance predictors from text	FastSpeech2-based	HiFi-GAN	✔
	IMS (Toucan)	University of Stuttgart, Institute for Natural Language Processing	DE	eSpeak + CamemBERT (POS)	Prosody predictor (GST) from input Variance predictors from text + GST	FastSpeech2-based (conformers)	BigVGAN	✔
	MuLanTTS	Microsoft	CN	Own L2S + BERT (liaisons and homographs)	Prosody predictor (GST) from text Variance predictors from text	FastSpeech2-based (conformers)	HiFi-GAN	✔
	Samsung TTS	Samsung Electronics HQ and Samsung Research China, Beijing	KR	CART + CamemBERT (breaks, liaisons, POS) + ChatGPT (some homographs)	Prosody predictor (GST/VAE) from text + CamemBERT + Speech type	FastSpeech2-based (conformers)	HiFi-GAN	✔
	AudioLabs	International Audio Laboratories Erlangen	DE	Lexicons + eSpeak	Variance predictors from text	Forward Tacotron / FastTacotron	StyleMelGAN	
	TTS-Cube	Adobe Systems, SCC	RO	Data-driven L2S	Variance predictors from text + CamemBERT	RNN-based	(HiFi-GAN)	✔
	La Forge	Ubisoft	CA	eSpeak + CamemBERT (POS)	Prosody predictor (VAE) from text	VAE-Tacotron	HiFi-GAN	✔
	FireRedTTS	Xiaohongshu Inc.	CN	Lexicon + CamemBERT (POS, DEP)	Prosody predictor (RNN) from text Rhythmic rules predictor from POS, NER, DEP	Non-attentive Tacotron	HiFi++	✔
	DeepZen	DeepZen Ltd.	GB	Lexicons + FlauBERT (POS)	Prosody predictor (GST/LST) from FlauBERT	Non-attentive Tacotron	HiFi-GAN-based	✔
	CASIA Speech (VIBVG)	Institute of Automation, Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(BigVGAN)	
	Fruit shell 2023	University of Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	BIGAI	Beijing Institute of General Artificial Intelligence	CN	eSpeak + pBART	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	Xiaomi-ASLP	Xiaomi AI Lab and Audio Speech and Language Processing Group (ASLP@NPU), Northwestern Polytechnical University	CN	eSpeak	Prosody predictor (Flow) from text + GPT-3	VITS	(HiFi-GAN)	✔
	10AI (Xpress)	Beijing Yiling Intelligence Technology Co., Ltd.	CN	/	Prosody predictor (Flow) from text	Flow-VAE	BigVGAN	
	IOA-ThinkIT	Institute of Acoustics of the Chinese Academy of Sciences	CN	Own L2S + BERT (word embeding)	Prosody predictor (H-VAE) from text	Hierarchical VAE	/	✔
	Idiap	Idiap Research Institute, Martigny	CH	eSpeak + CamemBERT (POS)	Variance predictors from text	Diffusion transformer	FastDiff	✔

Participants

A	Natural speech

Skip audio

	Team	Affiliation	Country	L2S	Prosody control (inference)	Acoustic model	Vocoder	LLM
BF	FastSpeech benchmark			eSpeak	Variance predictors from text	FastSpeech2	HiFi-GAN	
BT	Tacotron benchmark			/		Tacotron2	HiFi-GAN	
	LIUM-TTS	Laboratoire d'Informatique Le Mans Université	FR	Data-driven L2S	Variance predictors from text	FastSpeech2 (TTS) + WavLM-Tacotron2 (VC)	WaveGlow	
	GIPSA-lab	Univ. Grenoble Alpes, CNRS, Grenoble INP	FR	Phonetic prediction task in encoder	Variance predictors from text	FastSpeech2-based	WaveGlow	
	SCUT SCSE	South China University of Technology	CN	eSpeak	Prosody predictor (VQ-VAE) from FlauBERT Variance predictors from text	FastSpeech2-based	HiFi-GAN	✔
	IMS (Toucan)	University of Stuttgart, Institute for Natural Language Processing	DE	eSpeak + CamemBERT (POS)	Prosody predictor (GST) from input Variance predictors from text + GST	FastSpeech2-based (conformers)	BigVGAN	✔
	MuLanTTS	Microsoft	CN	Own L2S + BERT (liaisons and homographs)	Prosody predictor (GST) from text Variance predictors from text	FastSpeech2-based (conformers)	HiFi-GAN	✔
	Samsung TTS	Samsung Electronics HQ and Samsung Research China, Beijing	KR	CART + CamemBERT (breaks, liaisons, POS) + ChatGPT (some homographs)	Prosody predictor (GST/VAE) from text + CamemBERT + Speech type	FastSpeech2-based (conformers)	HiFi-GAN	✔
	AudioLabs	International Audio Laboratories Erlangen	DE	Lexicons + eSpeak	Variance predictors from text	Forward Tacotron / FastTacotron	StyleMelGAN	
	TTS-Cube	Adobe Systems, SCC	RO	Data-driven L2S	Variance predictors from text + CamemBERT	RNN-based	(HiFi-GAN)	✔
	La Forge	Ubisoft	CA	eSpeak + CamemBERT (POS)	Prosody predictor (VAE) from text	VAE-Tacotron	HiFi-GAN	✔
	FireRedTTS	Xiaohongshu Inc.	CN	Lexicon + CamemBERT (POS, DEP)	Prosody predictor (RNN) from text Rhythmic rules predictor from POS, NER, DEP	Non-attentive Tacotron	HiFi++	✔
	DeepZen	DeepZen Ltd.	GB	Lexicons + FlauBERT (POS)	Prosody predictor (GST/LST) from FlauBERT	Non-attentive Tacotron	HiFi-GAN-based	✔
	CASIA Speech (VIBVG)	Institute of Automation, Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(BigVGAN)	
	Fruit shell 2023	University of Chinese Academy of Sciences	CN	eSpeak	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	BIGAI	Beijing Institute of General Artificial Intelligence	CN	eSpeak + pBART	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
	Xiaomi-ASLP	Xiaomi AI Lab and Audio Speech and Language Processing Group (ASLP@NPU), Northwestern Polytechnical University	CN	eSpeak	Prosody predictor (Flow) from text + GPT-3	VITS	(HiFi-GAN)	✔
	10AI (Xpress)	Beijing Yiling Intelligence Technology Co., Ltd.	CN	/	Prosody predictor (Flow) from text	Flow-VAE	BigVGAN	
	IOA-ThinkIT	Institute of Acoustics of the Chinese Academy of Sciences	CN	Own L2S + BERT (word embeding)	Prosody predictor (H-VAE) from text	Hierarchical VAE	/	✔
	Idiap	Idiap Research Institute, Martigny	CH	eSpeak + CamemBERT (POS)	Variance predictors from text	Diffusion transformer	FastDiff	✔

Participants

Task completion

Tasks

- Hub task 2023-FH1 - *French TTS*
  - **18 participants**
- Spoke task 2023-FS1 - *Speaker adaptation*
  - **14 participants**

Participants

Task completion

Tasks

- Hub task 2023-FH1 - *French TTS*
  - **18 participants**
  - Reproducibility criteria 1 and 2
- Spoke task 2023-FS1 - *Speaker adaptation*
  - **14 participants**
  - No reproducibility criteria
- Reproducibility criterion 3 encouraged for all tasks

Reproducibility requirements

- Reproducibility criteria
  1. Used external models are **publicly-available off-the-shelf pre-trained models**, and references are given
  2. Any audio data used for training models (including for fine-tuning pre-trained models) is **publicly available** and reported
  3. Source code is **provided**

Criterion	Hub task	Spoke task
1	All teams	11/14
2	All teams	11/14
3	4/18	4/14

Participants

French level

In the team:

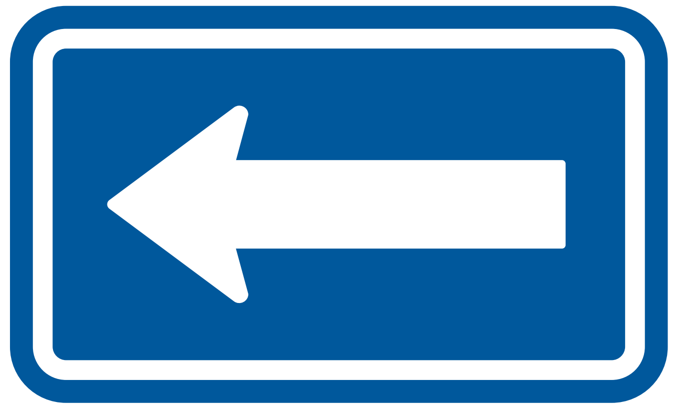
No French speaker	11
At least one French speaker but not native	2
At least one native speaker	5

Listening test design

Quality, Similarity, Intelligibility

Types of evaluation

- Speech naturalness
  - **MOS** evaluation for global assessment

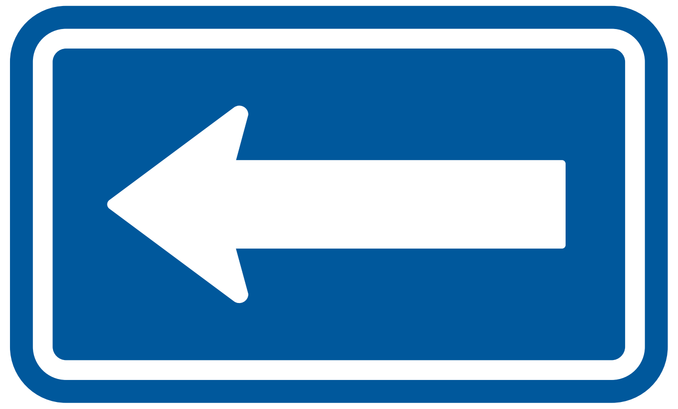


Standard Blizzard test

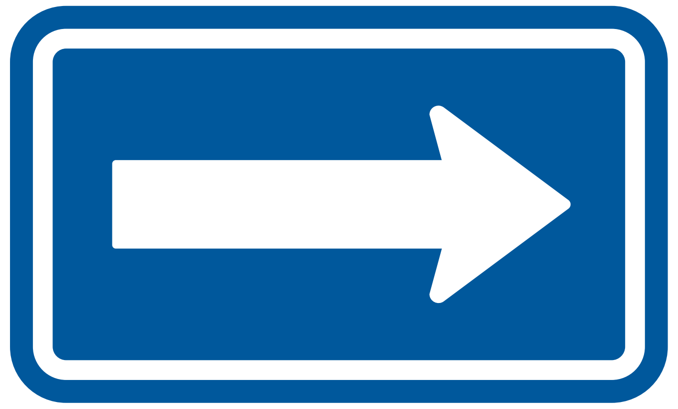
- Speaker similarity
  - **MOS** evaluation for global assessment
- Speech intelligibility
  - **SUS transcription** for global assessment

Types of evaluation

- Speech naturalness
  - **MOS** evaluation for global assessment
- Speaker similarity
  - **MOS** evaluation for global assessment
- Speech intelligibility
  - **SUS transcription** for global assessment



Standard Blizzard test



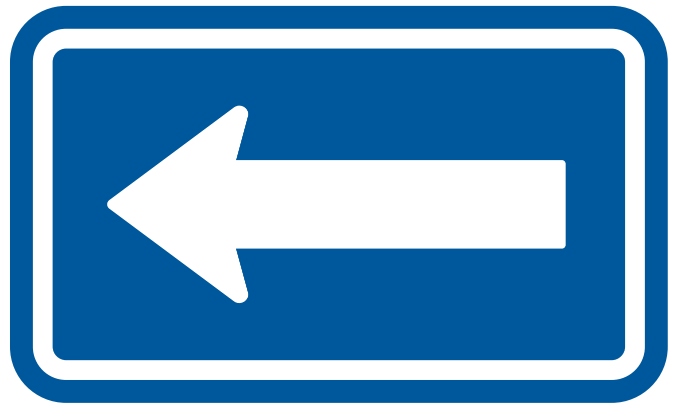
Most recent speech synthesis evaluation papers (IS, SSW)

Didn't have the time to reference them properly here, sorry for that

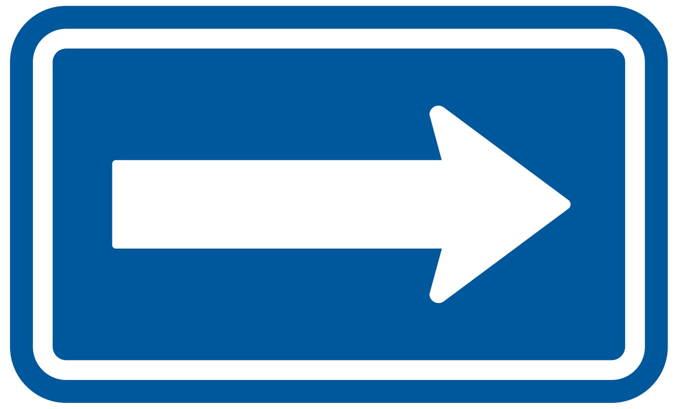
MOS

Types of evaluation

- Speech naturalness
  - **MOS** evaluation for global assessment
- Speaker similarity
  - **MOS** evaluation for global assessment
- Speech intelligibility
  - **SUS transcription** for global assessment



Standard Blizzard test



Most recent speech synthesis evaluation papers (IS, SSW)

Didn't have the time to reference them properly here, sorry for that

MOS

➡ Keep these tests for continuity but with addition of some refinements

Types of evaluation

- Speech quality
  - **MOS** evaluation for global assessment
  - Instead of speech naturalness
    - More intuitive for participants, do not affect the relative rankings of systems **A. Kirkland et al. (2023), SSW**

- Speaker similarity
  - **MOS** evaluation for global assessment

- Speech intelligibility
  - **SUS transcription** for global assessment

Types of evaluation

- Speech quality
  - **MOS** evaluation for global assessment
  - Instead of speech naturalness
    - More intuitive for participants, do not affect the relative rankings of systems A. Kirkland et al. (2023), SSW

New **MUSHRA** evaluation to refine the ranking of the best rated systems in the MOS evaluation E. Cooper et al. (2023), Interspeech

- Speaker similarity
  - **MOS** evaluation for global assessment

- Speech intelligibility
  - **SUS transcription** for global assessment

New Heterophonic **homographs recognition** for assessment of local behaviours

Types of evaluation

- Speech quality
  - **MOS** evaluation for global assessment
  - Instead of speech naturalness
    - More intuitive for participants, do not affect the relative rankings of systems A. Kirkland et al. (2023), SSW

New **MUSHRA** evaluation to refine the ranking of the best rated systems in the MOS evaluation E. Cooper et al. (2023), Interspeech

- Speaker similarity
  - **MOS** evaluation for global assessment
- Speech intelligibility
  - **SUS transcription** for global assessment

New Heterophonic **homographs recognition** for assessment of local behaviours

For all tests, selection of the ***utterances that maximise the dispersion of the systems***

Types of evaluation

- Speech quality
  - **MOS** evaluation for global assessment
  - Instead of speech naturalness
    - More intuitive for participants, do not affect the relative rankings of systems A. Kirkland et al. (2023), SSW
- New **MUSHRA** evaluation to refine the ranking of the best rated systems in the MOS evaluation E. Cooper et al. (2023), Interspeech
- Speaker similarity
  - **MOS** evaluation for global assessment
- Speech intelligibility
  - **SUS transcription** for global assessment
- New Heterophonic **homographs recognition** for assessment of local behaviours
- Expressivity evaluation wish list P. Wagner et al. (2019), SSW
  - Comprehensibility (enumeration, paragraphs) M. Grice (2023), Keynote Interspeech ; D. B. Pisoni et al. (1987), Comp. Speech and Language
  - Speech in context (paragraphs) R. Clark et al. (2019), SSW ; J. O'Mahony et al. (2021), SSW
  - ➡ Lack of time and good ideas to do this this year

For all tests, selection of the ***utterances that maximise the dispersion of the systems***

Types of evaluation

- Speech quality
  - **MOS** evaluation for global assessment
  - Instead of speech naturalness
    - More intuitive for participants, do not affect the relative rankings of systems   A. Kirkland et al. (2023), SSW

New **MUSHRA** evaluation to refine the ranking of the best rated systems in the MOS evaluation   E. Cooper et al. (2023), Interspeech

- Speaker similarity
  - **MOS** evaluation for global assessment

For all tests, selection of the ***utterances that maximise the dispersion of the systems***

- Speech intelligibility
  - **SUS transcription** for global assessment

New Heterophonic **homographs recognition** for assessment of local behaviours

Under evaluation

- Expressivity evaluation wish list   P. Wagner et al. (2019), SSW
  - Comprehensibility (enumeration, paragraphs)   M. Grice (2023), Keynote Interspeech ; D. B. Pisoni et al. (1987), Comp. Speech and Language
  - Speech in context (paragraphs)   R. Clark et al. (2019), SSW ; J. O'Mahony et al. (2021), SSW

➡ Lack of time and good ideas to do this this year

Test set

Hub task

- MOS
  - 1000 distinct utterances, to be used for quality and speaker similarity evaluation, from the same source corpus as the training data

- *.Je plaide le crétinisme, l'irresponsabilité, et je réclame l'acquittement!*
- *§Le premier entretien s'arrêta là.*

Test set

Hub task

- MOS
  - 1000 distinct utterances, to be used for quality and speaker similarity evaluation, from the same source corpus as the training data
- INT
  - 216 utterances including heterophonic homographs (36 pairs in 3 different contexts) **M.-L. Hajj et al. (2023), SPECOM**

- *.Le messager <u>but</u> de la bière et du vin. [by]*
  The messenger <u>drank</u> some beer and wine.
- *.Le <u>but</u> de l'opération est néanmoins humanitaire: [byt]*
  The <u>aim</u> of the operation is nonetheless humanitarian.

Test set

Hub task

- MOS
  - 1000 distinct utterances, to be used for quality and speaker similarity evaluation, from the same source corpus as the training data
- INT
  - 216 utterances including heterophonic homographs (36 pairs in 3 different contexts) **M.-L. Hajj et al. (2023), SPECOM**
  - 110 semantically unpredictable sentences (SUS) **C. Benoït et al. (1994), Speech Comm. 18(4)**

- *.Le champ vit contre le mot drôle.*
  The field lives against the funny word
- *.Le fils lourd souhaite le seuil.*
  The heavy son wishes the threshold.

Test set

Hub task

- MOS
  - 1000 distinct utterances, to be used for quality and speaker similarity evaluation, from the same source corpus as the training data
- INT
  - 216 utterances including heterophonic homographs (36 pairs in 3 different contexts) M.-L. Hajj et al. (2023), SPECOM
  - 110 semantically unpredictable sentences (SUS) C. Benoït et al. (1994), Speech Comm. 18(4)
- EXP
  - 100 enumerations of 4 objects

- *§Dans mon panier, il y a: un **livre** noir, une **boule** blanche, un **éléphant** bleu et une **poupée** verte.*

  *In my basket, there are: a black **book**, a white **ball**, a blue **elephant** and a green **doll**.*

Test set

Hub task

- MOS
  - 1000 distinct utterances, to be used for quality and speaker similarity evaluation, from the same source corpus as the training data
- INT
  - 216 utterances including heterophonic homographs (36 pairs in 3 different contexts) M.-L. Hajj et al. (2023), SPECOM
  - 110 semantically unpredictable sentences (SUS) C. Benoït et al. (1994), Speech Comm. 18(4)
- EXP
  - 100 enumerations of 4 objects
  - 213 paragraphs, from the same source corpus as the training data

- *§L'aéronef fit un crochet à droite pour éviter les hautes tours de l'Observatoire et de la grande usine électrique du mont Valérien, puis d'un seul bond au-dessus du quartier industriel de Nanterre, elle arriva au tournant de la Seine.§*

Test set

Hub task

- MOS
  - 1000 distinct utterances, to be used for quality and speaker similarity evaluation, from the same source corpus as the training data
- INT
  - 216 utterances including heterophonic homographs (36 pairs in 3 different contexts)
  - 110 semantically unpredictable sentences (SUS)
- EXP
  - 100 enumerations of 4 objects
  - 213 paragraphs, from the same source corpus as the training data

Spoke task

- MOS
  - 400 distinct utterances, to be used for quality and speaker similarity evaluation, from the same source corpus as the training data
  - *§Les mots ont leur importance, monsieur le rapporteur.§*
  - *§C'est pourquoi il faut rejeter les amendements de suppression.§*

Listening test structure

	Task	Dimension	Test	Systems	# Utt.	Implementation	Duration
1.a	FH1	Quality	Mean Opinion Score *(5pt scale)*	**21** = A + BF + BT + 18 systems	42	Latin square *(2 utterances per system | 21 groups)*	20 min
1.b	FH1	Similarity	Mean Opinion Score *(5pt scale)*		42	Latin square *(2 utterances per system | 21 groups)*	
2	FH1	Quality	MUSHRA	**5** = A + BF + 3 best systems	20	Same test for all	27 min
3.a	FH1	Intelligibility	Transcription (SUS)	**20** = BF + BT + 18 systems	20	Latin square *(1 utterances per system | 20 groups)*	22 min
3.b	FH1	Intelligibility	ABX (Homographs)		72 + 72	Latin square *(36 pairs of homo. per system | 20 groups)*	
4.a	FS1	Quality	Mean Opinion Score *(5pt scale)*	**17** = A + BF + BT + 14 systems	34	Latin square *(2 utterances per system | 17 groups)*	13 min
4.b	FS1	Similarity	Mean Opinion Score *(5pt scale)*		34	Latin square *(2 utterances per system | 17 groups)*	
5	FS1	Quality	MUSHRA	**6** = A + BF + 4 best systems	20	Same test for all	30 min

Listening test interfaces

- All tests were implemented on the Web Audio Evaluation Toolbox M.D.Jilling et al. (2015), SMC
- Listeners could participate once per block but could participate to several blocks

Listening test interfaces

Quality MOS

- Familiarisation
  - Listening of 1 utterance synthesised by 10 different systems
  - At the beginning of the test
- Task
  - Listen and rate 1 utterance at a time using the following instruction:

Instruction (EN): Please evaluate the quality of the audio.

Instruction (FR): Veuillez évaluer la qualité de la synthèse.

Scale (EN |FR):

	EN	FR
1.	Very Poor	Très mauvaise
2.	Poor	Mauvaise
3.	Fair	Passable
4.	Good	Bonne
5.	Excellent	Excellente

Listening test interfaces

Similarity MOS

- Familiarisation
  - Listening of 4 reference samples of the original speaker
  - Compulsory at the beginning of the test and every 7 stimuli ; facultative at anytime
- Task
  - Listen and rate 1 utterance at a time using the following instructions:

Instruction (EN): Please evaluate the similarity between the reference speaker and the voice in the present audio.

Instruction (FR): Veuillez évaluer la similarité entre la locutrice de l'extrait audio présenté, et la locutrice de référence.

Scale (EN |FR):

	EN	FR
1.	Completely different person	Personne totalement différente
2.	Probably a different person	Personne probablement différente
3.	Similar	Proche
4.	Probably the same person	Probablement la même personne
5.	Exactly the same person	Exactement la même personne

Listening test interfaces

MUSHRA Quality

- Task
  - One explicit reference (natural speech) to listen
  - 5 or 6 non-identified audio samples to rate on a continuous scale (0 to 100), among which:
    - One hidden reference (natural speech)
    - 3 or 4 systems
    - BF

Instructions (EN): Please evaluate the quality of speech synthesis:

1. Listen to the reference audio.
2. Listen to the other audio clips and rate them relative to one another using the rating scales.
3. Once you rated all [5/6] audios, click on the sort button to place your ratings in order.
4. Re-listen to the audios from worst to best (left to right) and refine your ratings.
5. You may re-order, re-listen and refine your ratings as many times as you like.

It is required to perform steps 1 to 4 to go to the next audio sample.

Instructions (FR): Veuillez évaluer la qualité de la synthèse de parole :

1. Ecoutez l'audio de référence.
2. Ecoutez les autres extraits audio et notez-les relativement aux autres en utilisant toute l'échelle de notation.
3. Une fois notés, cliquez sur "Ordonner" pour ordonner les extraits audios dans l'ordre croissant des notes que vous leurs avez attribuées.
4. Réécoutez chaque extrait dans l'ordre (de gauche à droite) et affinez votre jugement.
5. Vous pouvez réordonner les extraits, les réécouter et ajuster leurs notes autant de fois que vous le souhaitez.

Il est nécessaire de suivre les étapes 1-4 pour pouvoir passer à l'extrait suivant.

Scale:

0:	Very poor	Très mauvais
25:	Poor	Mauvais
50:	Fair	Passable
75:	Good	Bon
100:	Excellent	Excellent

Listening test interfaces

SUS Intelligibility

- Task
  - Listen to each utterance only once
  - Transcribe the words that are heard according to the spelling rules of French

Instruction (FR): Transcrivez ci-dessous les mots entendus, selon les règles orthographique du Français.

- Score extraction
  - Computation of word error rate (WER) per utterance and system
  - Automatic detection/correction of common spelling mistakes, typos, and homonymous words

Listening test interfaces

Homographs Intelligibility

- Task
  - Listen to 3 audio samples:
    - The synthesis that contains one homograph highlighted in the text content displayed on the screen
    - Two audio versions of the homograph as isolated words, uttered by a reference speaker
  - Select the reference audio that corresponds the best to the pronunciation of the homograph in the synthesis, regardless of the correctness of the pronunciation

Instruction (FR): Sélectionnez l'extrait audio (en cliquant sur A ou B) dont la prononciation du mot ressemble le plus à celle du mot en majuscule dans la phrase à évaluer. Fondez votre réponse sur la prononciation du mot uniquement, et indépendamment de la grammaire de la phrase.

- Score extraction
  - Listeners are annotators (objective answer)
  - Use of Fleiss' kappa test to obtain an inter-listener agreement value per block J. R. Landis et al. (1977), Biometrics 33(1)
  - Increase number of raters (from 4) until substantial agreement is reached
  - Select the homograph that received the majority of ratings, and derive a binary correct / non-correct pronunciation score per utterance and system

Listening test participants

1817 evaluation blocks completed

Listening test participants

Recruitment methods

- Paid listeners
  - Via the Prolific platform
  - Inclusion: Self-certified French native speakers and no self-reported hearing problems
  - Test instructions in French
- Online volunteers
  - Via URLs sent to mailing lists (one URL per block)
  - Inclusion: No self-reported hearing problems
  - Test instructions in English
  - Speech Quality and Speaker Similarity only (Tests 1, 2, 4 and 5)
- Screening
  - MOS: use > 2 levels of the scale across the whole test
  - MUSHRA: rate > 80% the hidden natural speech reference in average across the whole test

Listening test participants

Total number

Before screening / After screening

	Test ID	Prolific	Volunteers	Total	
FH1 - Quality MOS	1.a	322 / 324	39 / 39	361 / 363	(99%)
FH1 - Similarity MOS	1.b	316 / 317	32 / 32	348 / 349	(99%)
FH1 - Quality MUSHRA	2	30 / 43	17 / 20	47 / 63	(75%)
FH1 - SUS Intelligibility	3.a	228 / 228	/	228 / 228	(100%)
FH1 - Homographs Intelligibility	3.b	218 / 218	/	218 / 218	(100%)
FS1 - Quality MOS	4.a	257 / 260	25 / 25	282 / 285	(99%)
FS1 - Similarity MOS	4.b	255 / 258	31 / 31	286 / 289	(99%)
FS1 - Quality MUSHRA	5	30 / 46	17 / 18	47 / 64	(73%)

Listening test participants

Total number

Before screening / After screening

	Test ID	Prolific	Volunteers	Total	
FH1 - Quality MOS	1.a	322 / 324	39 / 39	361 / 363	(99%)
FH1 - Similarity MOS	1.b	316 / 317	32 / 32	348 / 349	(99%)
FH1 - Quality MUSHRA	2	30 / 43	17 / 20	47 / 63	(75%)
FH1 - SUS Intelligibility	3.a	228 / 228	/	228 / 228	(100%)
FH1 - Homographs Intelligibility	3.b	218 / 218	/	218 / 218	(100%)
FS1 - Quality MOS	4.a	257 / 260	25 / 25	282 / 285	(99%)
FS1 - Similarity MOS	4.b	255 / 258	31 / 31	286 / 289	(99%)
FS1 - Quality MUSHRA	5	30 / 46	17 / 18	47 / 64	(73%)
	Total	**1656**	**161**	**1817**	

Listening test participants

Some stats from listeners feedback

Listener type

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b	Total
SE	39	37	30	31	18	18	11	10	**194**
SP	312	305	245	243	29	28	217	208	**1587**
SR	10	6	7	12	0	1	0	0	**36**

SE = Speech expert (paid or volunteer)
SP = Paid participant, non expert
SR = Volunteer, non expert

French native / non-native speaker

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b	Total
native	336	328	269	275	40	39	228	218	**1733**
non-native	25	20	13	11	7	8	0	0	**84**

Gender

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b	Total
Female	149	144	123	123	22	19	113	104	**797**
Male	203	195	155	159	24	28	113	108	**985**
Non binary	9	9	4	4	1	0	2	2	**31**
Unanswered	0	0	0	0	0	0	0	4	**4**

Analysis methodology

Statistical analysis

Comparison by pairs (Wilcoxon) vs. Full statistical model

- Previous editions R. Clark et al. (2007), Blizzard Chalenge
  - Wilcoxon's signed rank test applied between each pair of systems given the factor levels under investigation

- Limitations of comparison by pairs
  - Multiplicity of statical tests
    - High number of statistical tests that are performed artificially increases the chance of getting significant results.
    - Bonferroni correction is applied but generally too strong: it inversely decreases the chance of getting significant results.
  - A Wilcoxon test compares pairs of distributions based on the ranking of the samples from both distributions
    - MOS can take only five different values: dramatic number of ties in the ranking

➡ A full statistical model only performs a single statistical test, and is adapted to the data

Statistical analysis

1. Selection of the factors under investigation
   - Listener type
     (SE = Speech expert ; SP = Paid participant ; SR = Volunteer)
   - Speech expertise
     (SE = Speech expert ; N-SE = SP+SR = Non speech expert)
   - Is French native speaker

Statistical analysis

1. Selection of the factors under investigation
   - Listener type
     (SE = Speech expert ; SP = Paid participant ; SR = Volunteer)
   - Speech expertise
     (SE = Speech expert ; N-SE = SP+SR = Non speech expert)
   - Is French native speaker

2. Descriptive statistics
   - Median, mean, standard deviation, etc. **to use with care**
   - Since data from MOS tests is **ordinal**, we should not say things like "*halfway between*" or "*closes half the gap to natural speech*"

Statistical analysis

1. Selection of the factors under investigation
   - Listener type
     (SE = Speech expert ; SP = Paid participant ; SR = Volunteer)
   - Speech expertise
     (SE = Speech expert ; N-SE = SP+SR = Non speech expert)
   - Is French native speaker
2. Descriptive statistics
   - Median, mean, standard deviation, etc. **to use with care**
   - Since data from MOS tests is **ordinal**, we should not say things like "*halfway between*" or "*closes half the gap to natural speech*"
3. Statistical models

Test ID	**1, 4**	**2, 5**	**3.a**	**3.b**
Score	MOS	MUSHRA	WER	Correct score
Data type	Ordinal	Proportion		Binary
Statistical model	Ordinal-	Beta-		Logistic-
	regression with random effects			
R function	`clmm`	`glmmTMB`		`glmer`
R package	`ordinal`	`glmmTMB`		`lme4`

Statistical analysis

1. Selection of the factors under investigation
   - Listener type
     (SE = Speech expert ; SP = Paid participant ; SR = Volunteer)
   - Speech expertise
     (SE = Speech expert ; N-SE = SP+SR = Non speech expert)
   - Is French native speaker

2. Descriptive statistics
   - Median, mean, standard deviation, etc. **to use with care**
   - Since data from MOS tests is **ordinal**, we should not say things like "*halfway between*" or "*closes half the gap to natural speech*"

3. Statistical models

Test ID	**1, 4**	**2, 5**	**3.a**	**3.b**
Score	MOS	MUSHRA	WER	Correct score
Data type	Ordinal	Proportion		Binary
Statistical model	Ordinal-	Beta-		Logistic-
	regression with random effects			
R function	`clmm`	`glmmTMB`		`glmer`
R package	`ordinal`	`glmmTMB`		`lme4`

4. Simplifying the models
   - Assessing the significance of each factor and their interaction
   - Remove non-significant ones from the model

e.g.,
- *if the listener type as a significant impact on the scores*
- *if their is an effect of the listener being native*

Statistical analysis

1. Selection of the factors under investigation
   - Listener type
     (SE = Speech expert ; SP = Paid participant ; SR = Volunteer)
   - Speech expertise
     (SE = Speech expert ; N-SE = SP+SR = Non speech expert)
   - Is French native speaker
2. Descriptive statistics
   - Median, mean, standard deviation, etc. **to use with care**
   - Since data from MOS tests is **ordinal**, we should not say things like "*halfway between*" or "*closes half the gap to natural speech*"
3. Statistical models
4. Simplifying the models
   - Assessing the significance of each factor and their interaction
   - Remove non-significant ones from the model
5. Multiple comparisons
   - Comparison between each pair of levels

Test ID	**1, 4**	**2, 5**	**3.a**	**3.b**
Score	MOS	MUSHRA	WER	Correct score
Data type	Ordinal	Proportion		Binary
Statistical model	Ordinal-	Beta- regression with random effects		Logistic-
R function	`clmm`	`glmmTMB`		`glmer`
R package	`ordinal`	`glmmTMB`		`lme4`
Post-hoc analysis	Estimated marginal means	Method from T. Hothorn et al. (2008), Biometric Journal 50(3)		
R function	`emmeans`	`glht`		
R package	`emmeans`	`mutlcomp`		

e.g.,
- *If systems A and X are rated differently by native experts and non-native non-experts*
- *If systems B and Z are rated differently by paid participants and volunteers*

Results

Finally!

Overall effects of the different factors

Significance of the different factors and their interactions (p < 0.01)

Test	1.a	4.a	2	5	1.b	4.b
system	✓	✓	✓	✓	✓	✓
sentence (random)	✓	✓	✓	✓	✓	✓
listener_ID (random)	✓	✓	✓	✓	✓	✓
listener_type (SE, SP, SR)	✓				✓	
listener_type × *system*	✓				✓	
speech_expert (SE, N-SE)	✓	✓	✓	✓	✓	✓
speech_expert × *system*			✓	✓		
is_native (native, non-native)	✓	✓		✓	✓	✓
is_native × *system*	✓			✓	✓	✓
speech_expert × *is_native*		✓				
speech_expert × *is_native* × *system*						
	FH1	*FS1*	*FH1*	*FS1*	*FH1*	*FS1*
	Quality MOS		*Quality MUSHRA*		*Similarity MOS*	

Overall effects of the different factors

- Effect of the systems (trivial)

Significance of the different factors and their interactions ($p < 0.01$)

Test	1.a	4.a	2	5	1.b	4.b
system	✓	✓	✓	✓	✓	✓
sentence (random)	✓	✓	✓	✓	✓	✓
listener_ID (random)	✓	✓	✓	✓	✓	✓
listener_type (SE, SP, SR)	✓				✓	
listener_type × *system*	✓				✓	
speech_expert (SE, N-SE)	✓	✓	✓	✓	✓	✓
speech_expert × *system*			✓	✓		
is_native (native, non-native)	✓	✓		✓	✓	✓
is_native × *system*	✓			✓	✓	✓
speech_expert × *is_native*		✓				
speech_expert × *is_native* × *system*						
	FH1	*FS1*	*FH1*	*FS1*	*FH1*	*FS1*
	Quality MOS		*Quality MUSHRA*		*Similarity MOS*	

Overall effects of the different factors

- Effect of the systems (trivial)
- Effect of sentence and listener ID

Significance of the different factors and their interactions (p < 0.01)

Test	**1.a**	**4.a**	**2**	**5**	**1.b**	**4.b**
system	✓	✓	✓	✓	✓	✓
sentence (random)	✓	✓	✓	✓	✓	✓
listener_ID (random)	✓	✓	✓	✓	✓	✓
listener_type (SE, SP, SR)	✓				✓	
listener_type × *system*	✓				✓	
speech_expert (SE, N-SE)	✓	✓	✓	✓	✓	✓
speech_expert × *system*			✓	✓		
is_native (native, non-native)	✓	✓		✓	✓	✓
is_native × *system*	✓			✓	✓	✓
speech_expert × *is_native*		✓				
speech_expert × *is_native* × *system*						
	FH1	*FS1*	*FH1*	*FS1*	*FH1*	*FS1*
	Quality MOS		*Quality MUSHRA*		*Similarity MOS*	

Overall effects of the different factors

- Effect of the systems (trivial)
- Effect of sentence and listener ID
- Effect of speech expertise
  - For all tests
  - Interaction with systems (2, 5)

Significance of the different factors and their interactions ($p < 0.01$)

Test	1.a	4.a	2	5	1.b	4.b
system	✓	✓	✓	✓	✓	✓
sentence (random)	✓	✓	✓	✓	✓	✓
listener_ID (random)	✓	✓	✓	✓	✓	✓
listener_type (SE, SP, SR)	✓				✓	
listener_type × *system*	✓				✓	
speech_expert (SE, N-SE)	✓	✓	✓	✓	✓	✓
speech_expert × *system*			✓	✓		
is_native (native, non-native)	✓	✓		✓	✓	✓
is_native × *system*	✓			✓	✓	✓
speech_expert × *is_native*		✓				
speech_expert × *is_native* × *system*						
	FH1	*FS1*	*FH1*	*FS1*	*FH1*	*FS1*
	Quality MOS		*Quality MUSHRA*		*Similarity MOS*	

Overall effects of the different factors

- Effect of the systems (trivial)
- Effect of sentence and listener ID
- Effect of speech expertise
  - For all tests
  - Interaction with systems (2, 5)
- Effect of listener type
  - Only for 1.a and 1.b
  - Little difference between SP and SR

Significance of the different factors and their interactions ($p < 0.01$)

Test	1.a	4.a	2	5	1.b	4.b
system	✓	✓	✓	✓	✓	✓
sentence (random)	✓	✓	✓	✓	✓	✓
listener_ID (random)	✓	✓	✓	✓	✓	✓
listener_type (SE, SP, SR)	✓				✓	
listener_type × *system*	✓				✓	
speech_expert (SE, N-SE)	✓	✓	✓	✓	✓	✓
speech_expert × *system*			✓	✓		
is_native (native, non-native)	✓	✓		✓	✓	✓
is_native × *system*	✓			✓	✓	✓
speech_expert × *is_native*		✓				
speech_expert × *is_native* × *system*						
	FH1	*FS1*	*FH1*	*FS1*	*FH1*	*FS1*
	Quality MOS		*Quality MUSHRA*		*Similarity MOS*	

Overall effects of the different factors

- Effect of the systems (trivial)
- Effect of sentence and listener ID
- Effect of speech expertise
  - For all tests
  - Interaction with systems (2, 5)
- Effect of listener type
  - Only for 1.a and 1.b
  - Little difference between SP and SR
- Effect of is native
  - For most tests

Significance of the different factors and their interactions ($p < 0.01$)

Test	1.a	4.a	2	5	1.b	4.b
system	✓	✓	✓	✓	✓	✓
sentence (random)	✓	✓	✓	✓	✓	✓
listener_ID (random)	✓	✓	✓	✓	✓	✓
listener_type (SE, SP, SR)	✓				✓	
listener_type × *system*	✓				✓	
speech_expert (SE, N-SE)	✓	✓	✓	✓	✓	✓
speech_expert × *system*			✓	✓		
is_native (native, non-native)	✓	✓		✓	✓	✓
is_native × *system*	✓			✓	✓	✓
speech_expert × *is_native*		✓				
speech_expert × *is_native* × *system*						
	FH1	*FS1*	*FH1*	*FS1*	*FH1*	*FS1*
	Quality MOS		*Quality MUSHRA*		*Similarity MOS*	

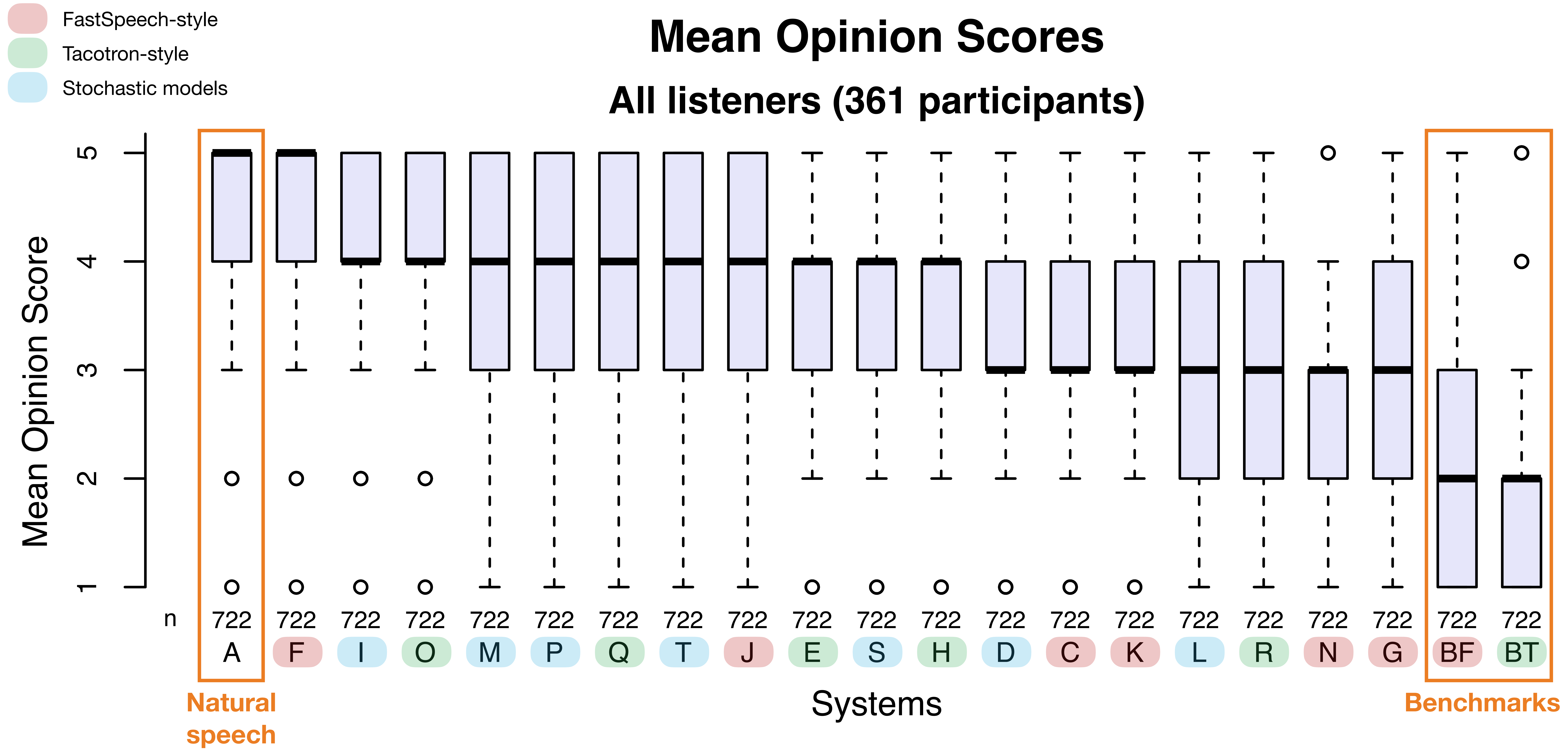
Results per factor

Effect of systems only

(all other factors combined)

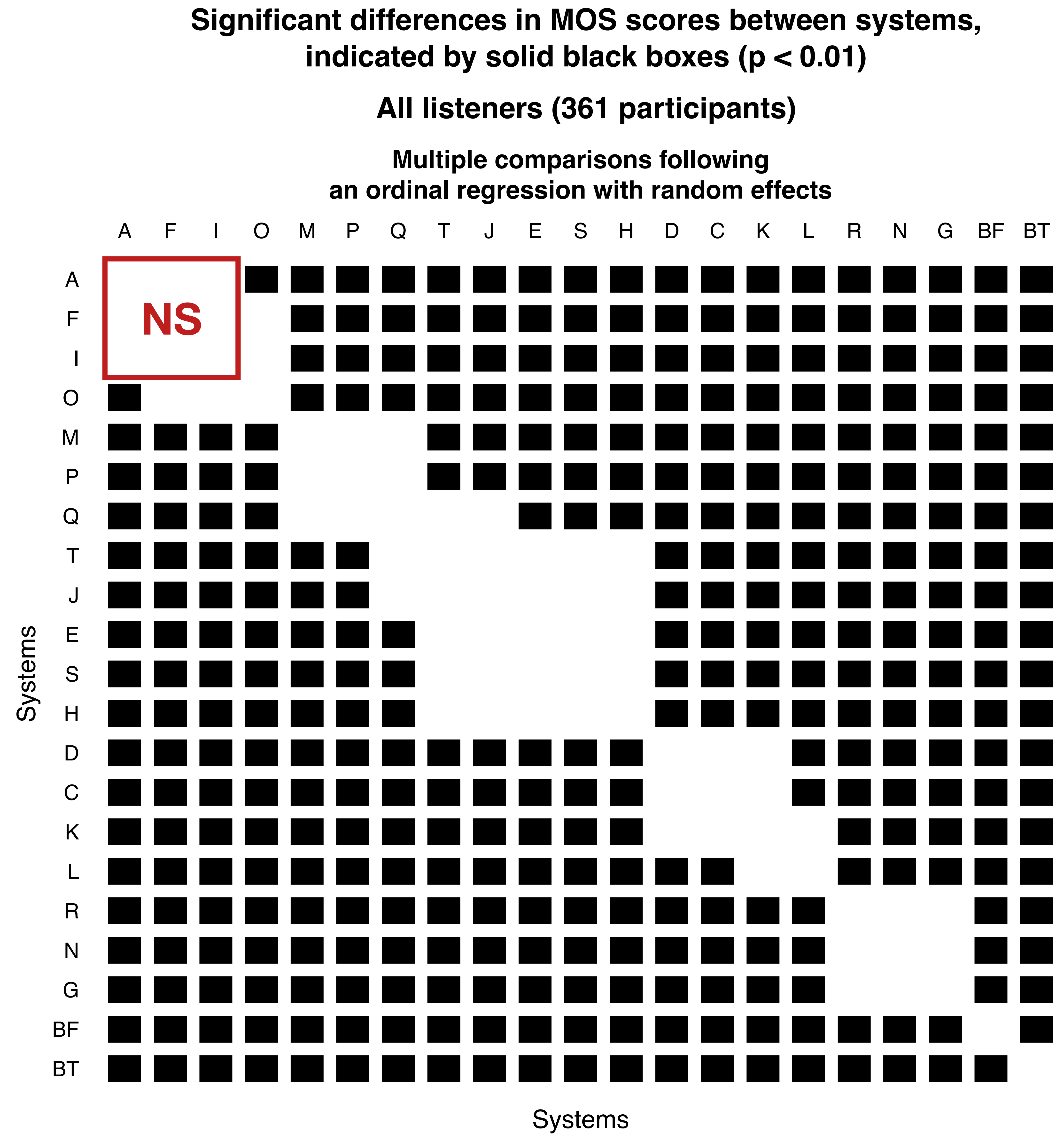
Speech quality | Hub task

Per system



Speech quality | Hub task

Per system



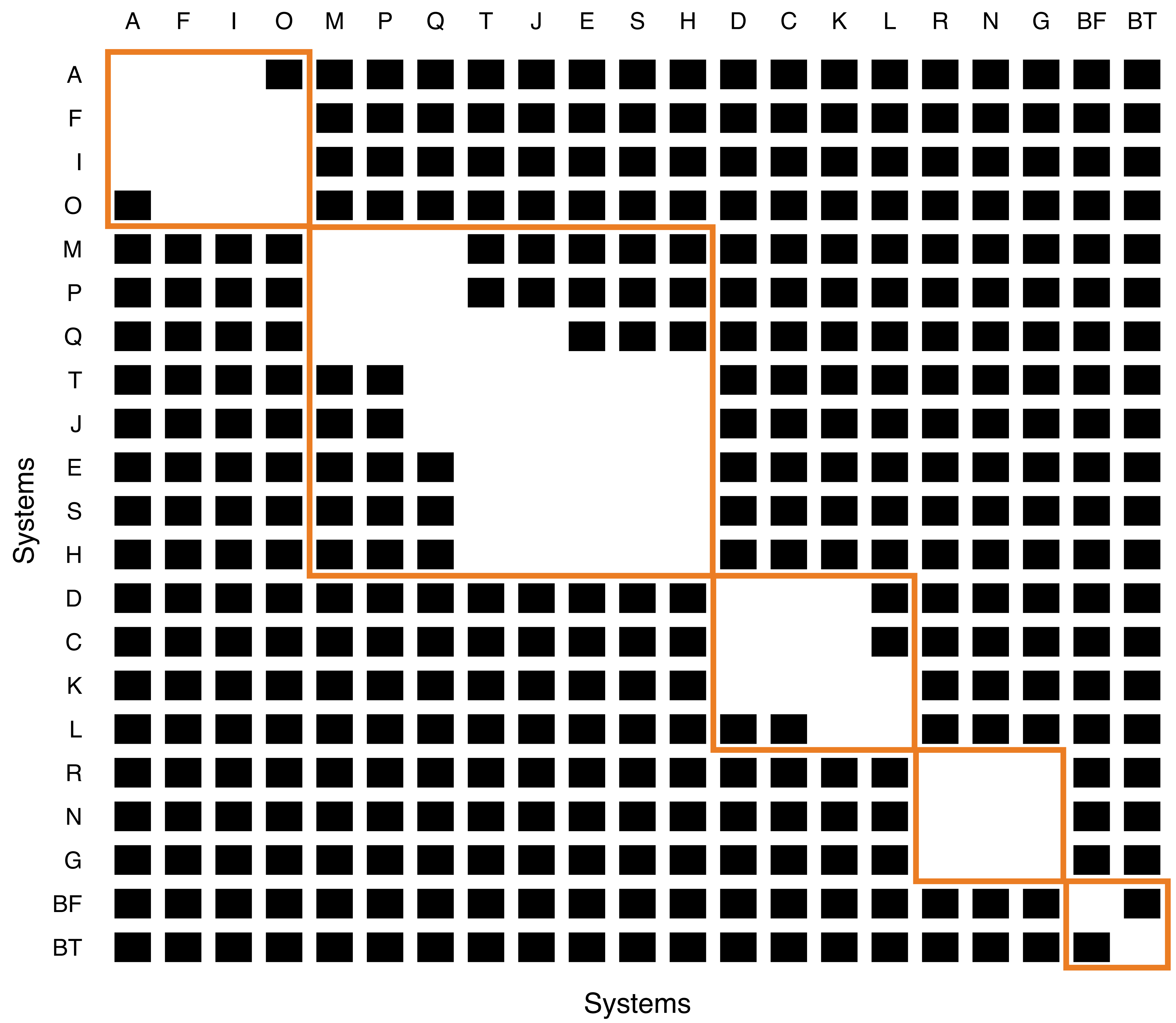
Speech quality | Hub task

Per system

Significant differences in MOS scores between systems, indicated by solid black boxes (p < 0.01)

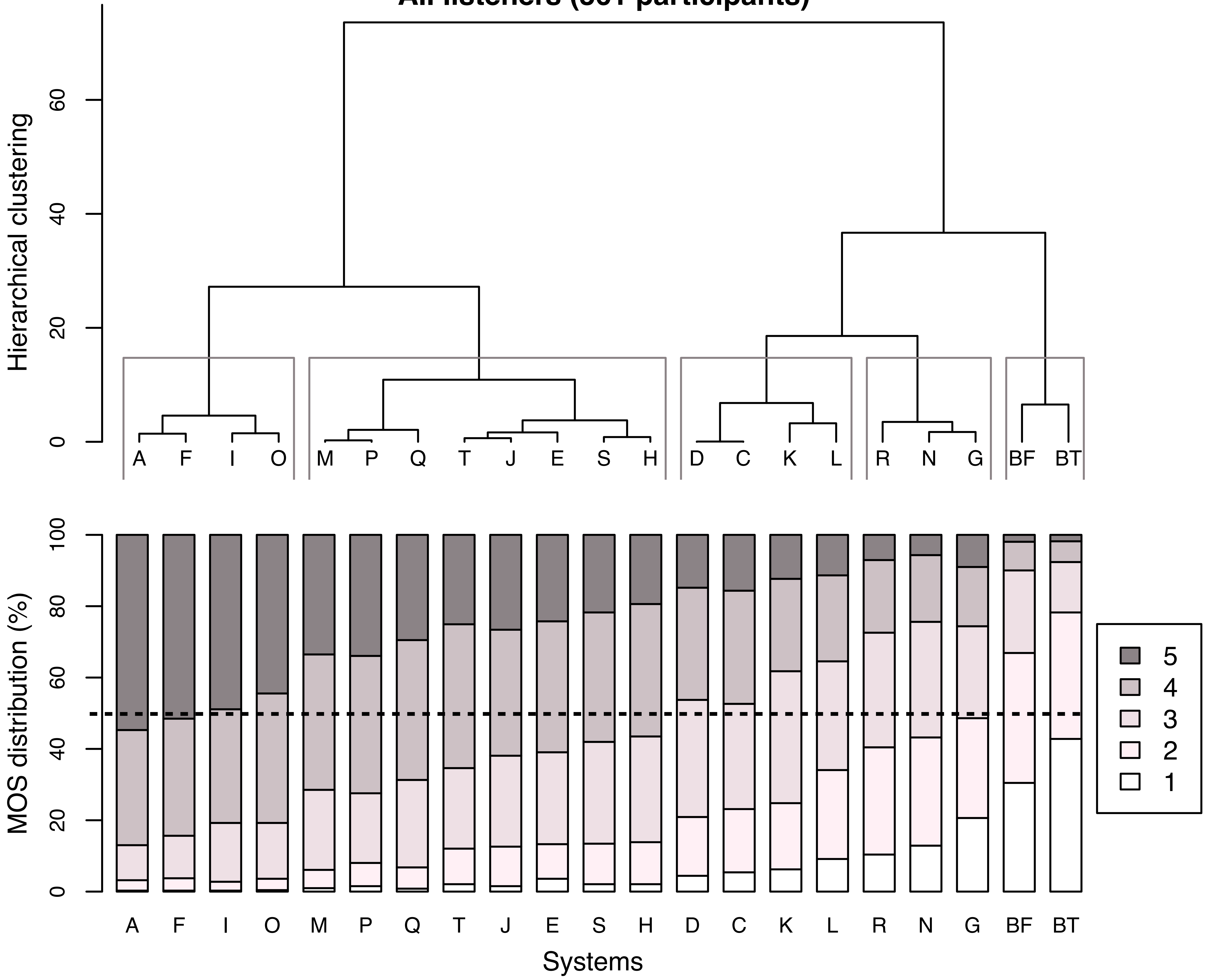
All listeners (361 participants)

Multiple comparisons following an ordinal regression with random effects



Hierarchical clustering of MOS distributions

All listeners (361 participants)



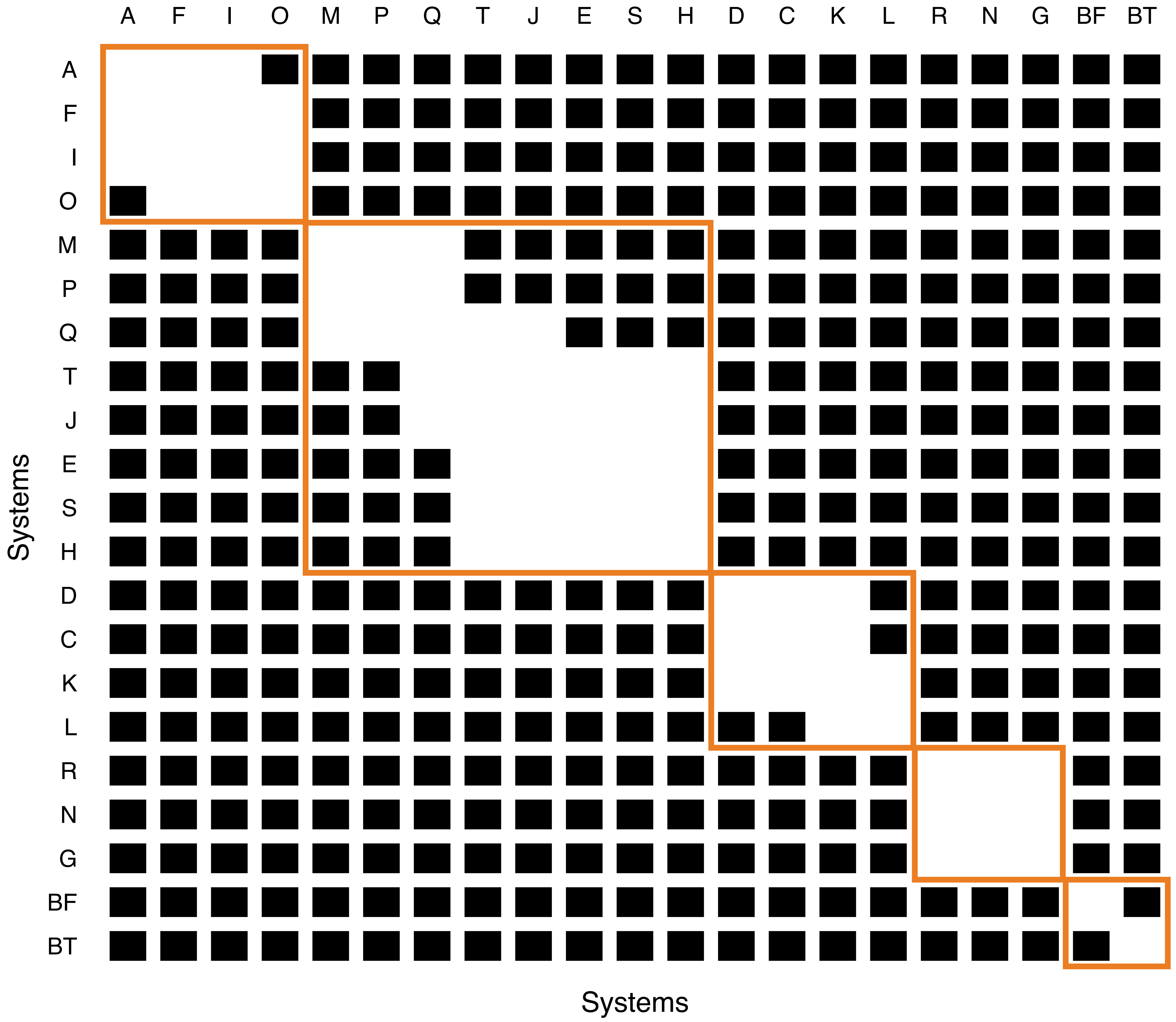
Speech quality | Hub task

Per system

Significant differences in MOS scores between systems, indicated by solid black boxes (p < 0.01)

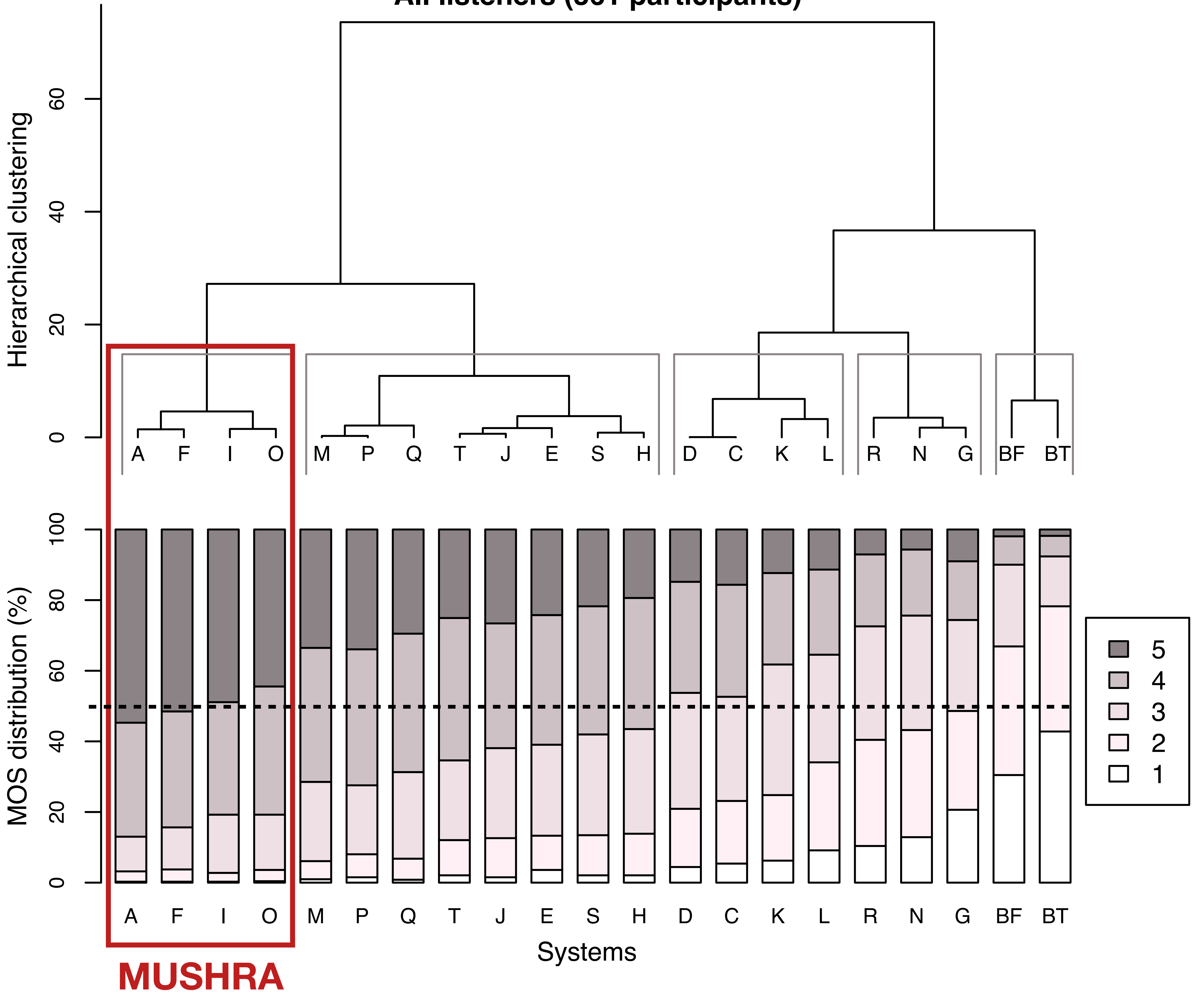
All listeners (361 participants)

Multiple comparisons following an ordinal regression with random effects

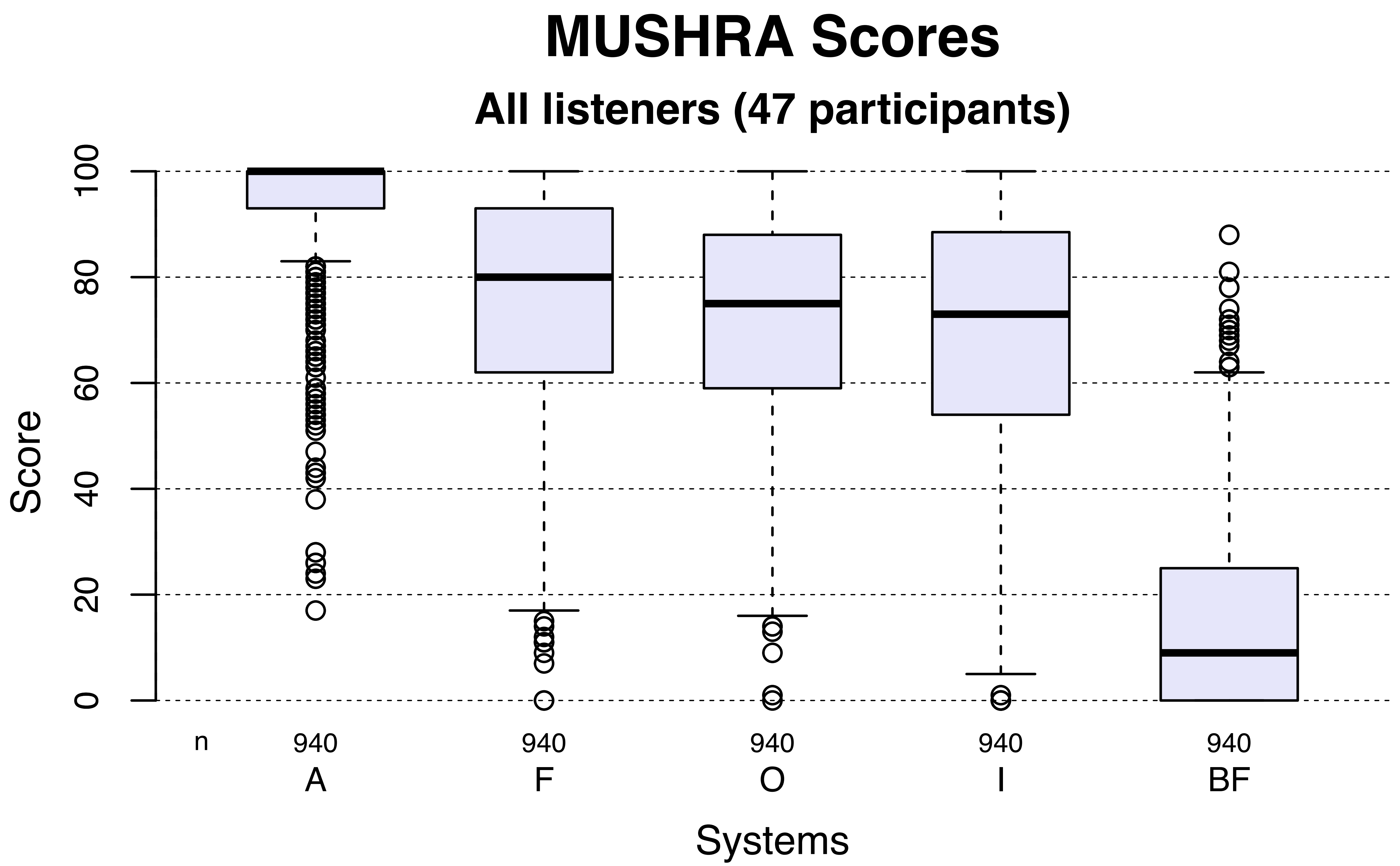


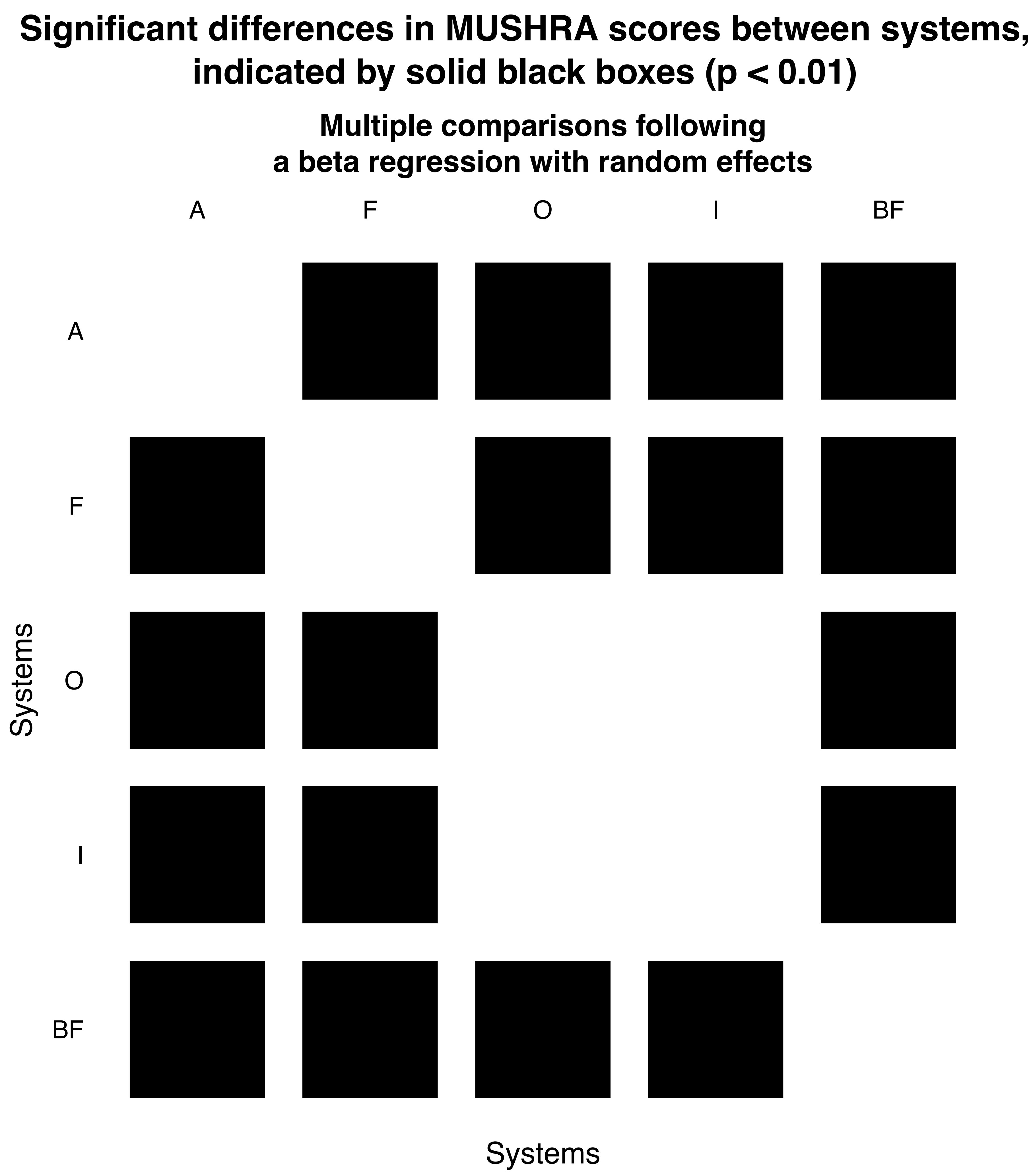
Hierarchical clustering of MOS distributions

All listeners (361 participants)



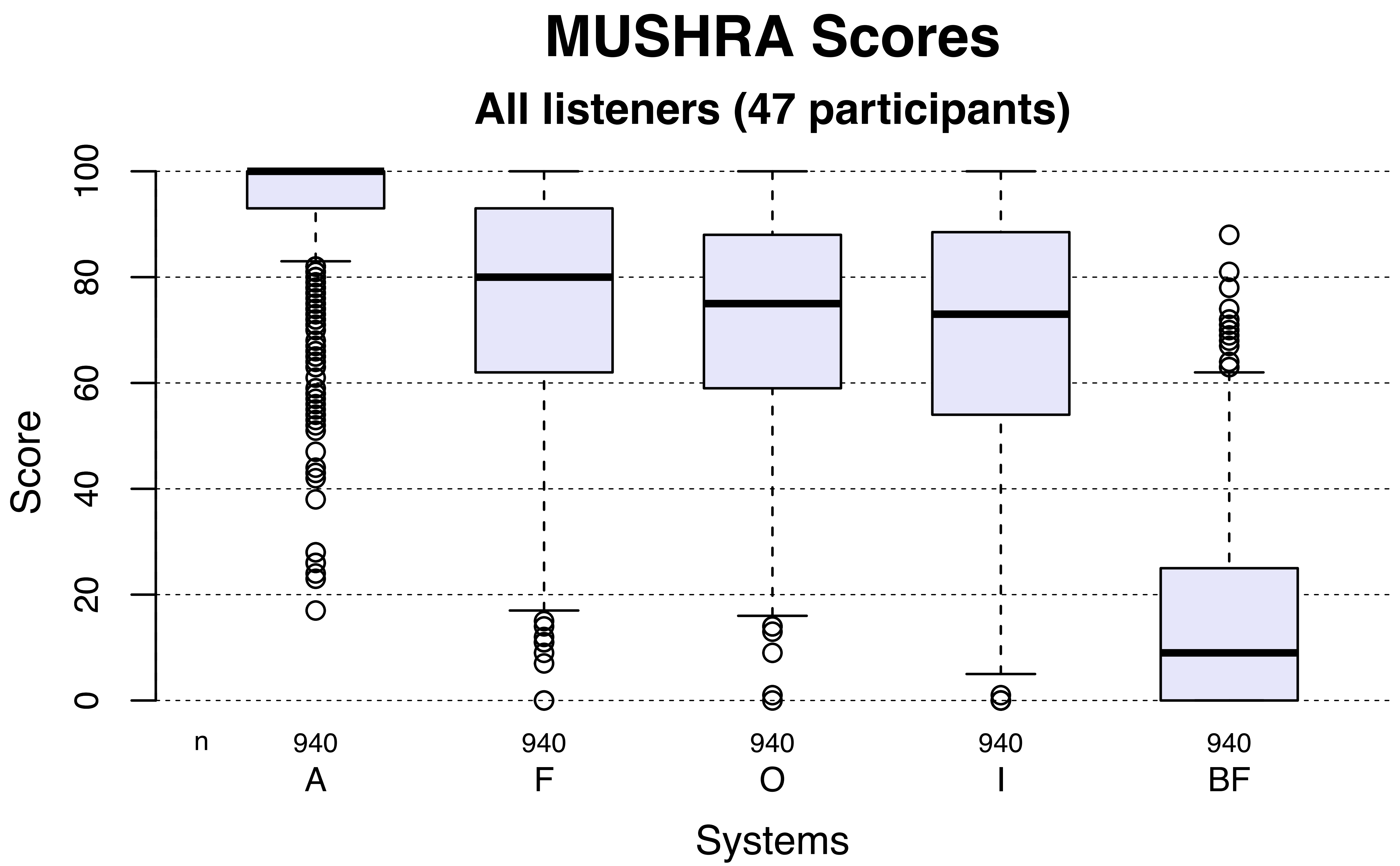
Speech quality | Hub task
Per system

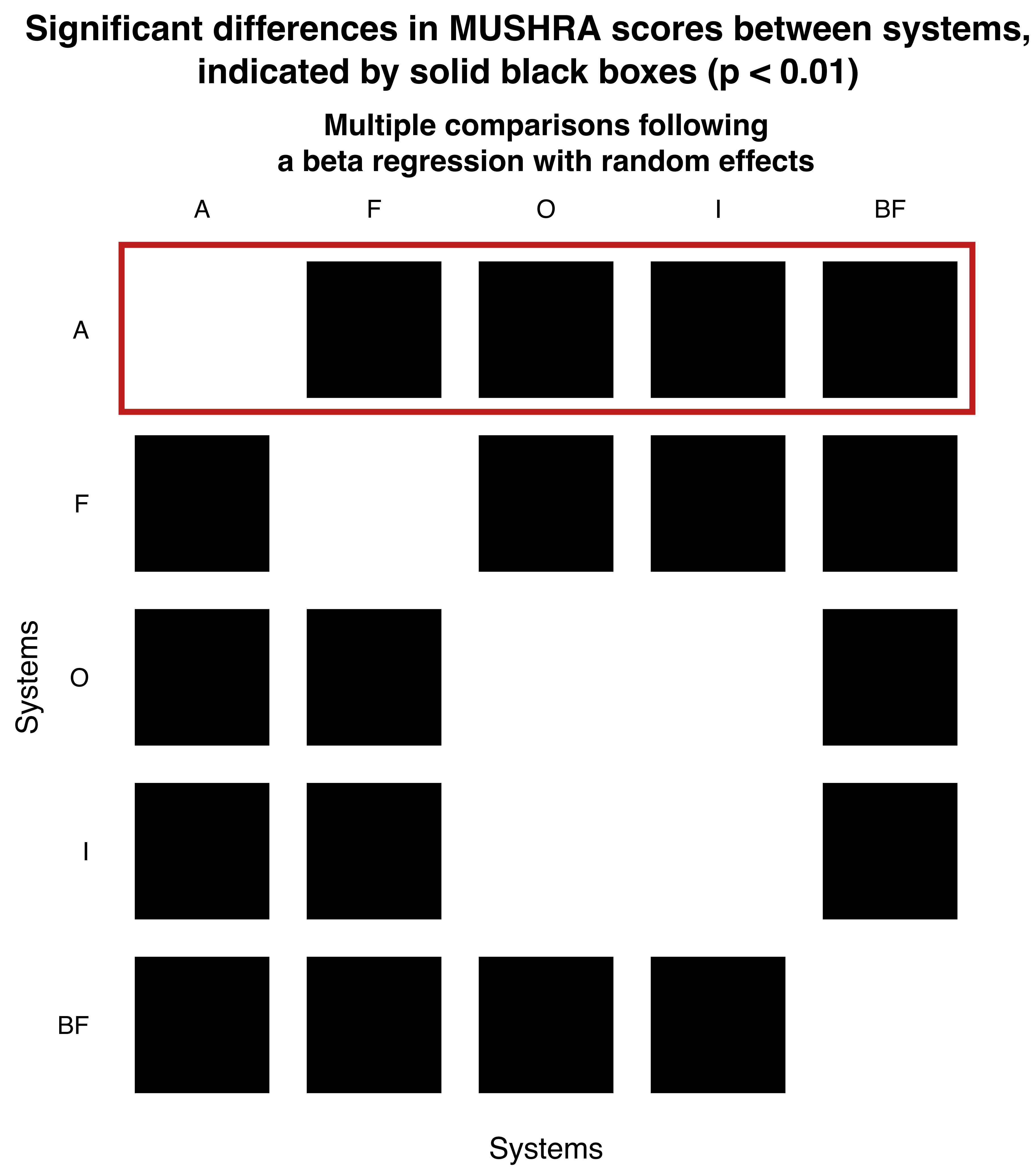




Speech quality | Hub task

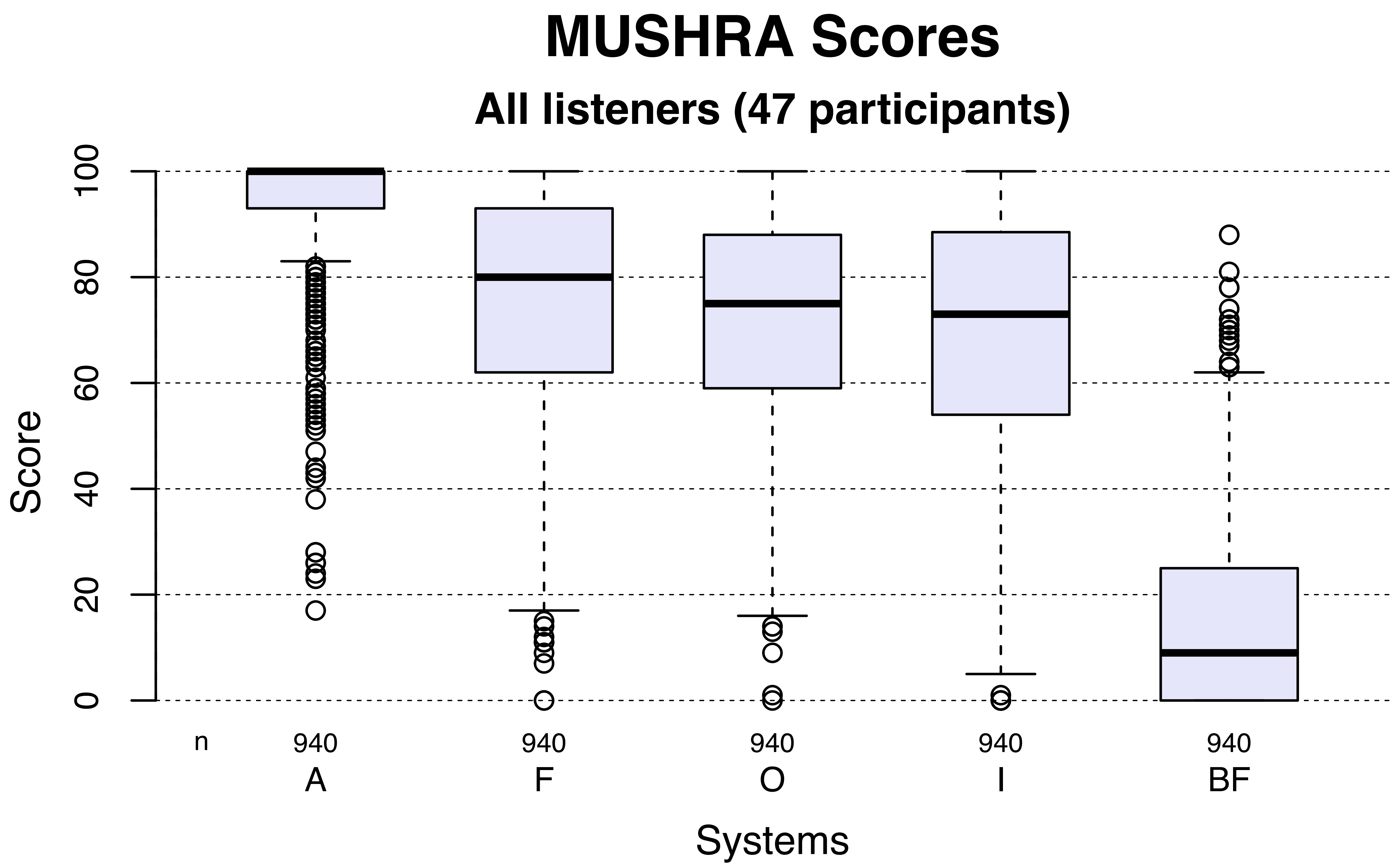
Per system

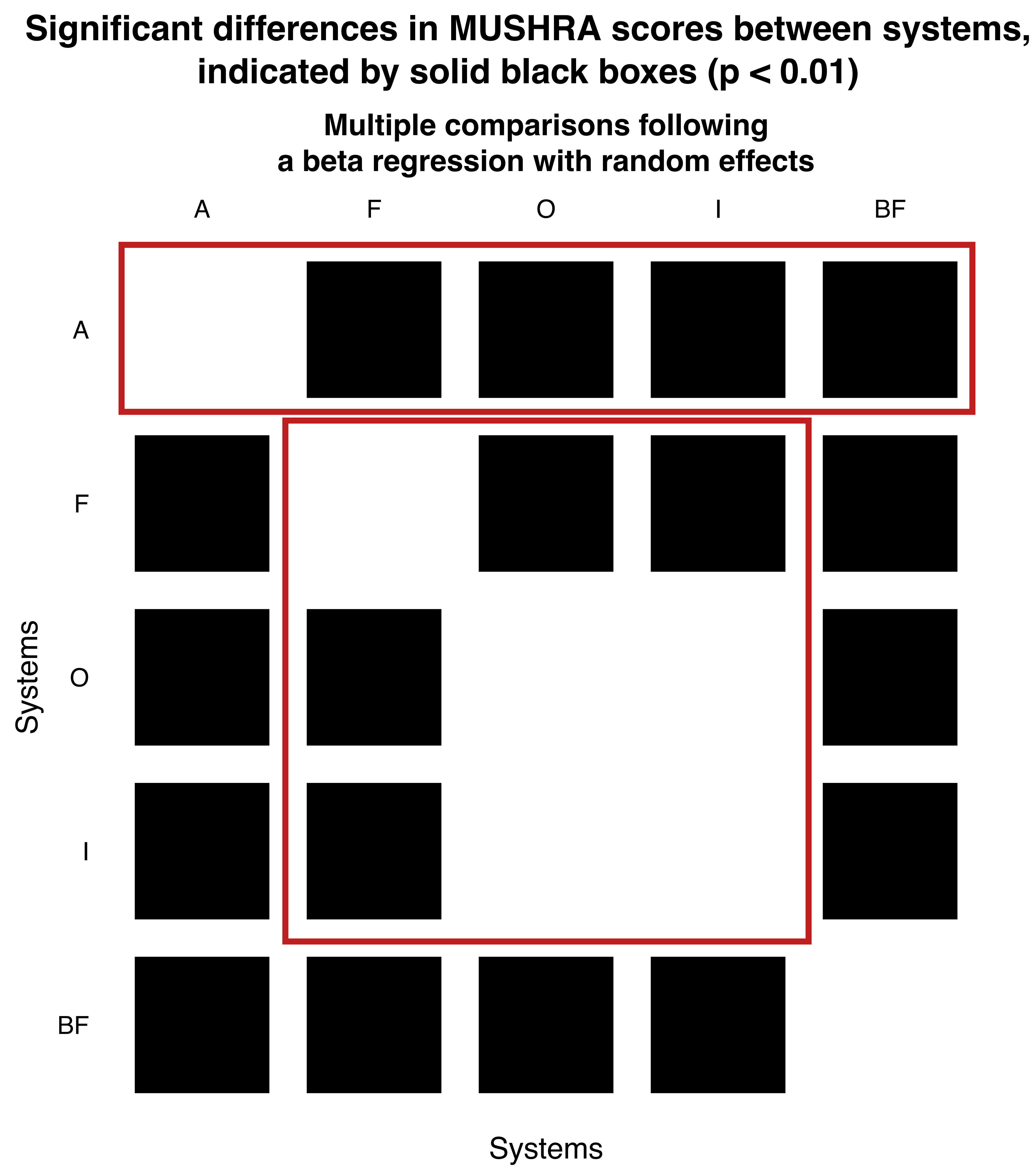




Speech quality | Hub task

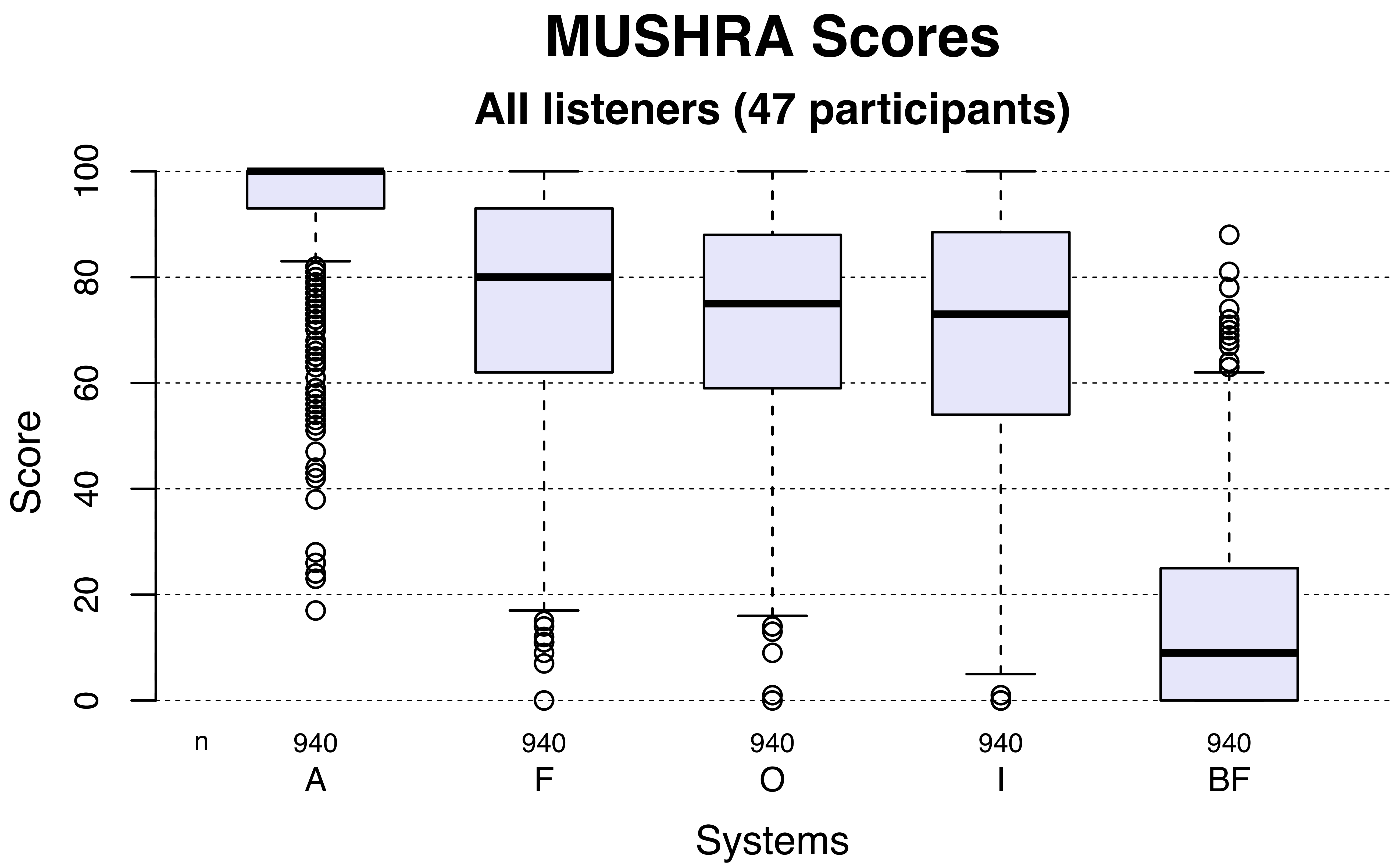
Per system

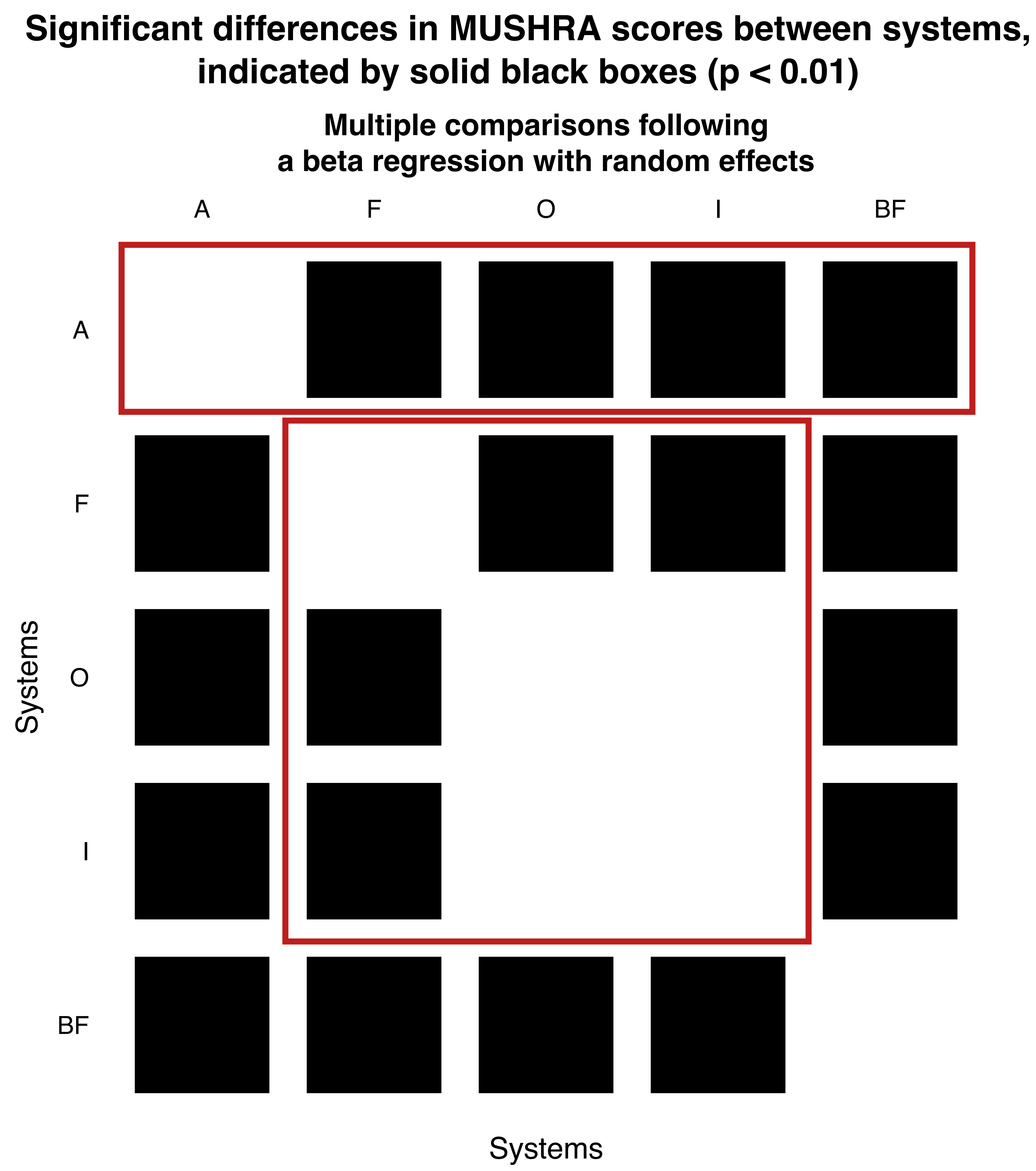




Speech quality | Hub task

Per system

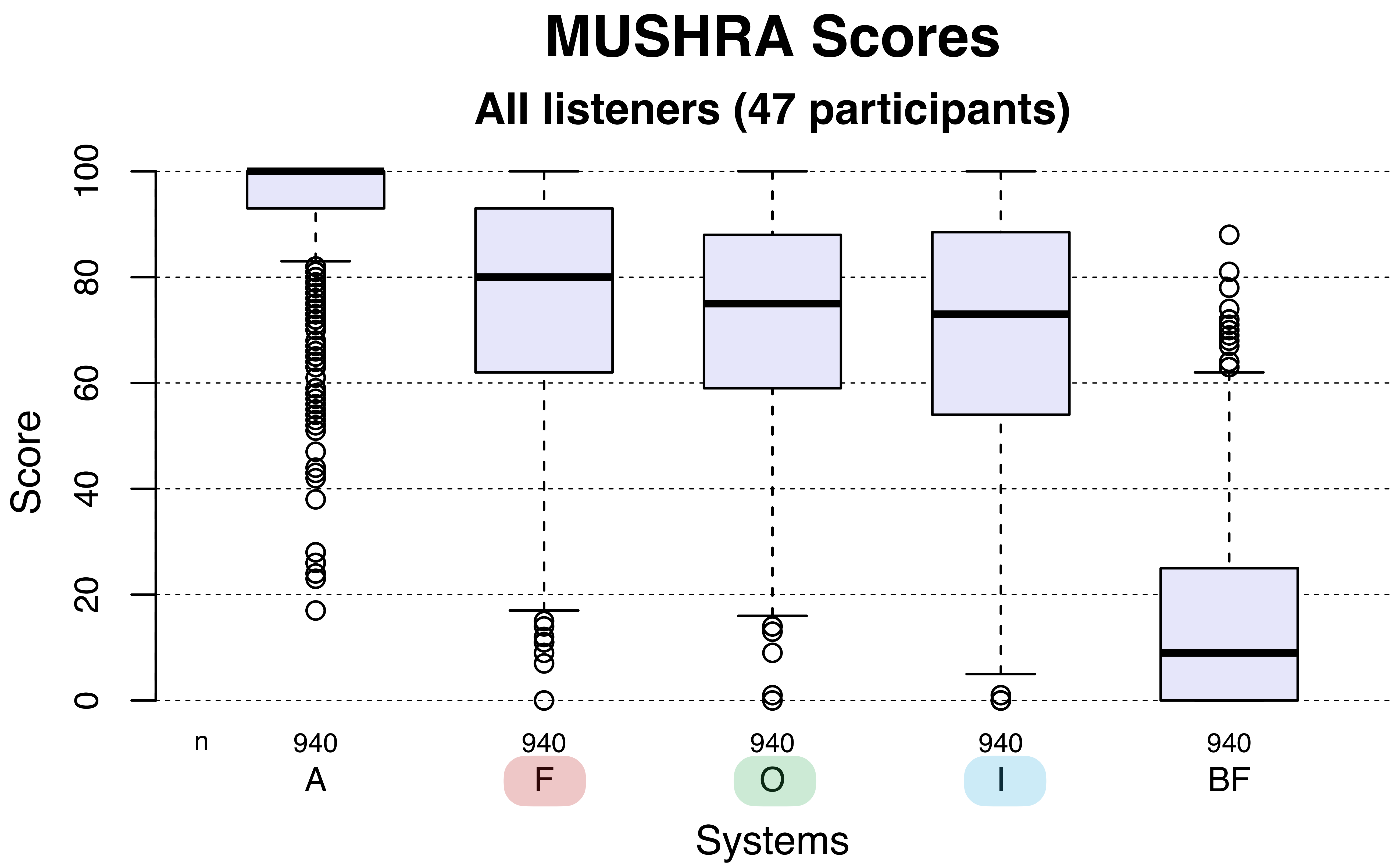


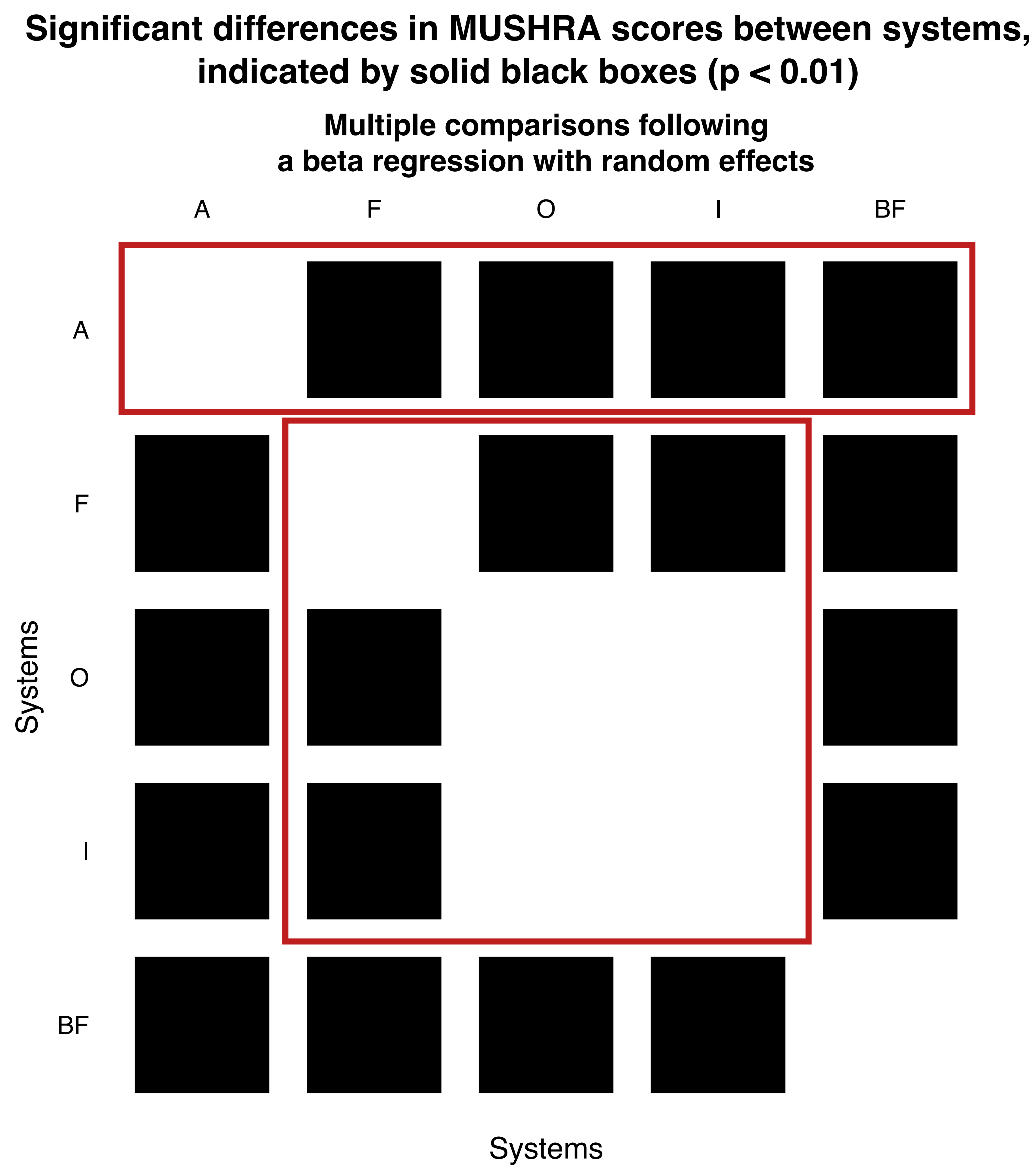


- The MUSHRA highlights the limits of the MOS test when reaching quality close to natural speech

➡ One should **NOT** conclude that the synthetic speech is 'as good as' or 'indistinguishable from' natural speech **in general** from a MOS test

Speech quality | Hub task
Per system





- The MUSHRA highlights the limits of the MOS test when reaching quality close to natural speech

➡ One should **NOT** conclude that the synthetic speech is ‘as good as’ or ‘indistinguishable from’ natural speech **in general** from a MOS test

- One of each architecture in the top 3

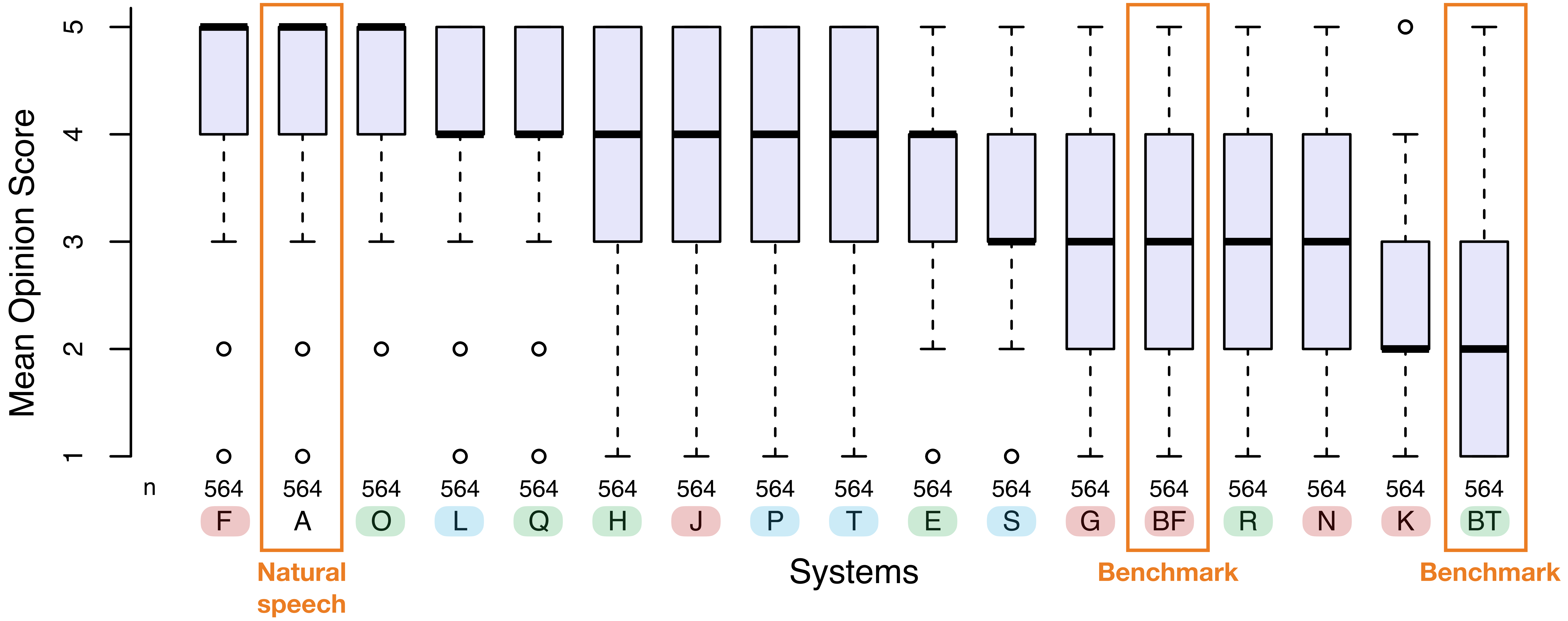
Speech quality | Spoke task

Per system

FastSpeech-style
Tacotron-style
Stochastic models

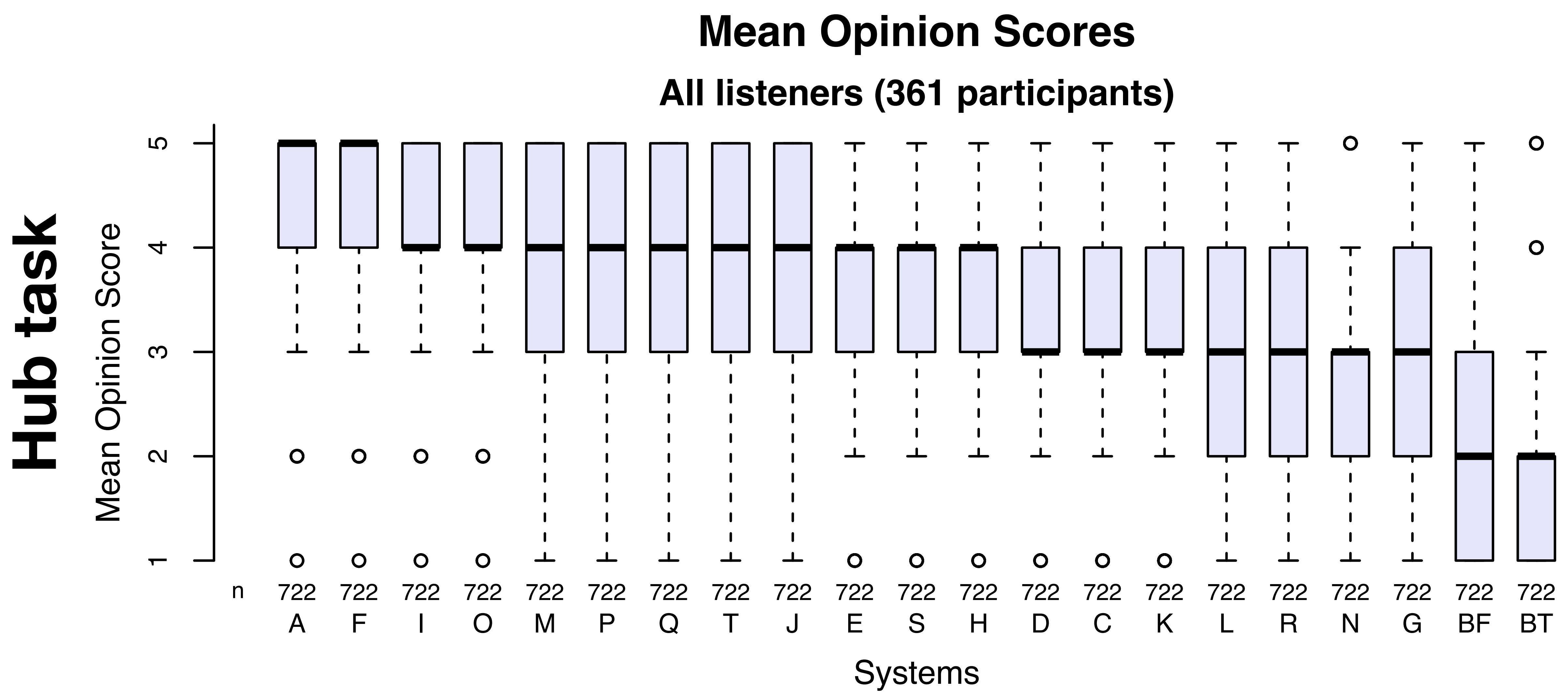
Mean Opinion Scores

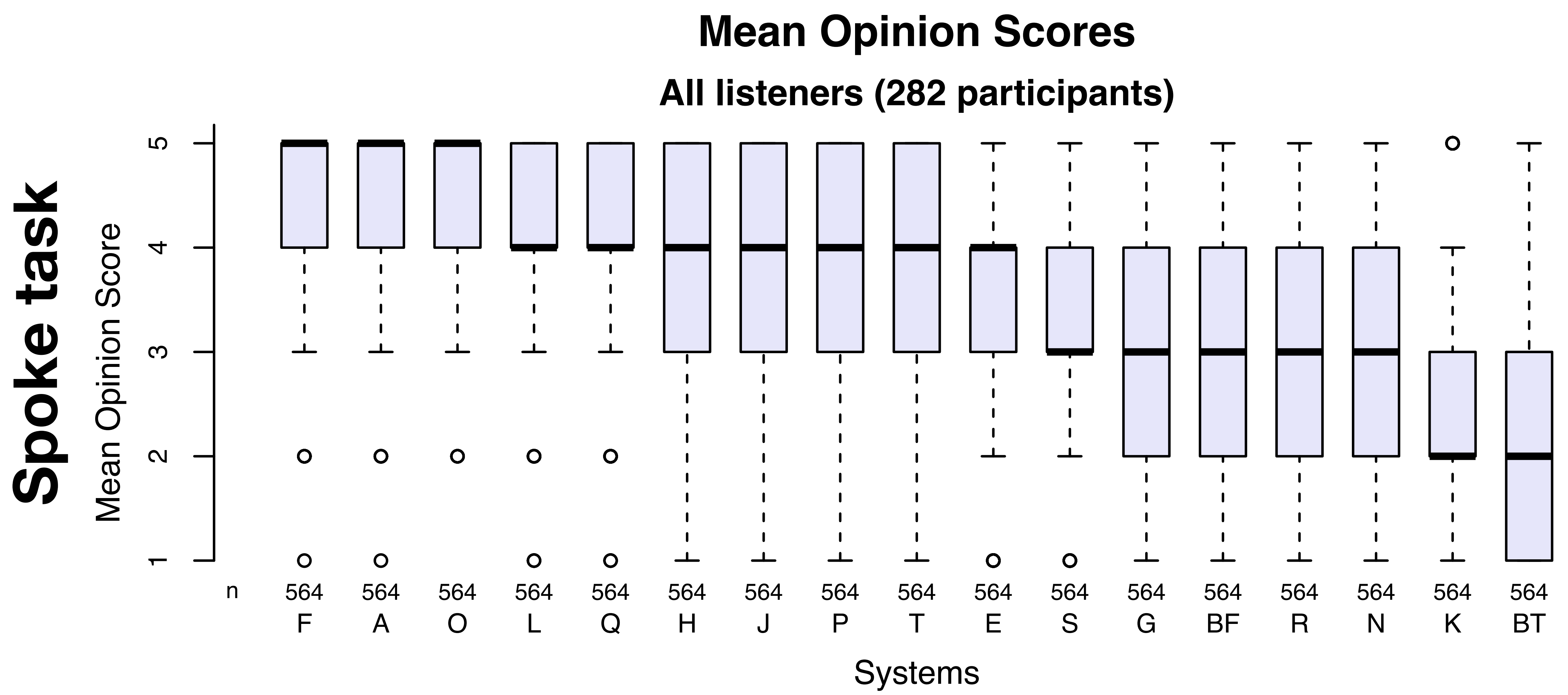
All listeners (282 participants)



Speech quality | Spoke task

Per system





- All systems have the same median in both tasks, except:
  - BF: 2 -> 3
  - L: 3 -> 4
  - O: 4 -> 5
  - K: 3 -> 2
  - S: 4 -> 3

➡ Similar score range and system repartition than Hub task

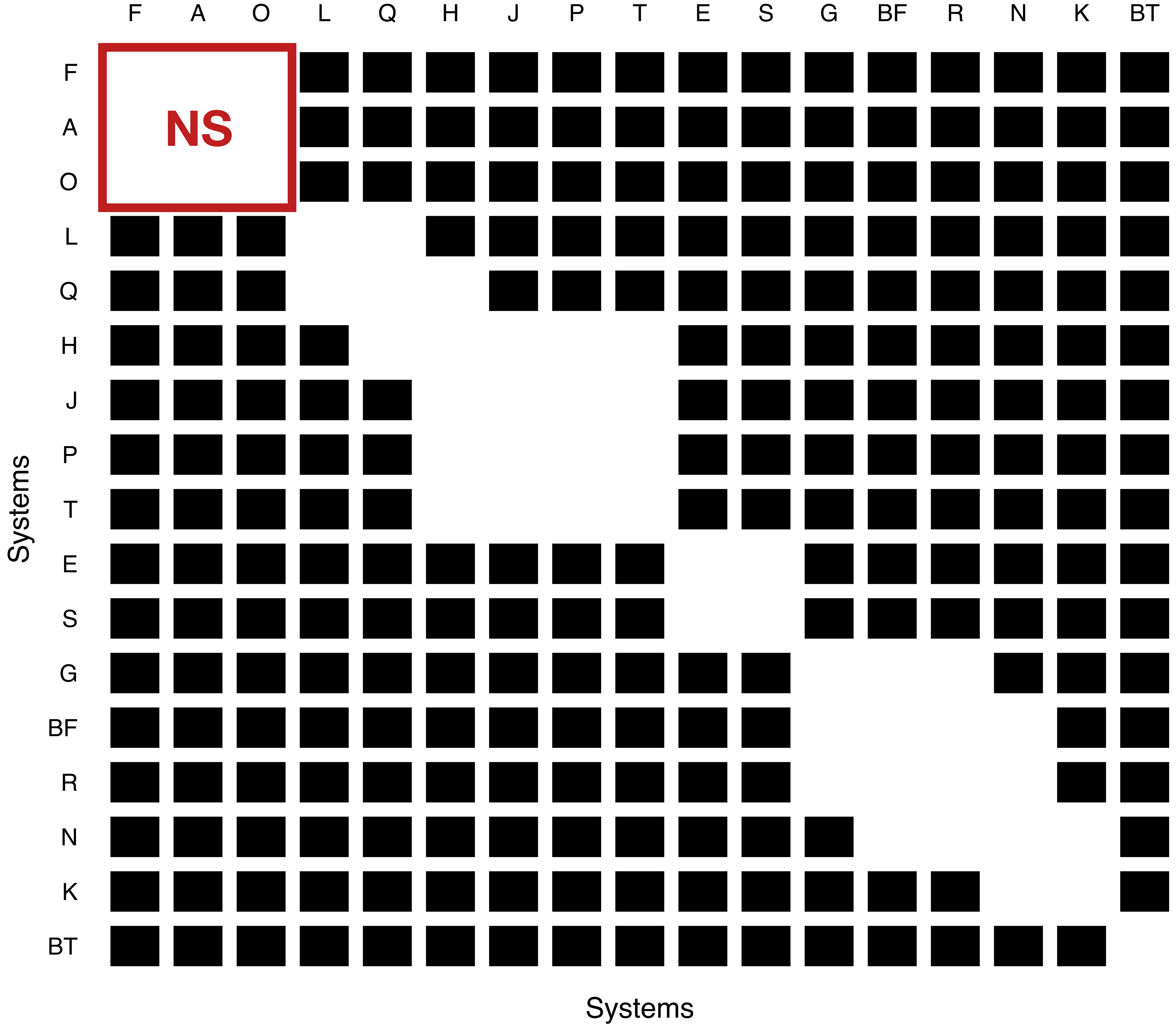
Speech quality | Spoke task

Per system

Significant differences in MOS scores between systems, indicated by solid black boxes ($p < 0.01$)

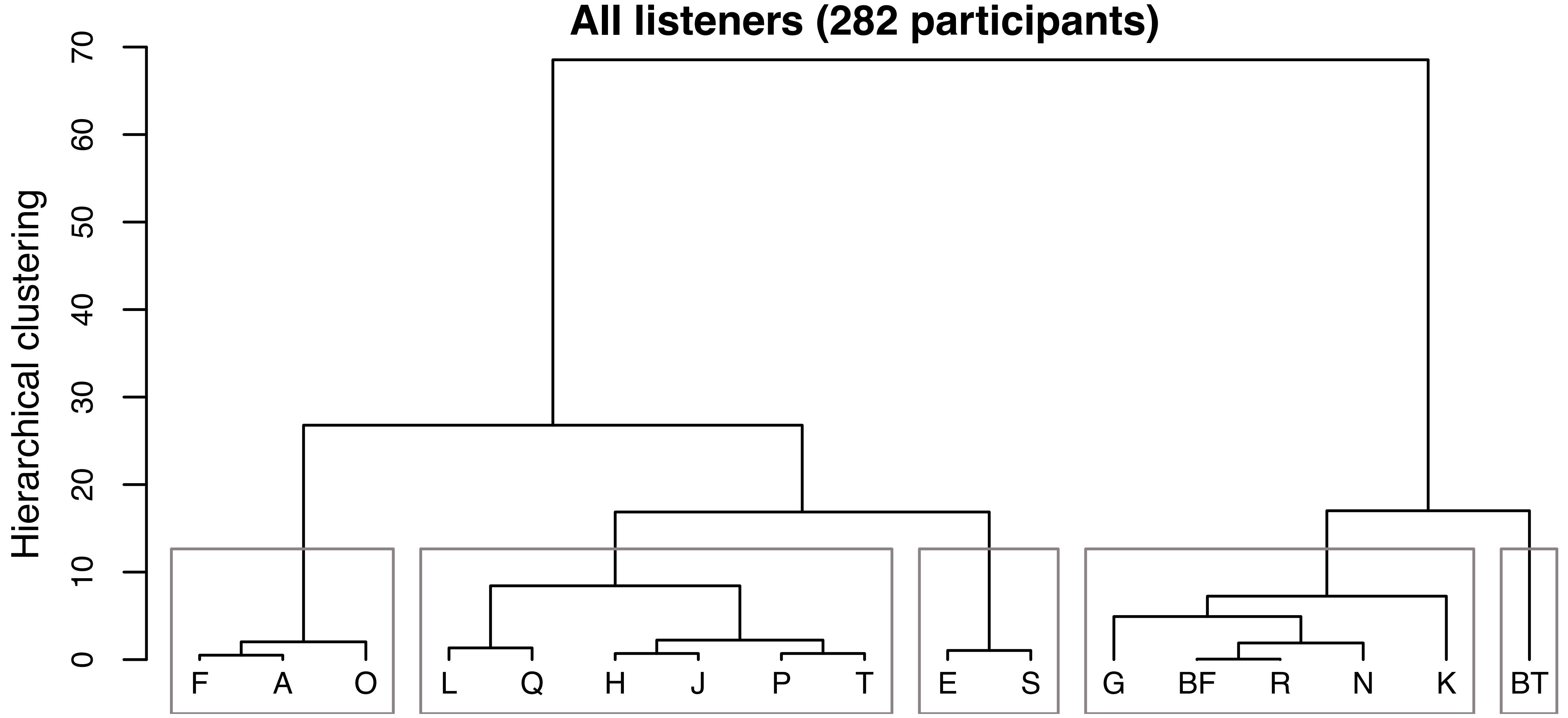
All listeners (282 participants)

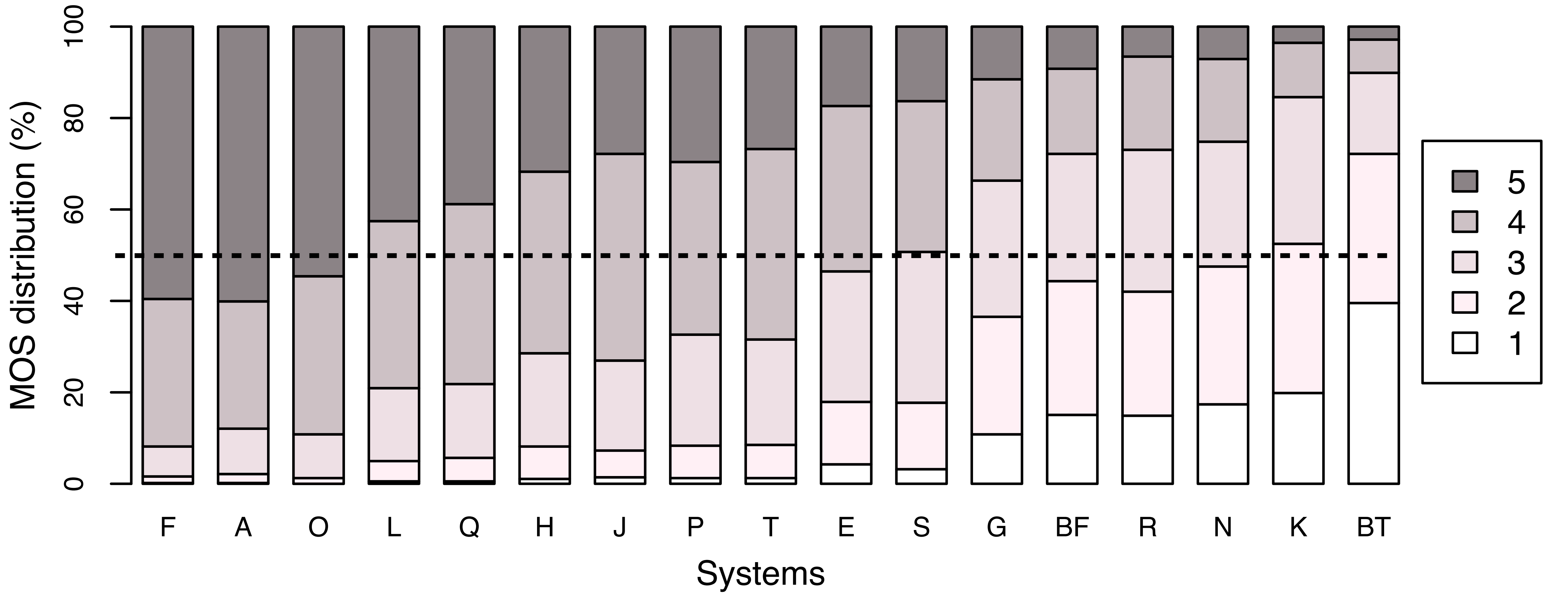
Multiple comparisons following an ordinal regression with random effects



Hierarchical clustering of MOS distributions

All listeners (282 participants)





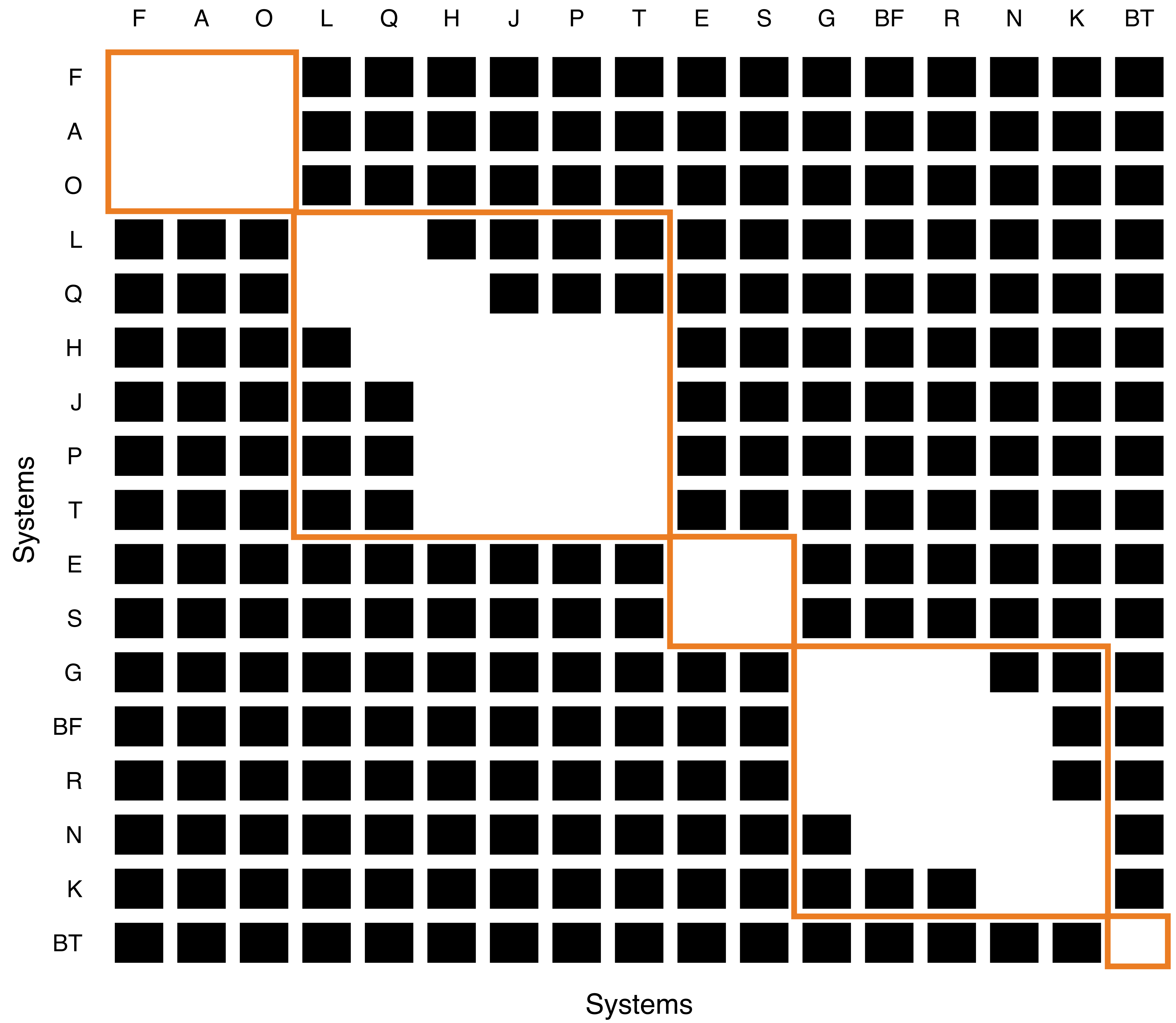
Speech quality | Spoke task

Per system

Significant differences in MOS scores between systems, indicated by solid black boxes ($p < 0.01$)

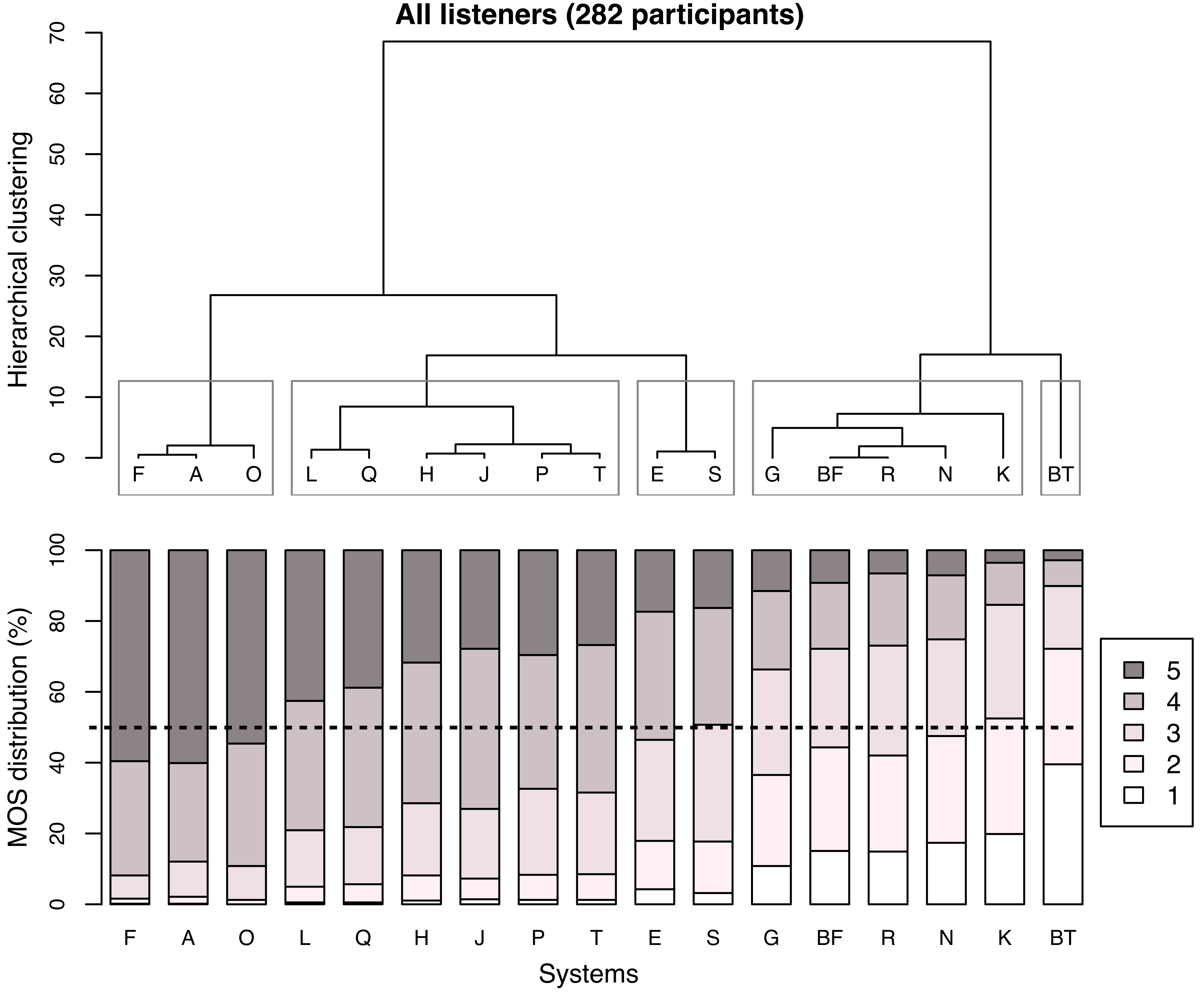
All listeners (282 participants)

Multiple comparisons following an ordinal regression with random effects



Hierarchical clustering of MOS distributions

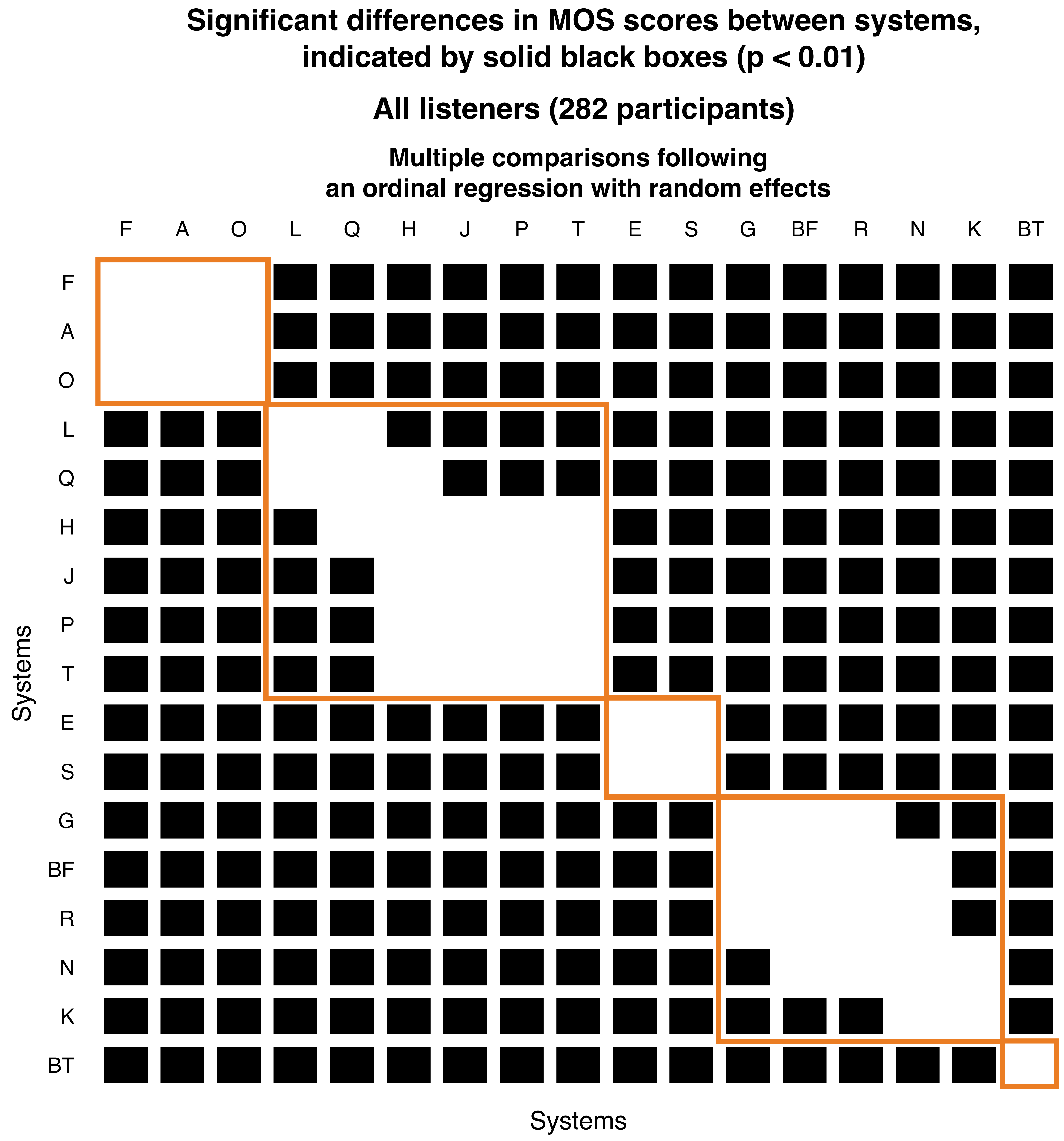
All listeners (282 participants)

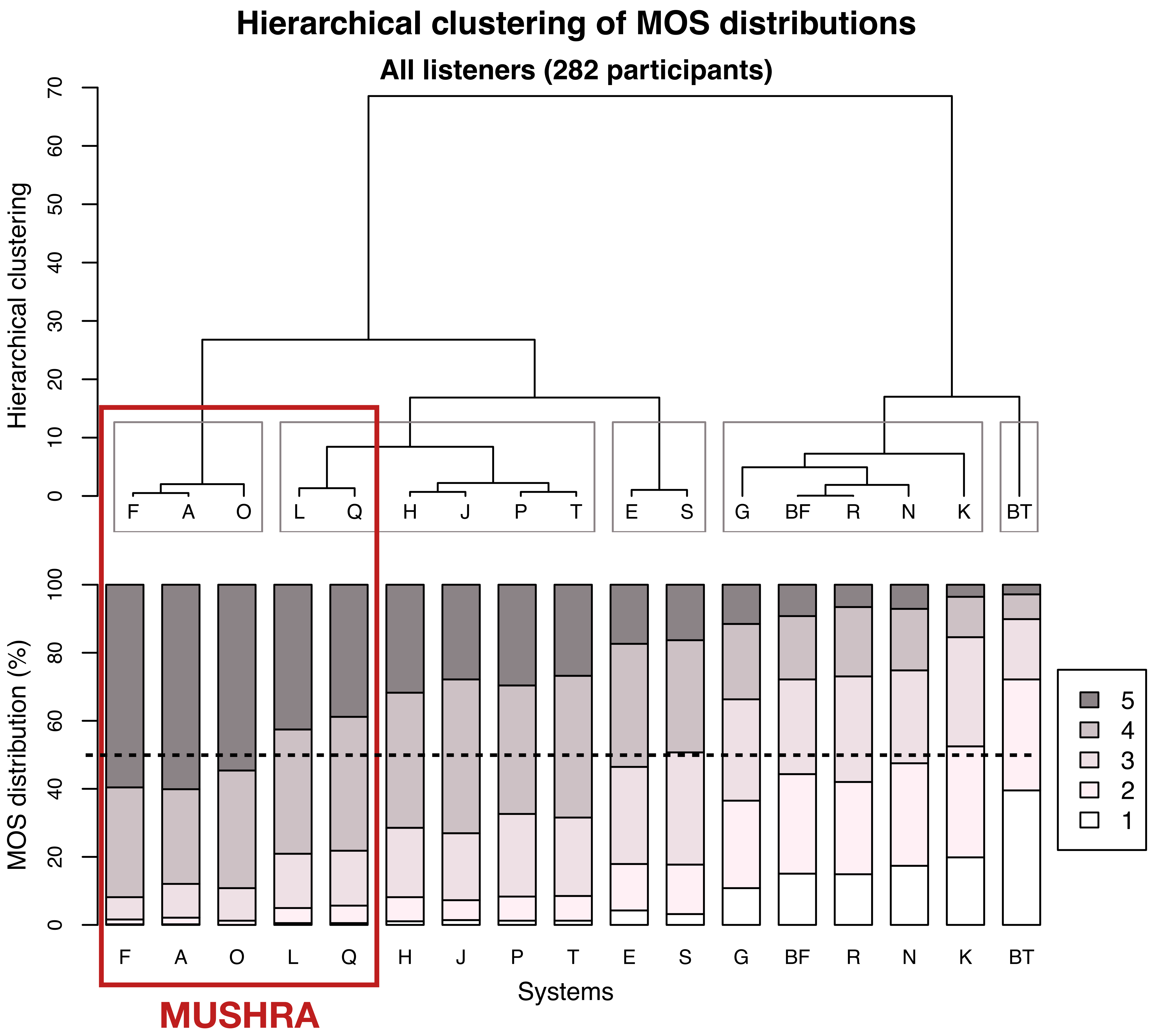


gipsa-lab

Speech quality | Spoke task

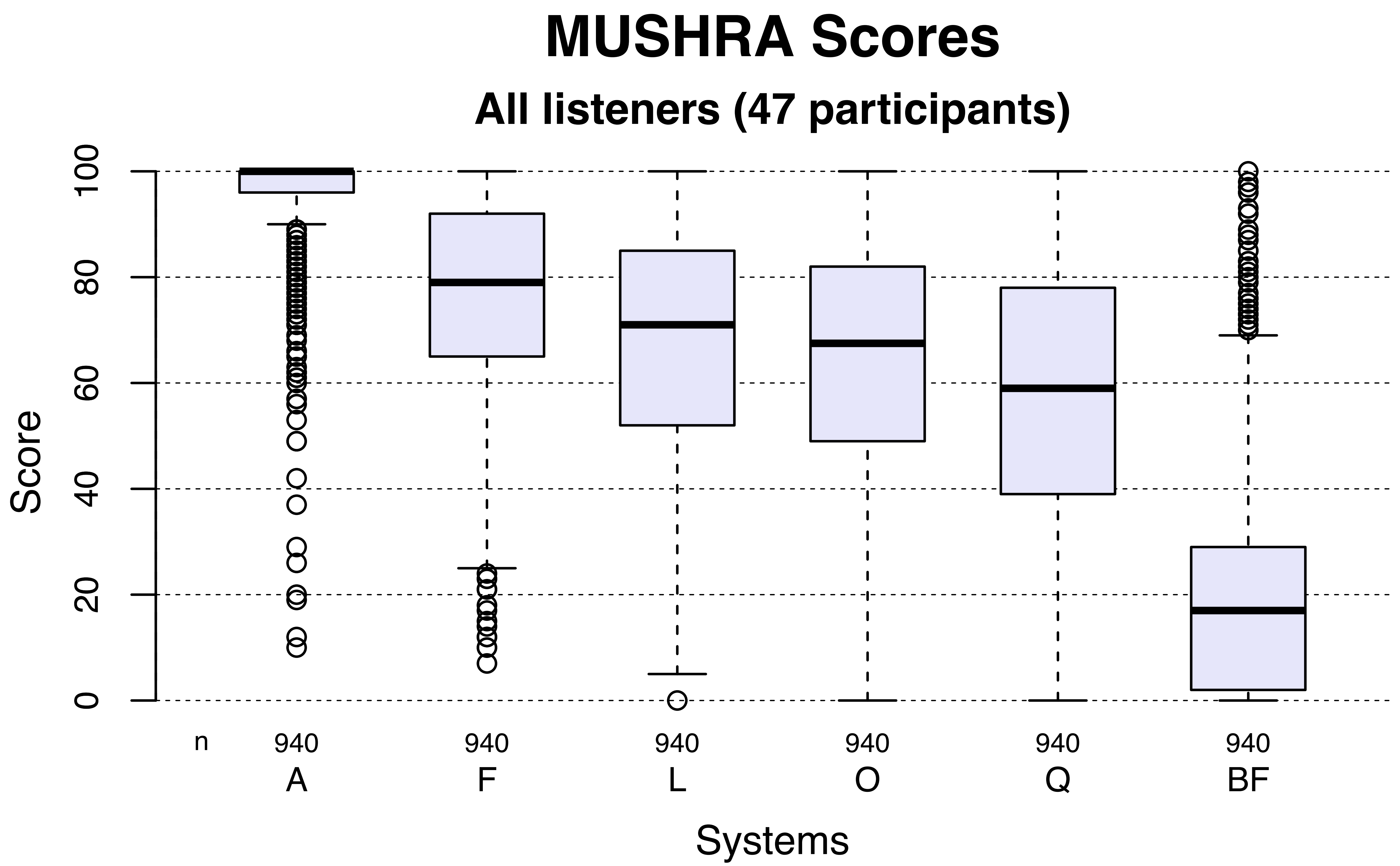
Per system

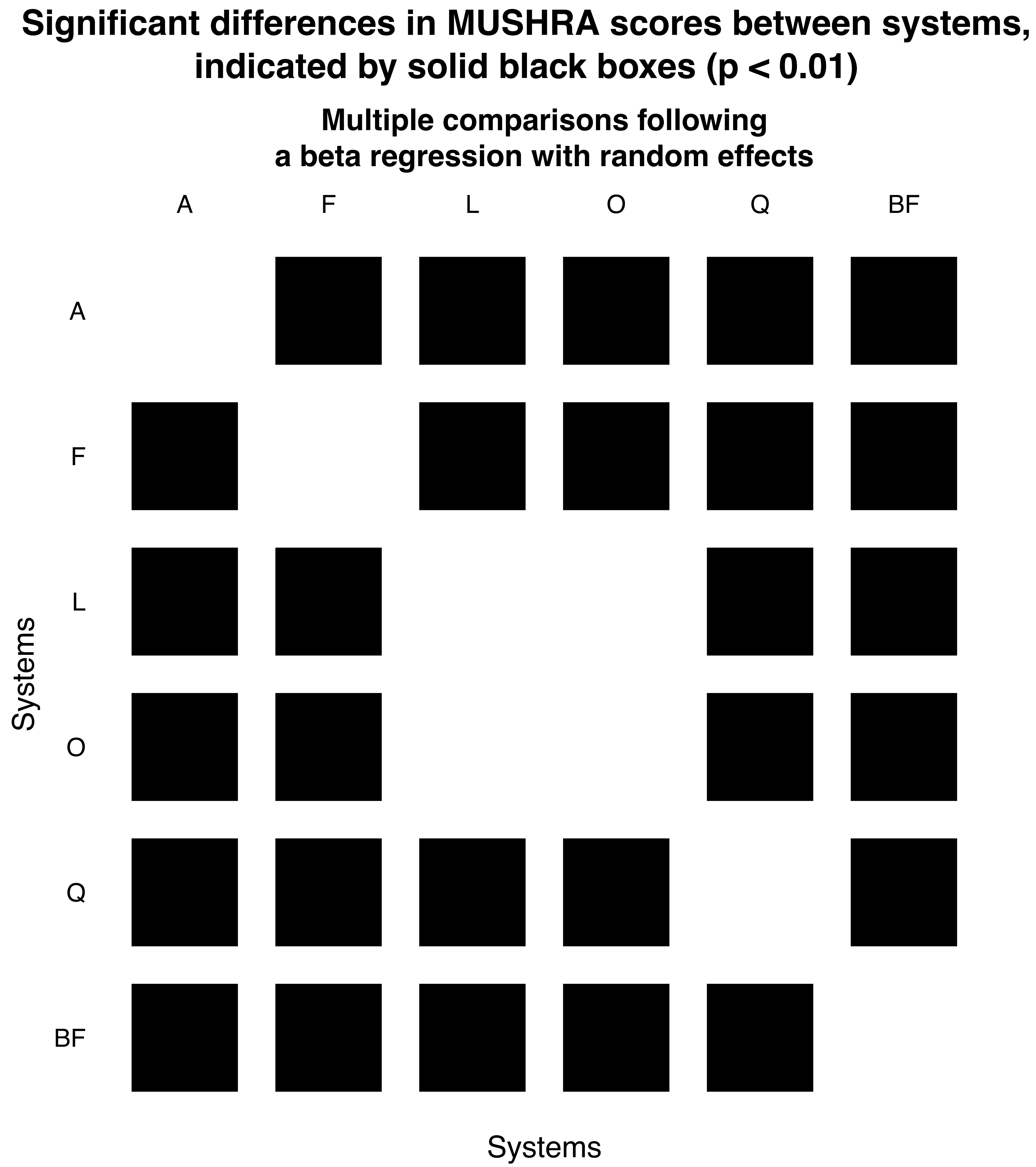




Speech quality | Spoke task

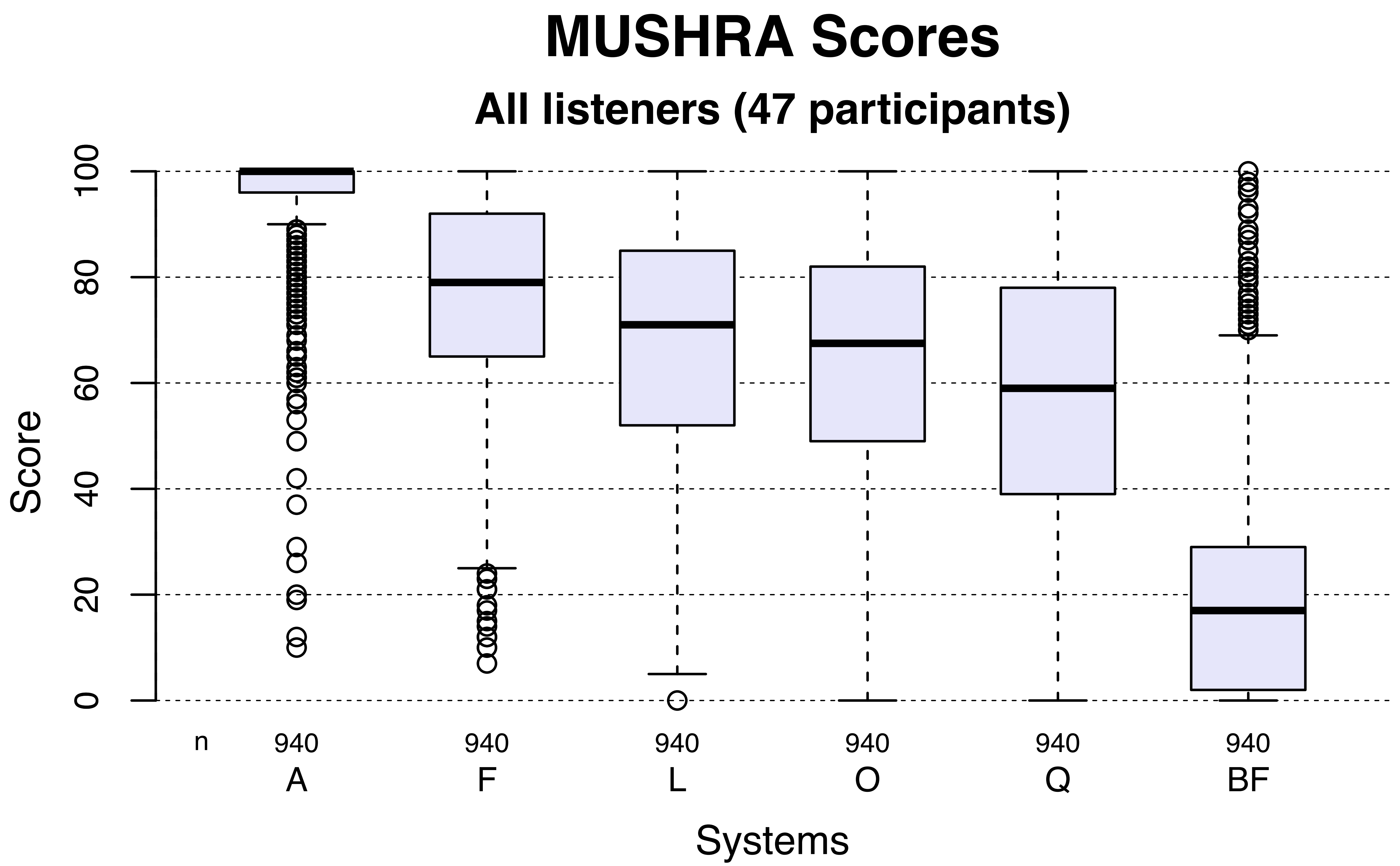
Per system

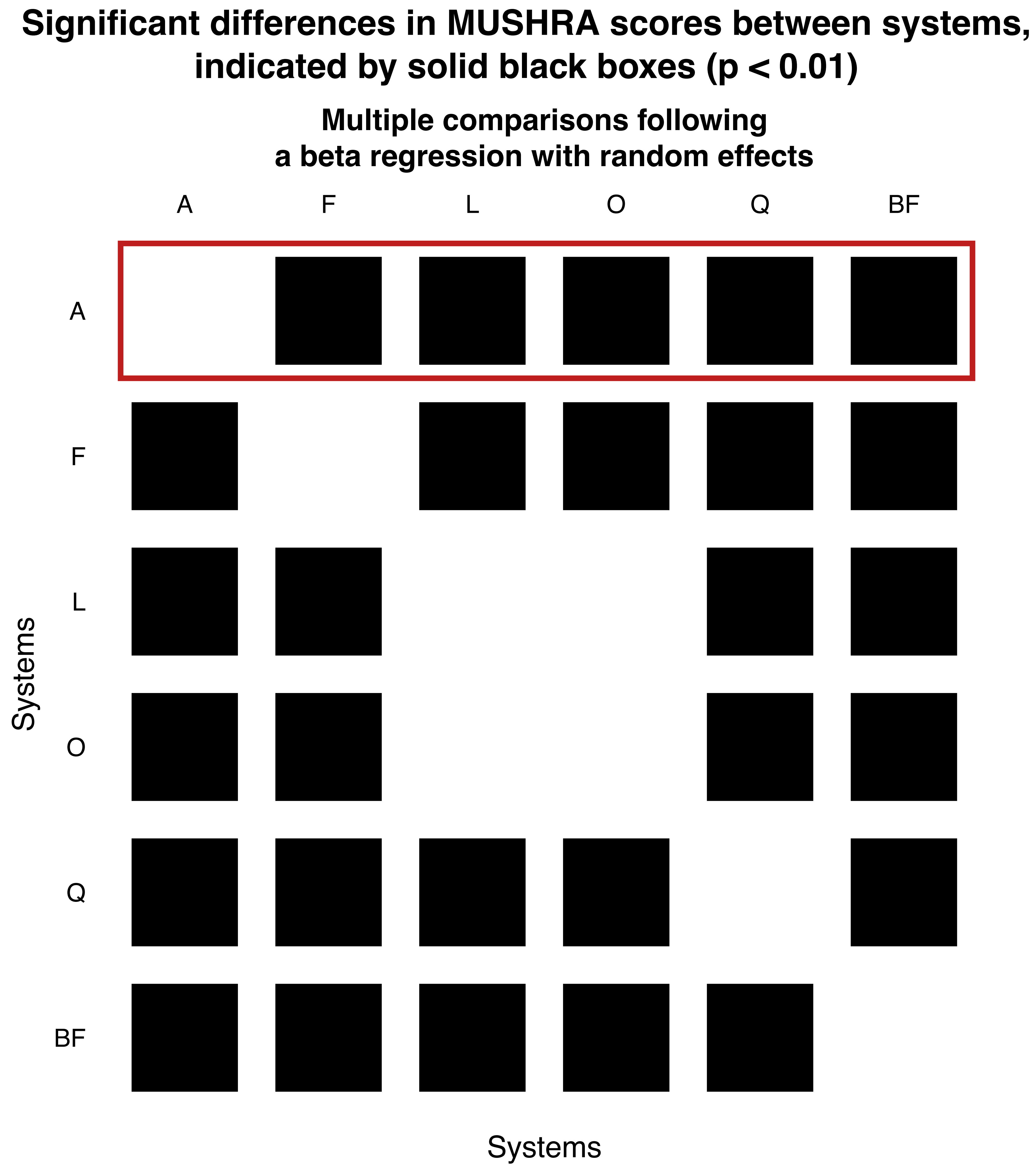




Speech quality | Spoke task

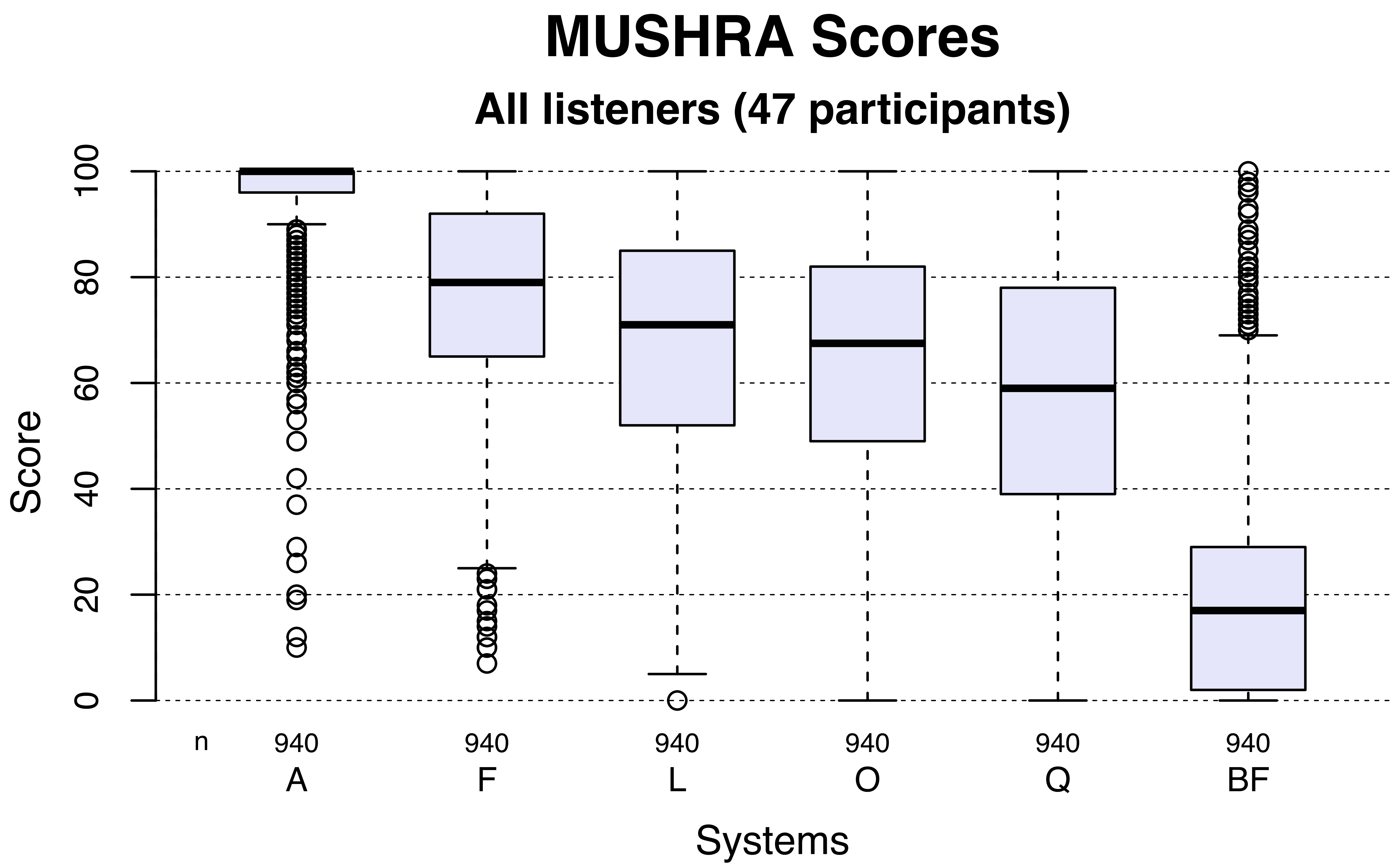
Per system

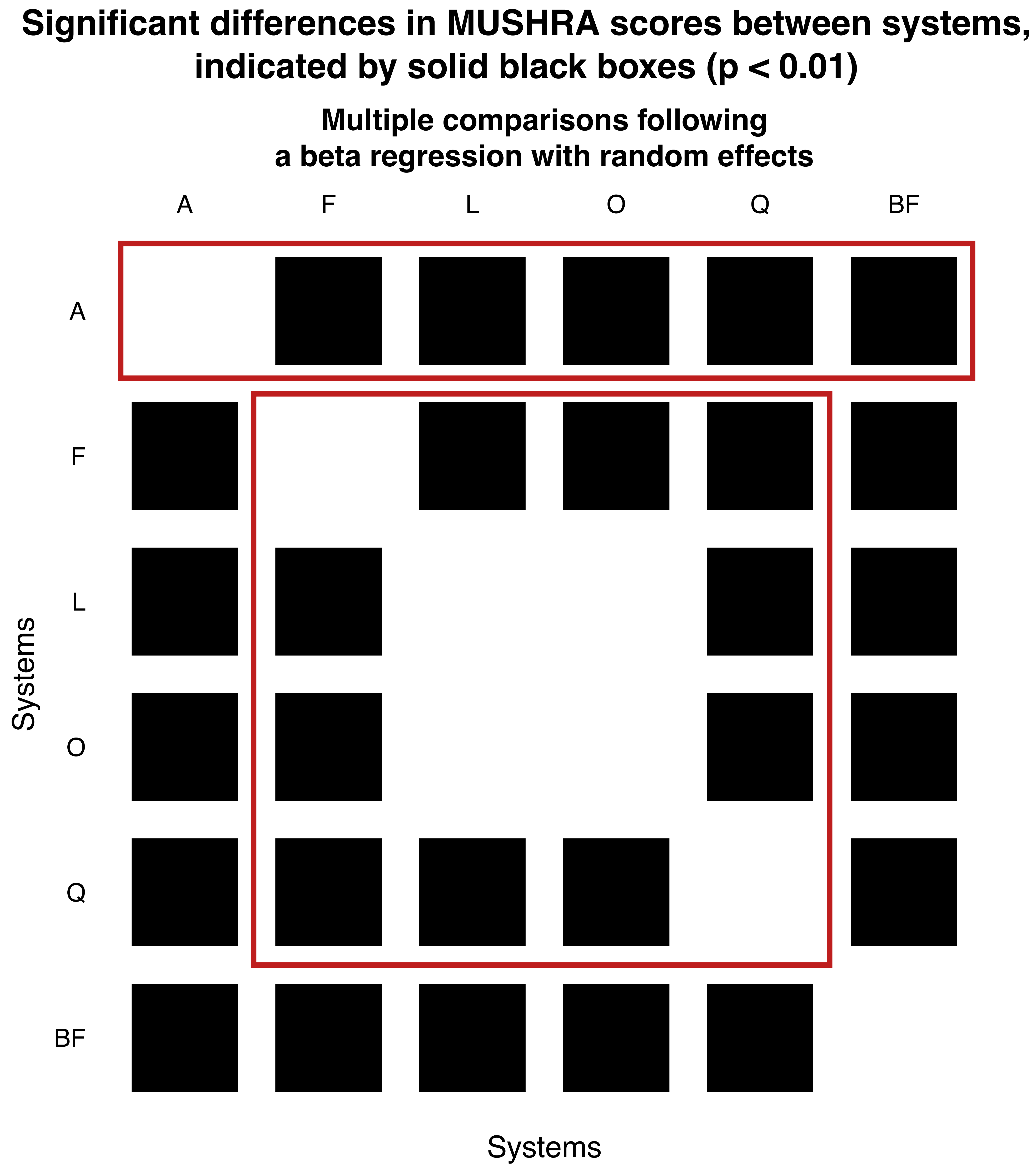




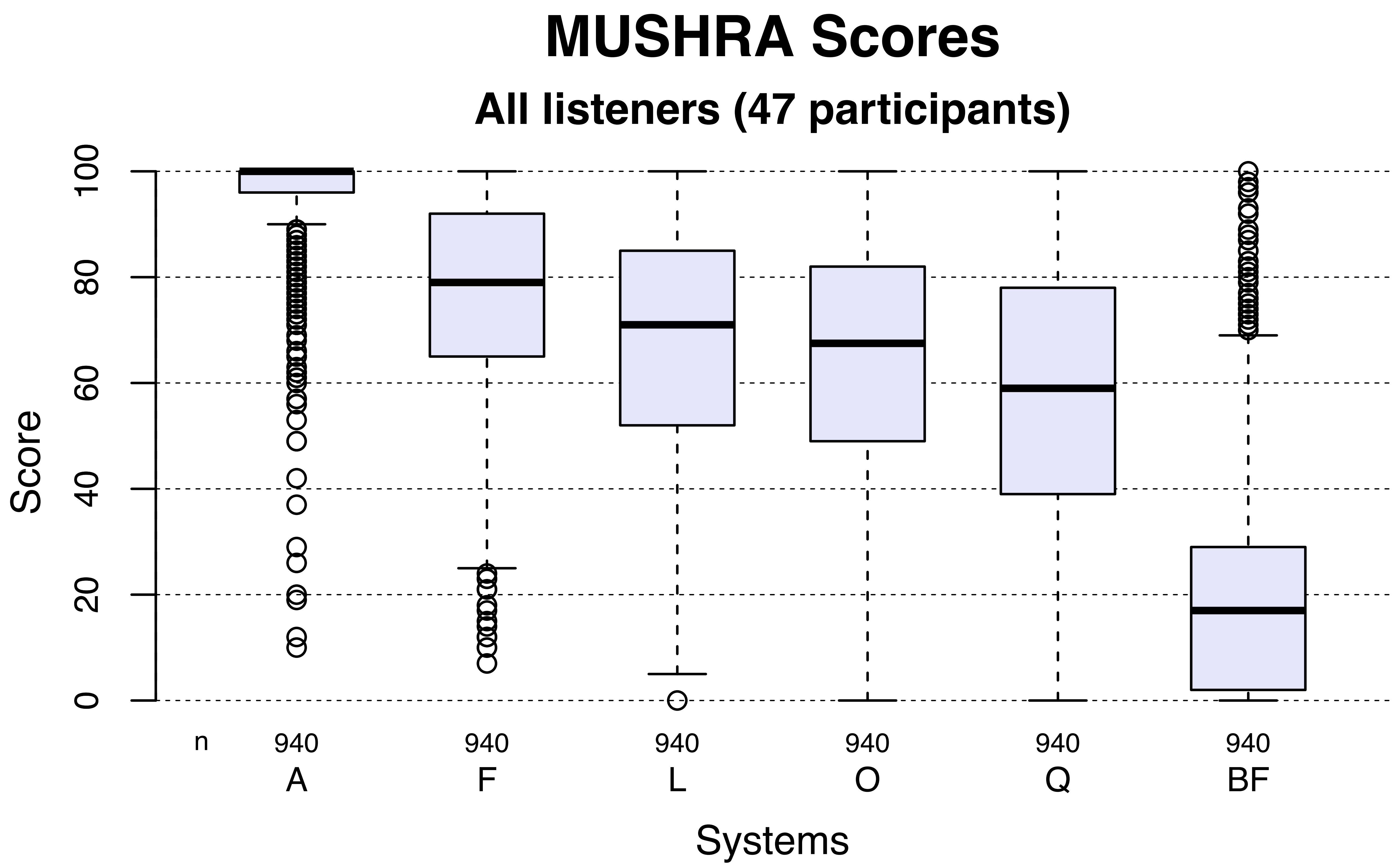
Speech quality | Spoke task

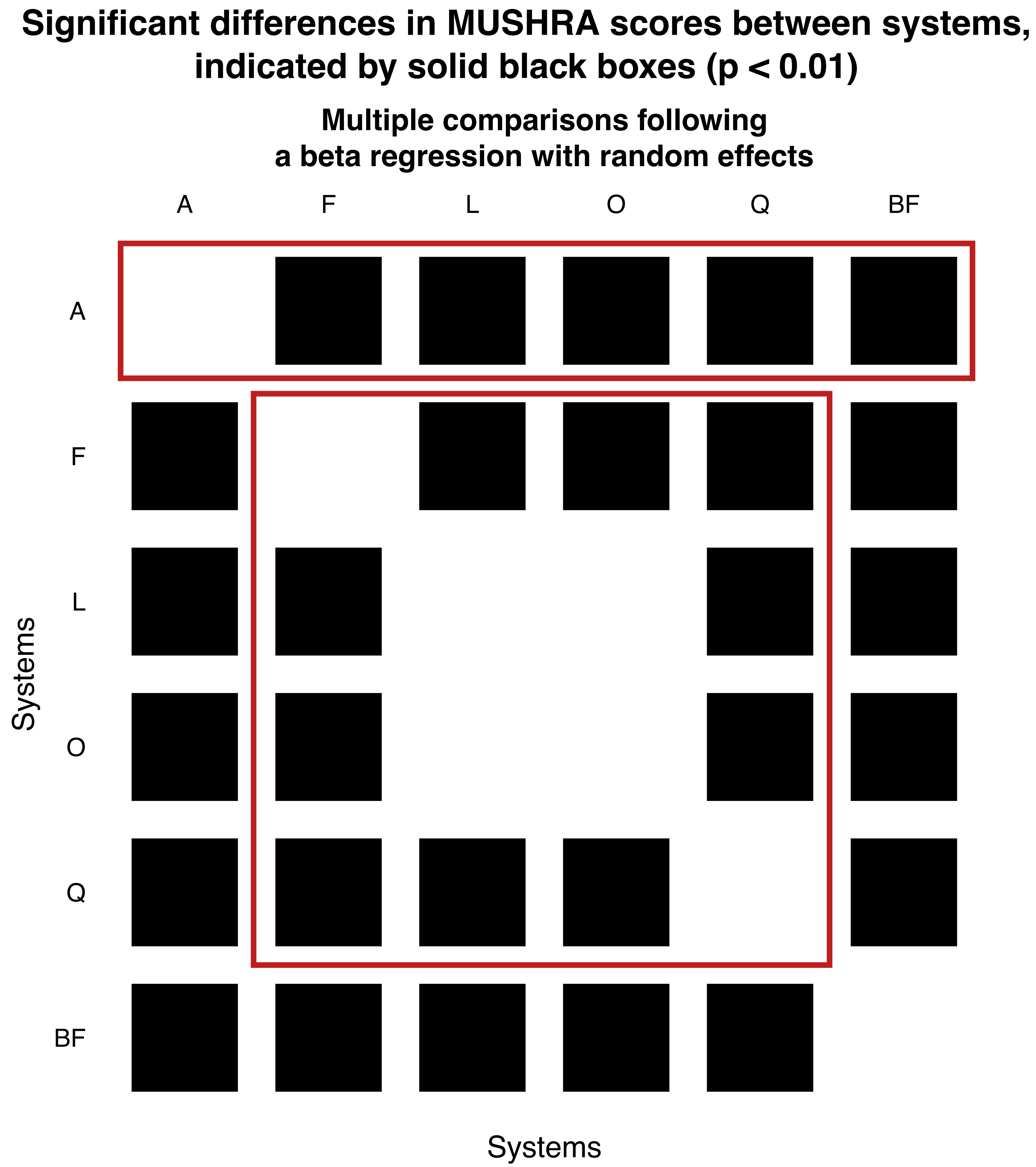
Per system





Speech quality | Spoke task Per system



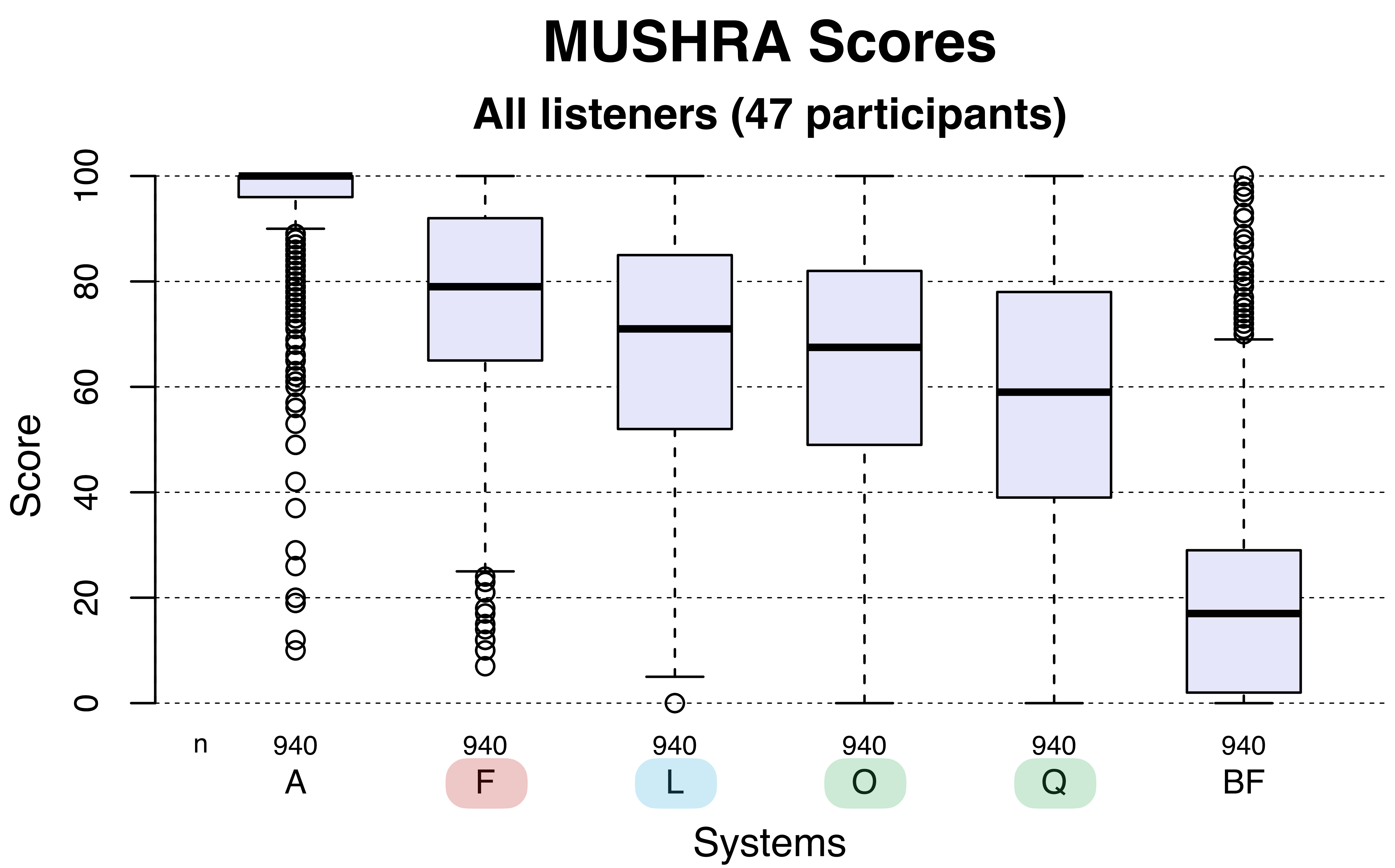


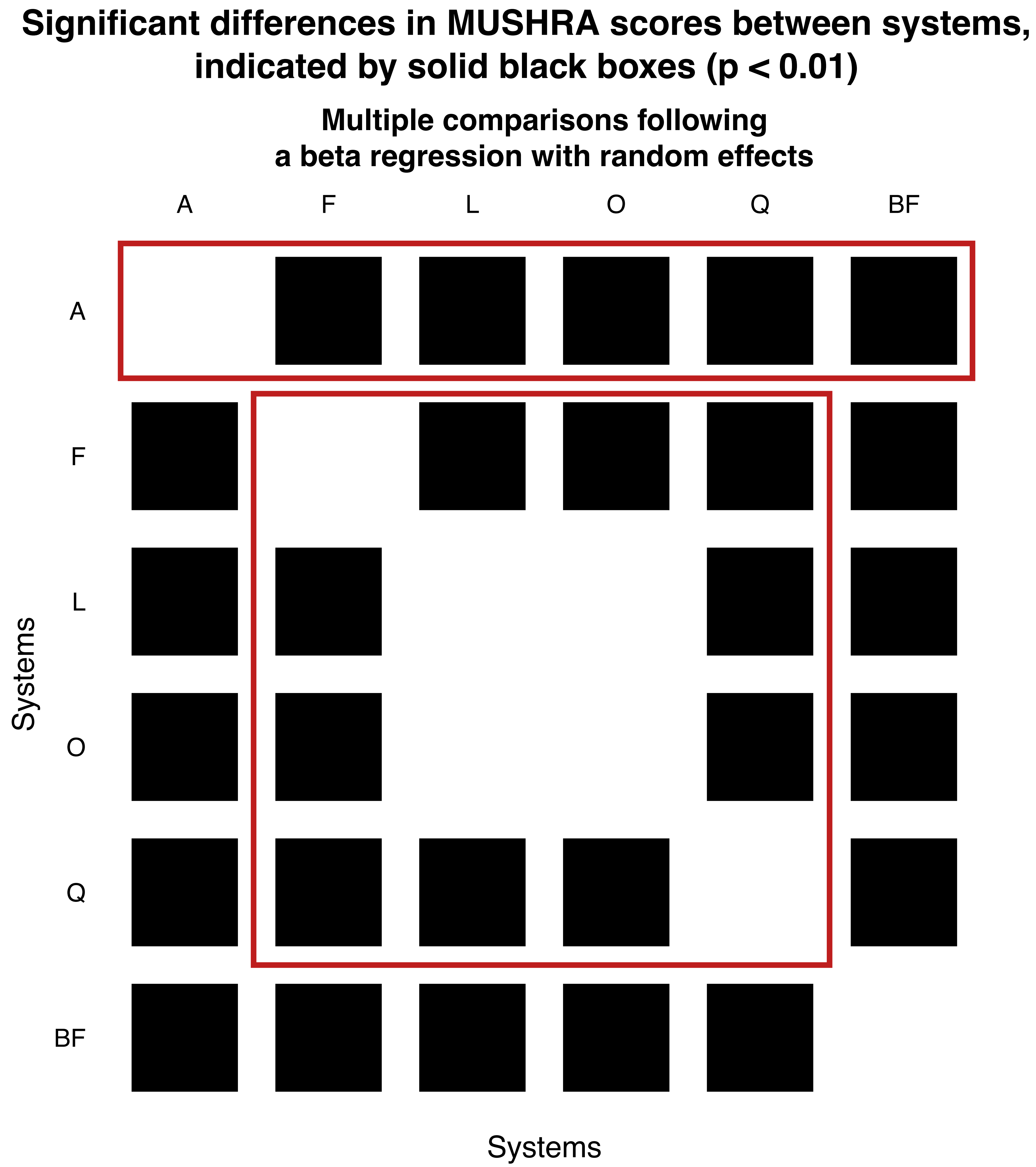
- The MUSHRA highlights the limits of the MOS test when reaching quality close to natural speech

➡ One should **NOT** conclude that the synthetic speech is 'as good as' or 'indistinguishable from' natural speech **in general** from a MOS test

Speech quality | Spoke task

Per system





- The MUSHRA highlights the limits of the MOS test when reaching quality close to natural speech

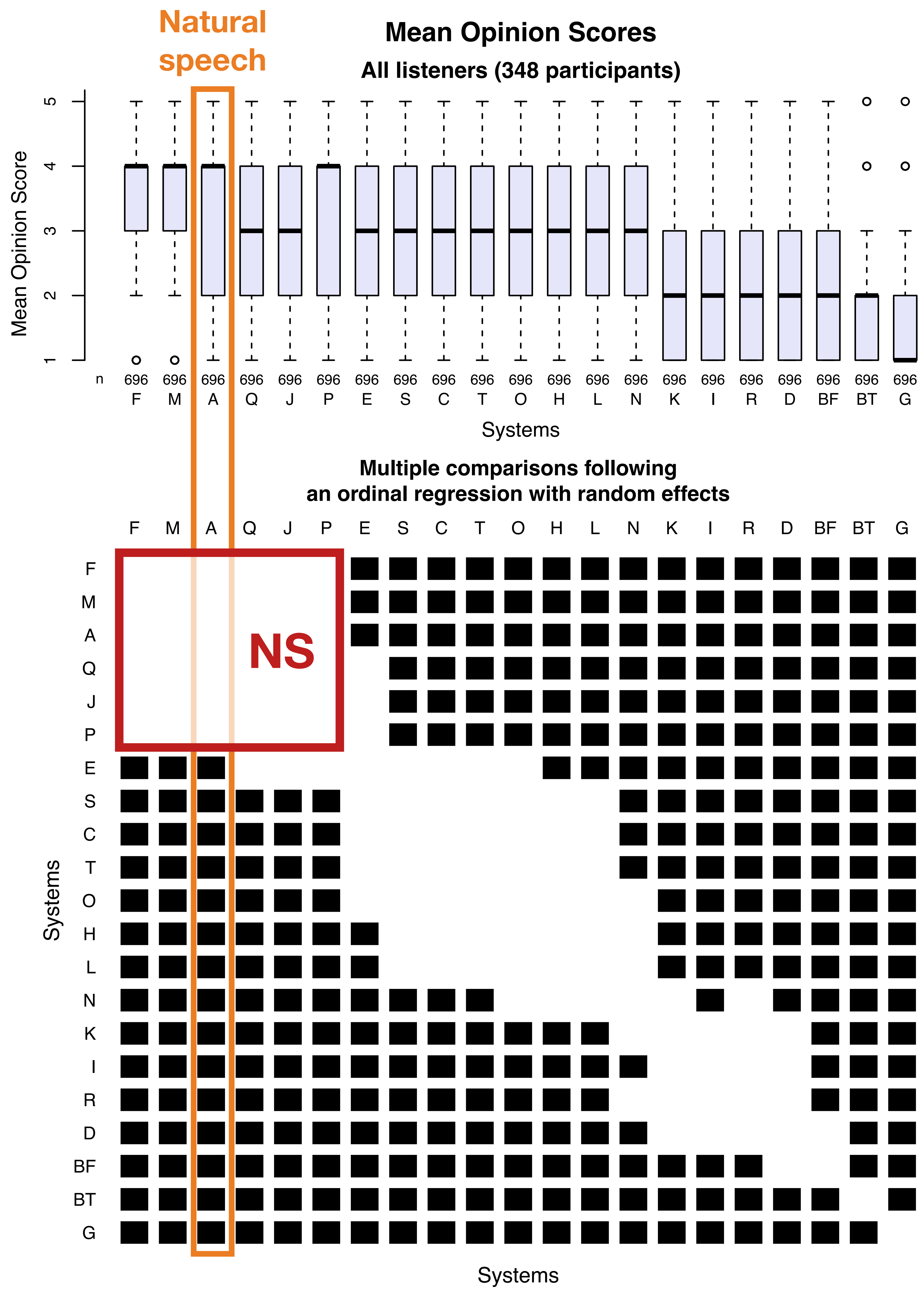
➡ One should **NOT** conclude that the synthetic speech is 'as good as' or 'indistinguishable from' natural speech **in general** from a MOS test

- One of each architecture in the top 4

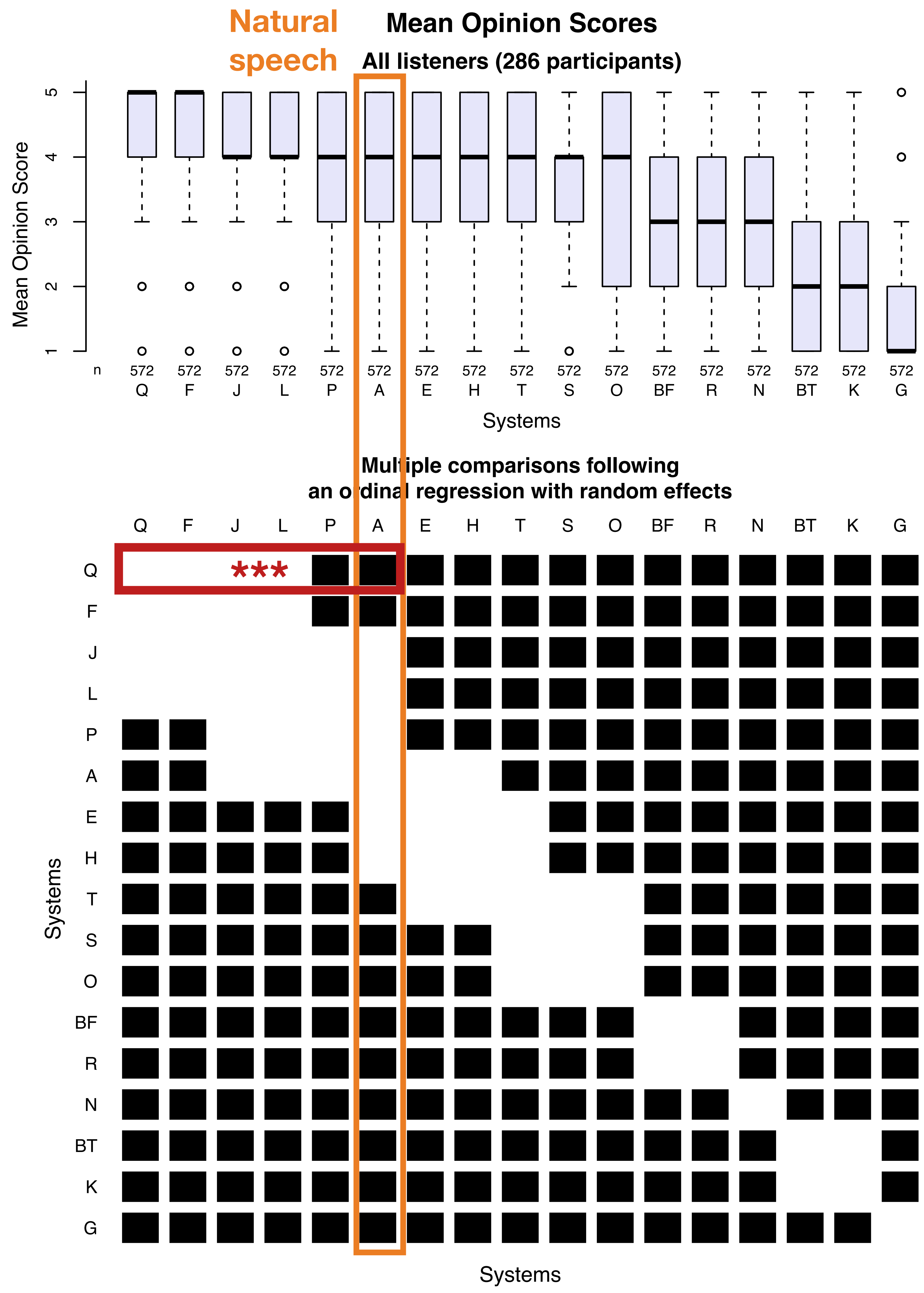
Speaker similarity | Both tasks

Per system

Hub task

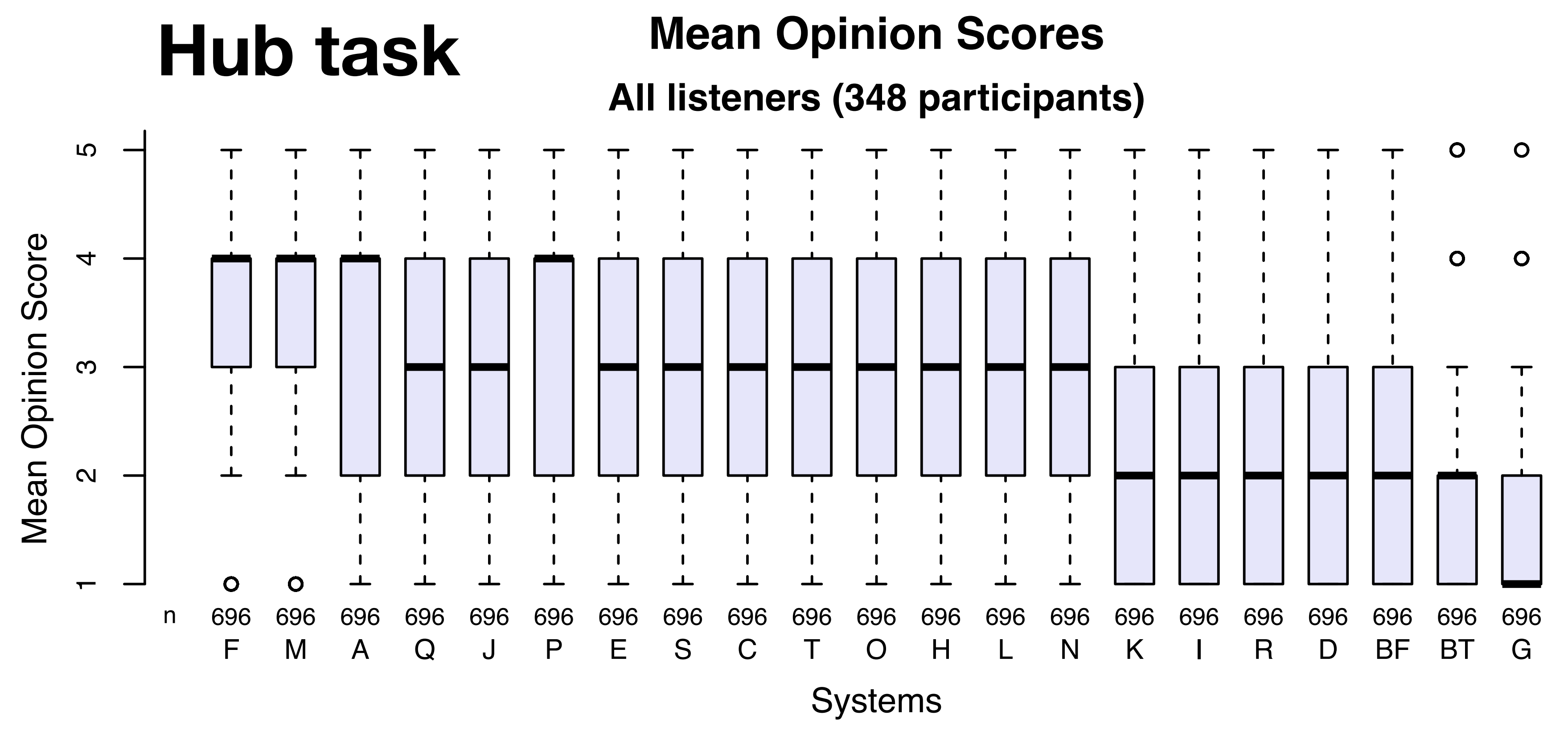


Spoke task



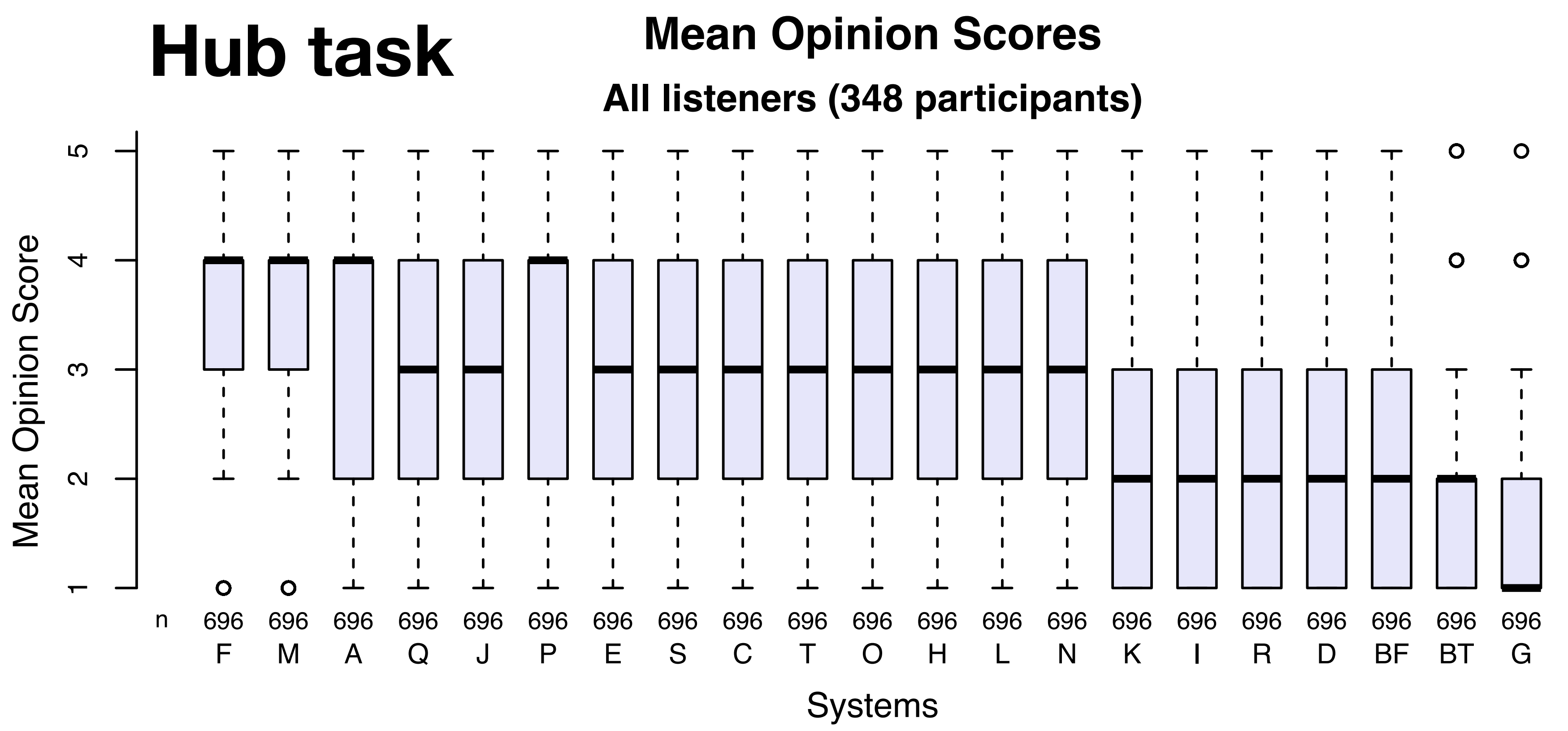
Speaker similarity | Both tasks Per system

- Some listeners and participants to the challenge reported that the reference signals sounded different from each other.
  - Intentional choice, to have reference samples that were representative of the speaker's voice range
  - But few high scores « Exactly the same person » were given for the Hub task



Speaker similarity | Both tasks   Per system

- Some listeners and participants to the challenge reported that the reference signals sounded different from each other.
  - Intentional choice, to have reference samples that were representative of the speaker's voice range
  - But few high scores « Exactly the same person » were given for the Hub task



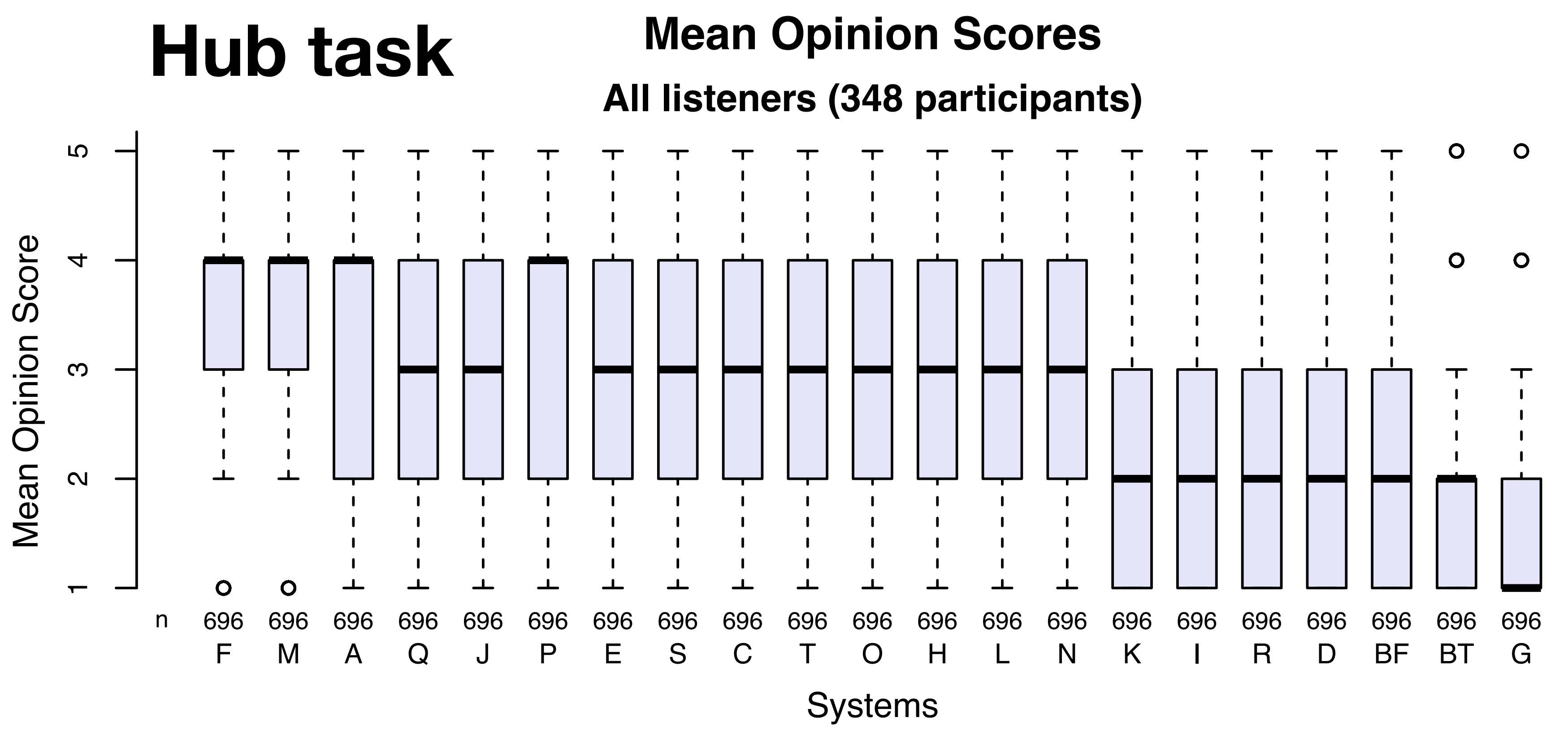
- What is speaker similarity?
  - Similarity to references which are in the centre of the distribution of the speaker's voice range of variation, to which the syntheses might be close
  - Similarity to references that are representative of the speaker's full voice range, with wide timbre variations
  - ➡ We chose the second option, more ecological speaker recognition task

Speaker similarity | Both tasks Per system

- Some listeners and participants to the challenge reported that the reference signals sounded different from each other.
  - Intentional choice, to have reference samples that were representative of the speaker's voice range
  - But few high scores « Exactly the same person » were given for the Hub task
- What is speaker similarity?
  - Similarity to references which are in the centre of the distribution of the speaker's voice range of variation, to which the syntheses might be close
  - Similarity to references that are representative of the speaker's full voice range, with wide timbre variations

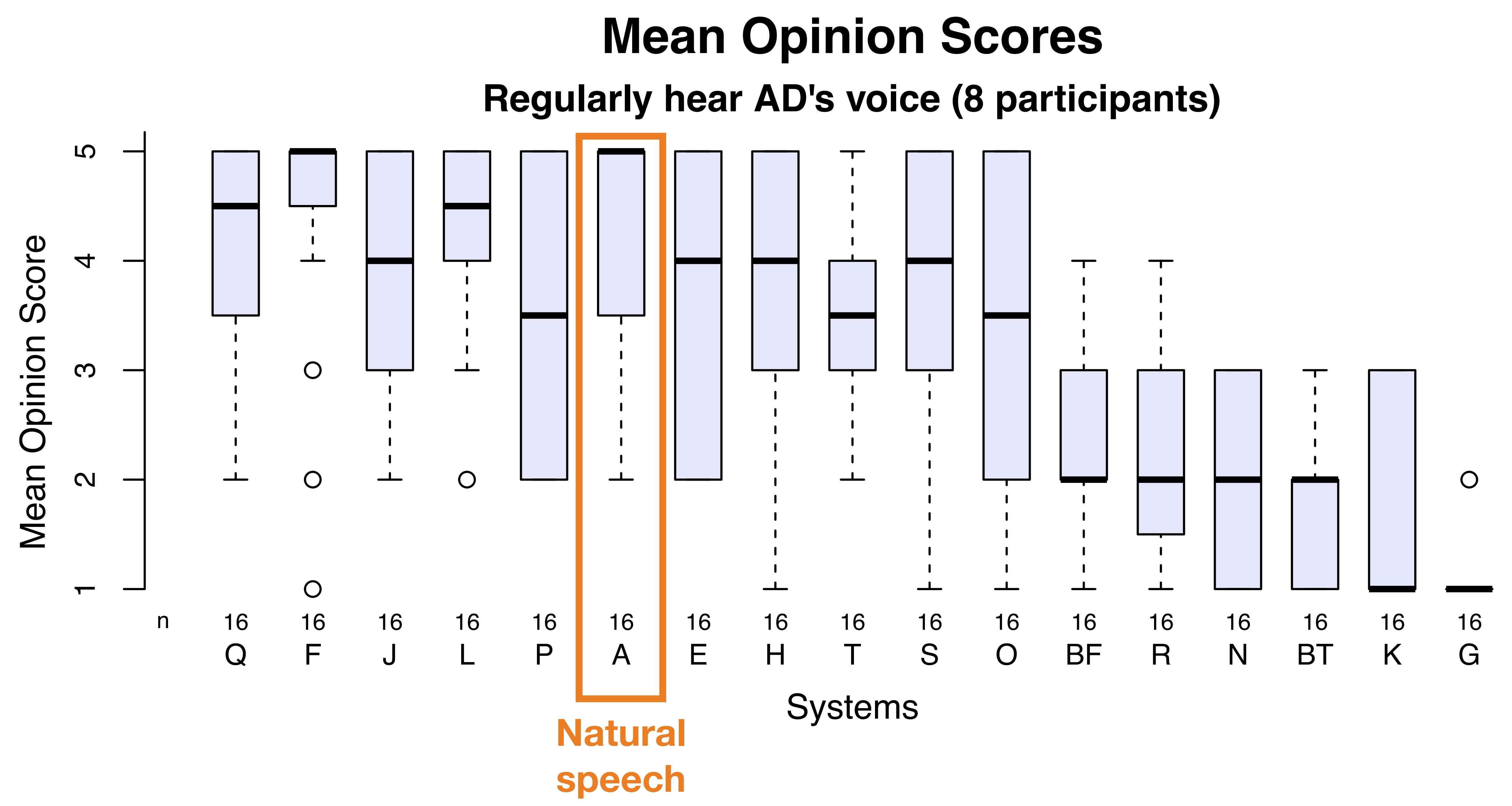
  ➡ We chose the second option, more ecological speaker recognition task
- In that case, can we ask listeners who have never heard the voice of the reference speaker before, if a sound sample could come from his/her voice?

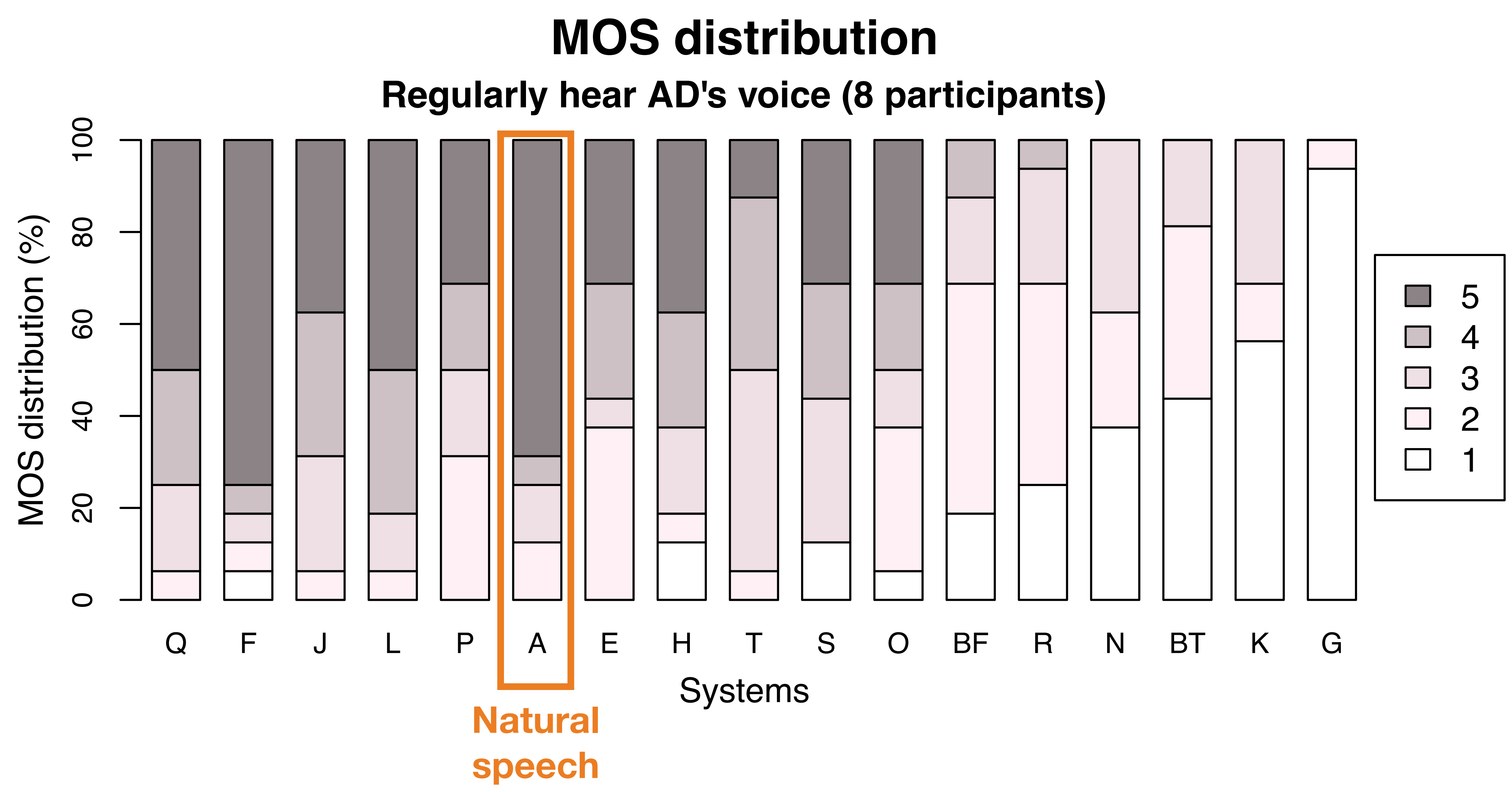
  ➡ Low score for natural speech suggest that they cannot create a mental representation of the speaker's full voice range



Speaker similarity | Both tasks   Per system

Spoke task (FS1) I Similarity assessment





- Listeners who are familiar with AD's voice (family and friends)
- Natural voice rated as « Exactly the same person in > 70% of the time.
- Only system F is equivalent, consistent with its high quality rating

gipsa-lab cnrs INP UGA

Speaker similarity | Both tasks Per system

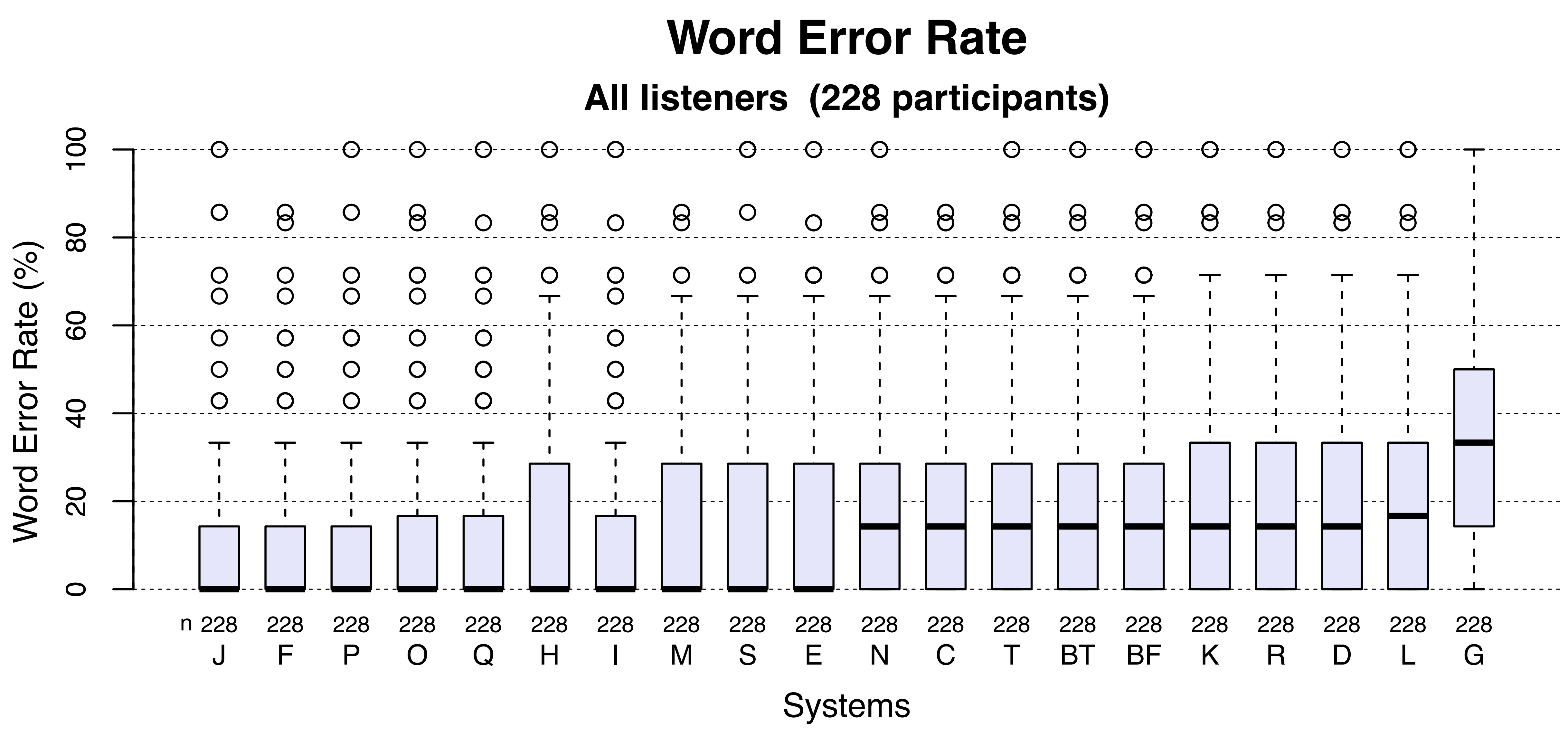
- Listeners that are **familiar** with the speaker's voice are able to correctly perform the speaker similarity task on the ground truth signal where the **references given have a wide range of variation**

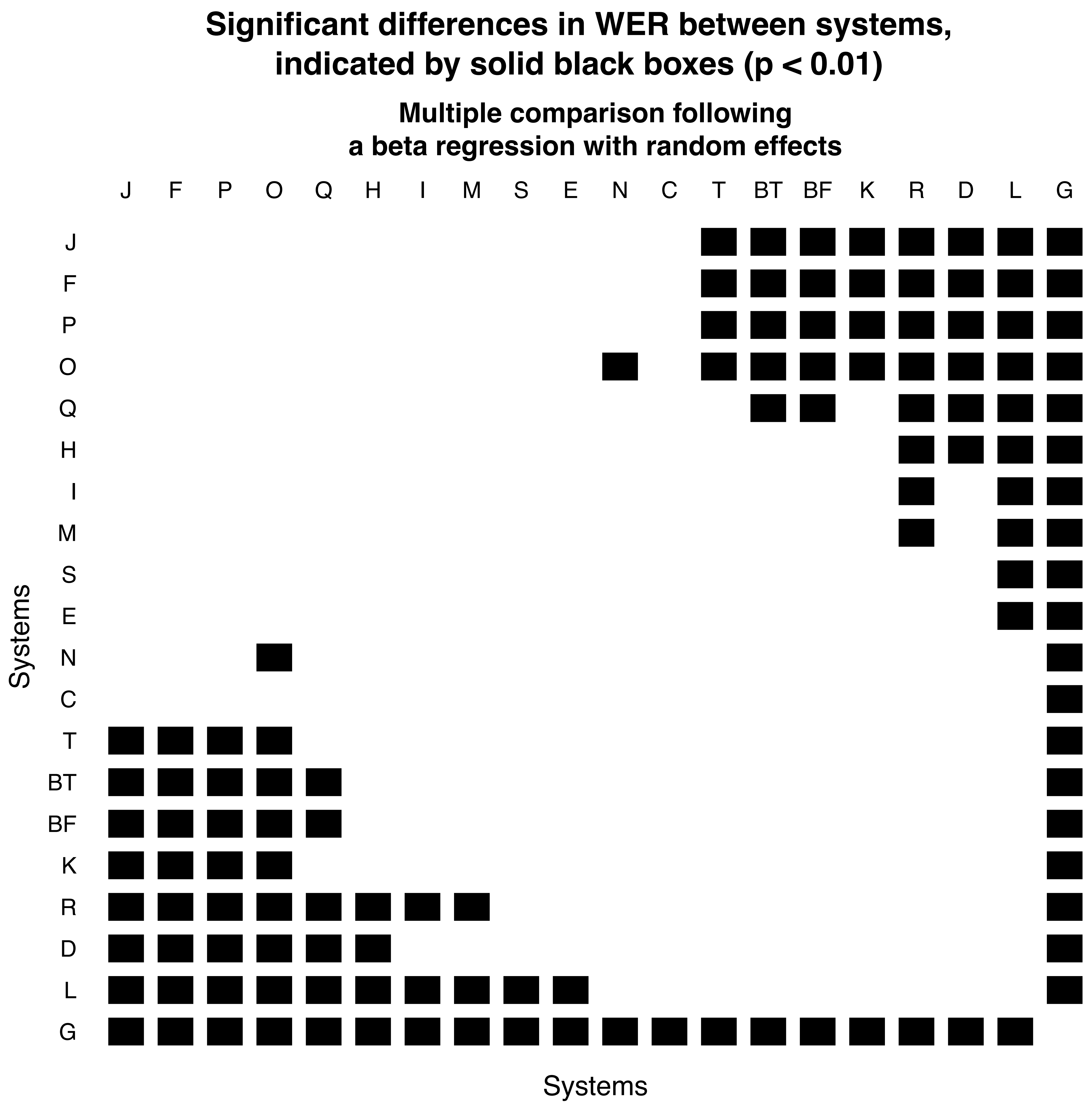
- Listeners that are **not familiar** with the speaker's voice may only be able to perform a speaker similarity task where the **reference given is in the centre of the distribution** of the speaker's voice range of variation

➡ **Redefinition of the speaker similarity task?**

Intelligibility | SUS

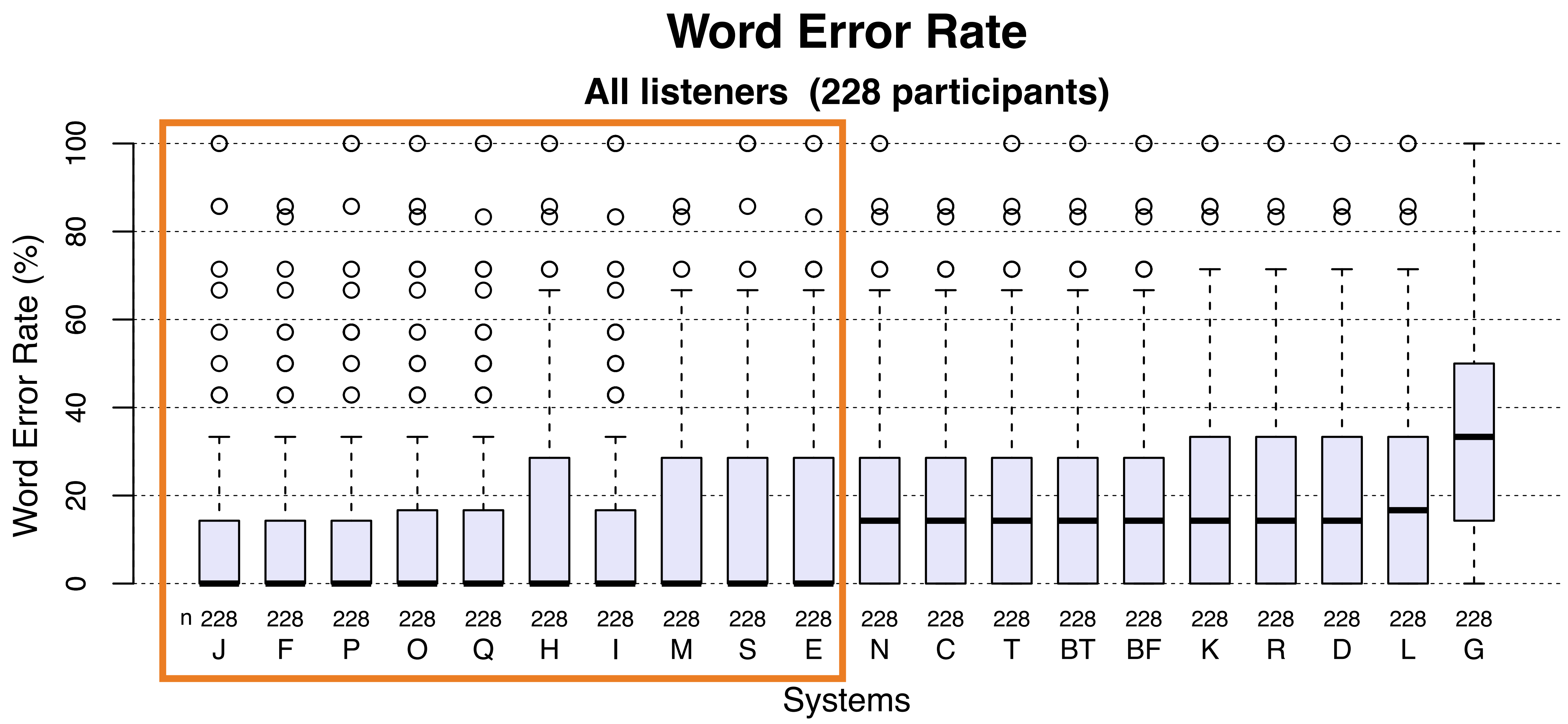
Per system



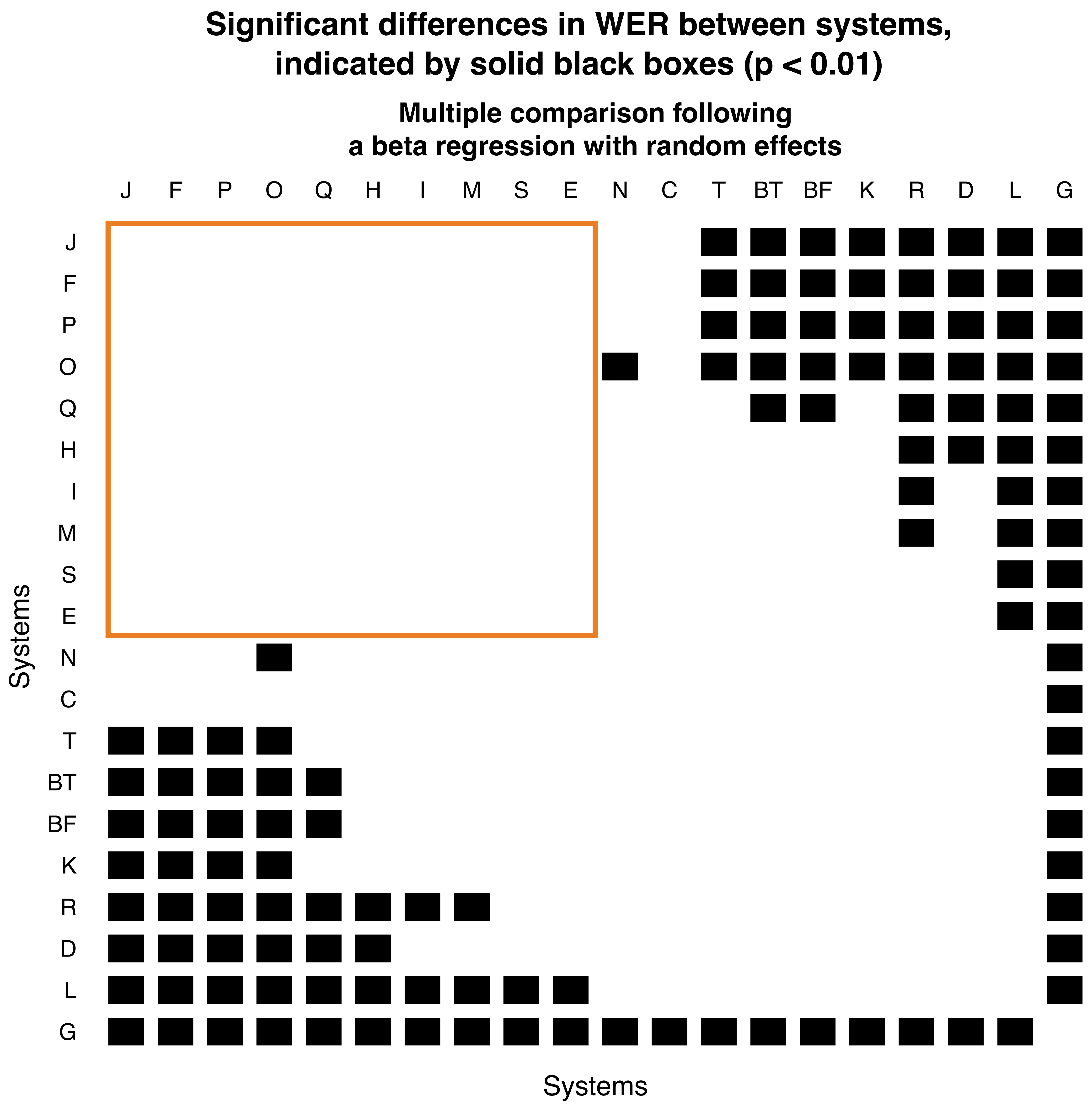


Intelligibility | SUS

Per system



Median of 0

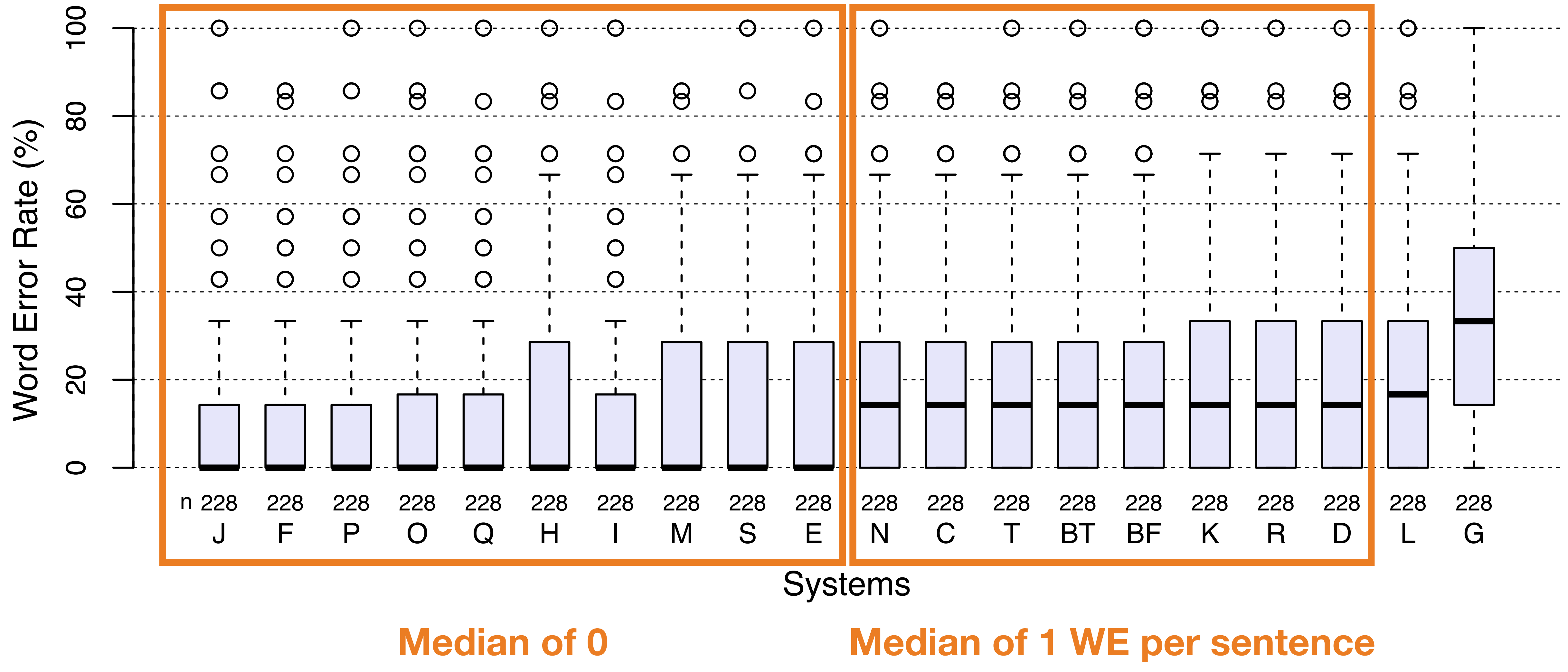


Intelligibility | SUS

Per system

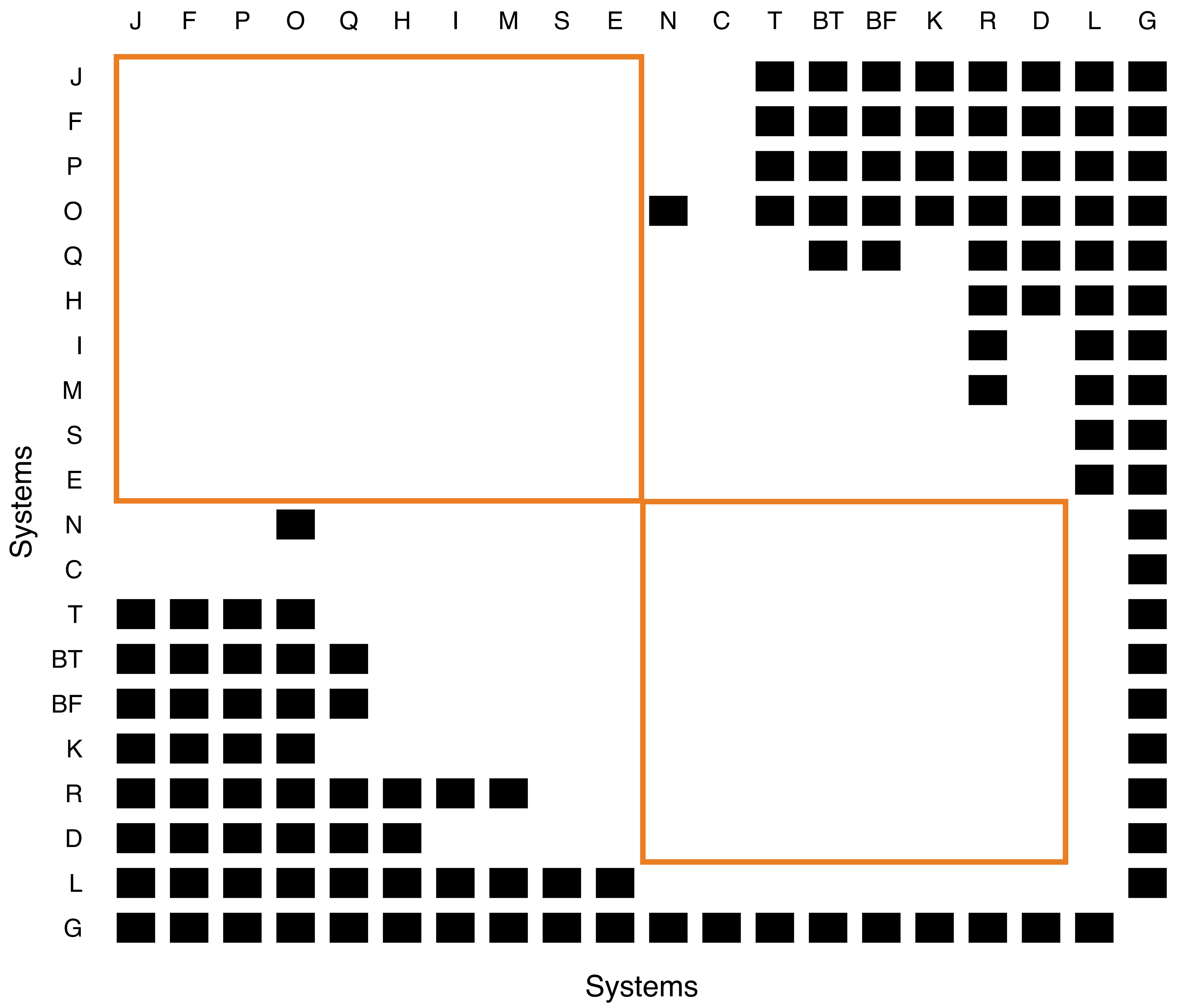
Word Error Rate

All listeners (228 participants)



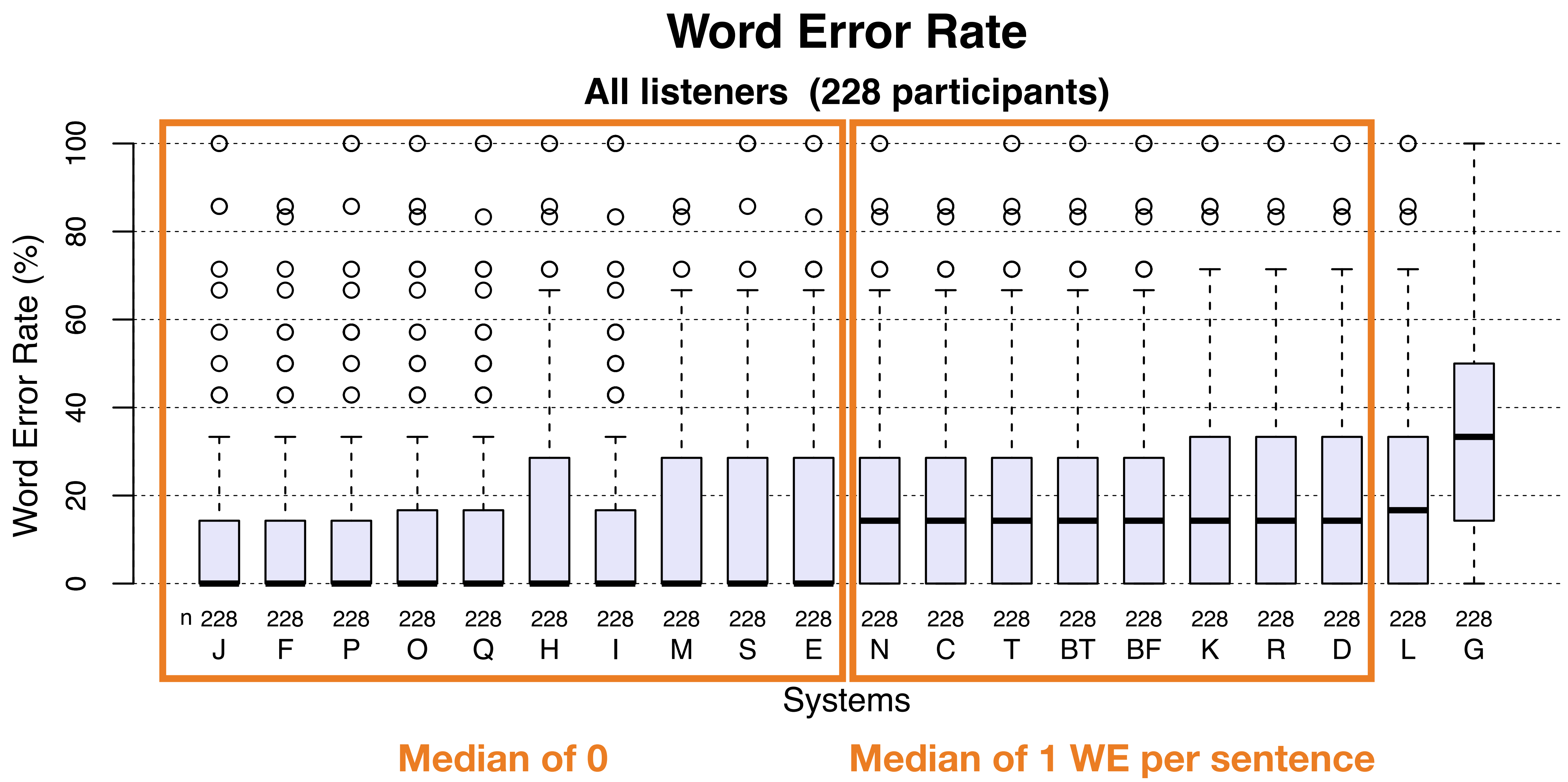
Significant differences in WER between systems, indicated by solid black boxes (p < 0.01)

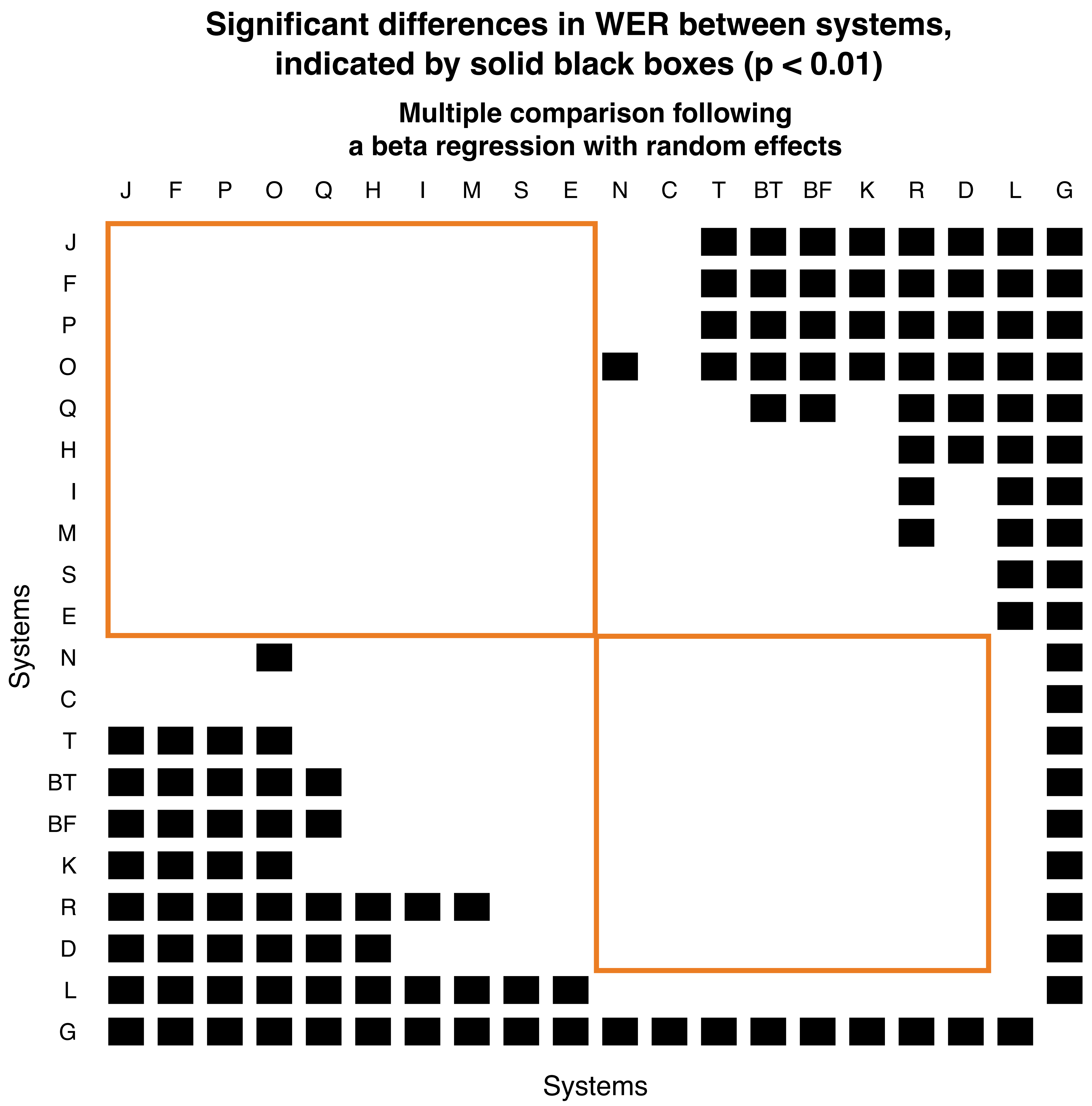
Multiple comparison following a beta regression with random effects



Intelligibility | SUS

Per system





- SUS synthesis globally well handled

➡ Need for finer tasks for the evaluation of intelligibility

Intelligibility | Homographs

Per system

gipsa-lab cnrs INP UGA

Intelligibility | Homographs

Per system

Intelligibility | Homographs

Per system

Intelligibility | Homographs

Per system

Use of a Large Language Model

Small jump of performance

➡ Only systems which used a LLM made few errors in synthesising homographs

Results per factor

Effect of:

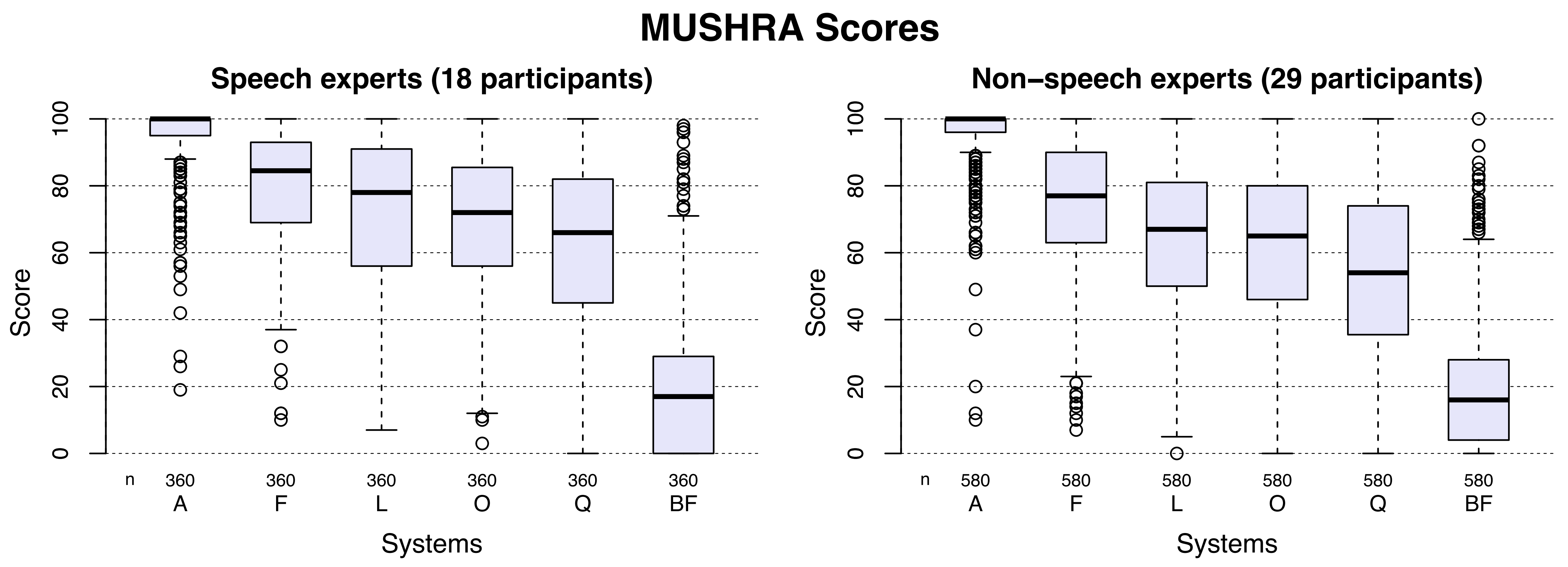
- systems x speech expertise
- systems x is native

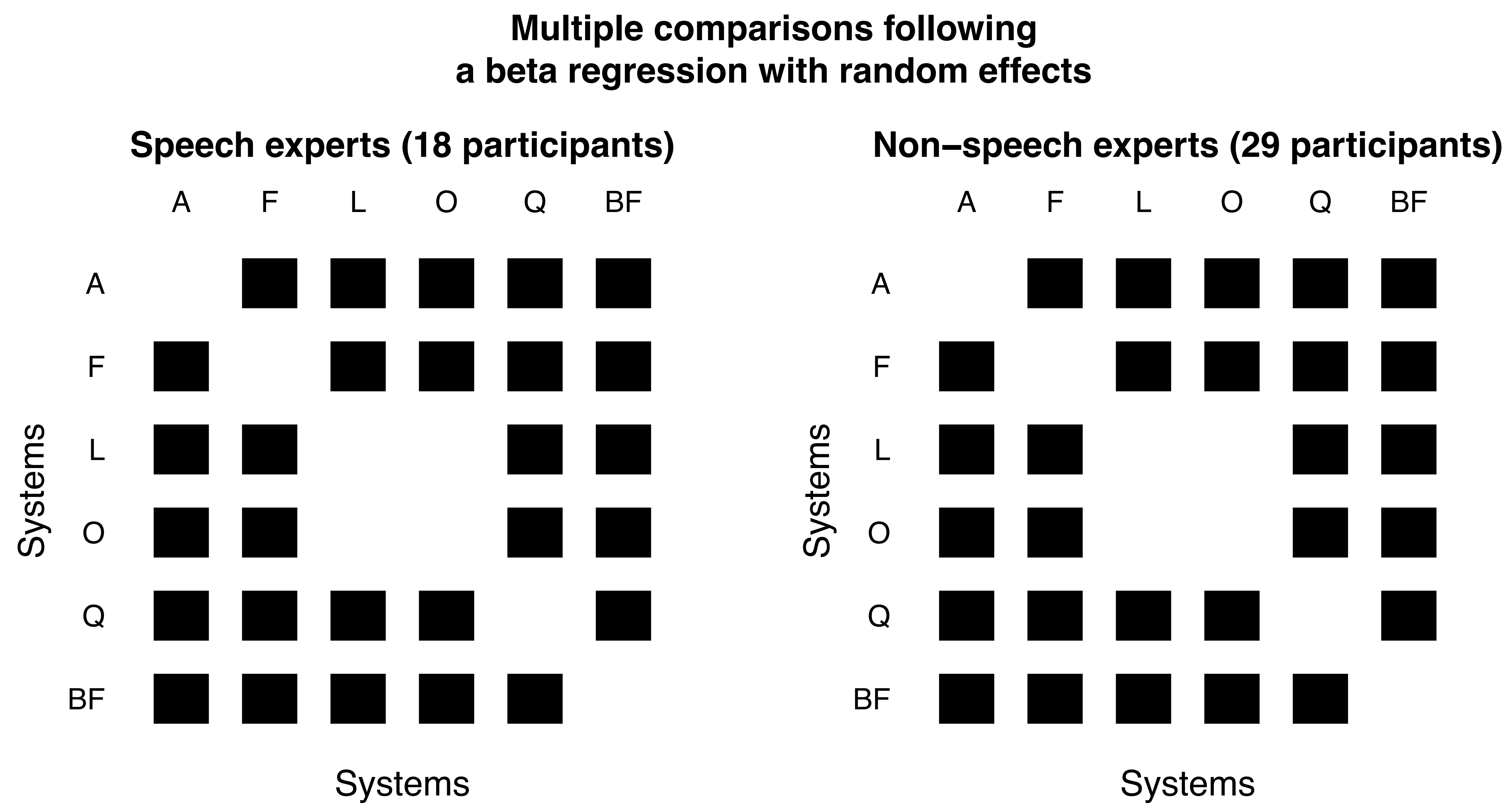
Speech quality | Spoke task

Effect of speech expertise

Significant effect of speech expertise

- Non-speech experts gave lower scores than speech experts
- Does not affect the relative difference between systems
- Does not affect the significance of the differences between systems



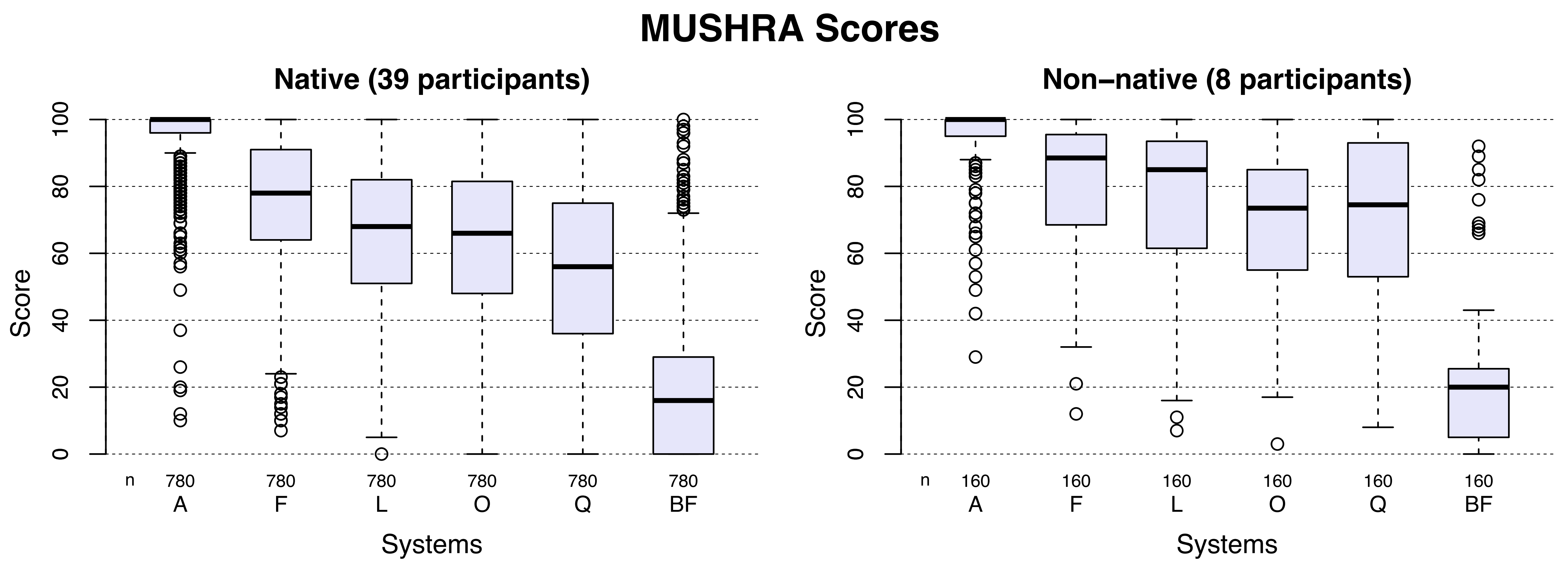


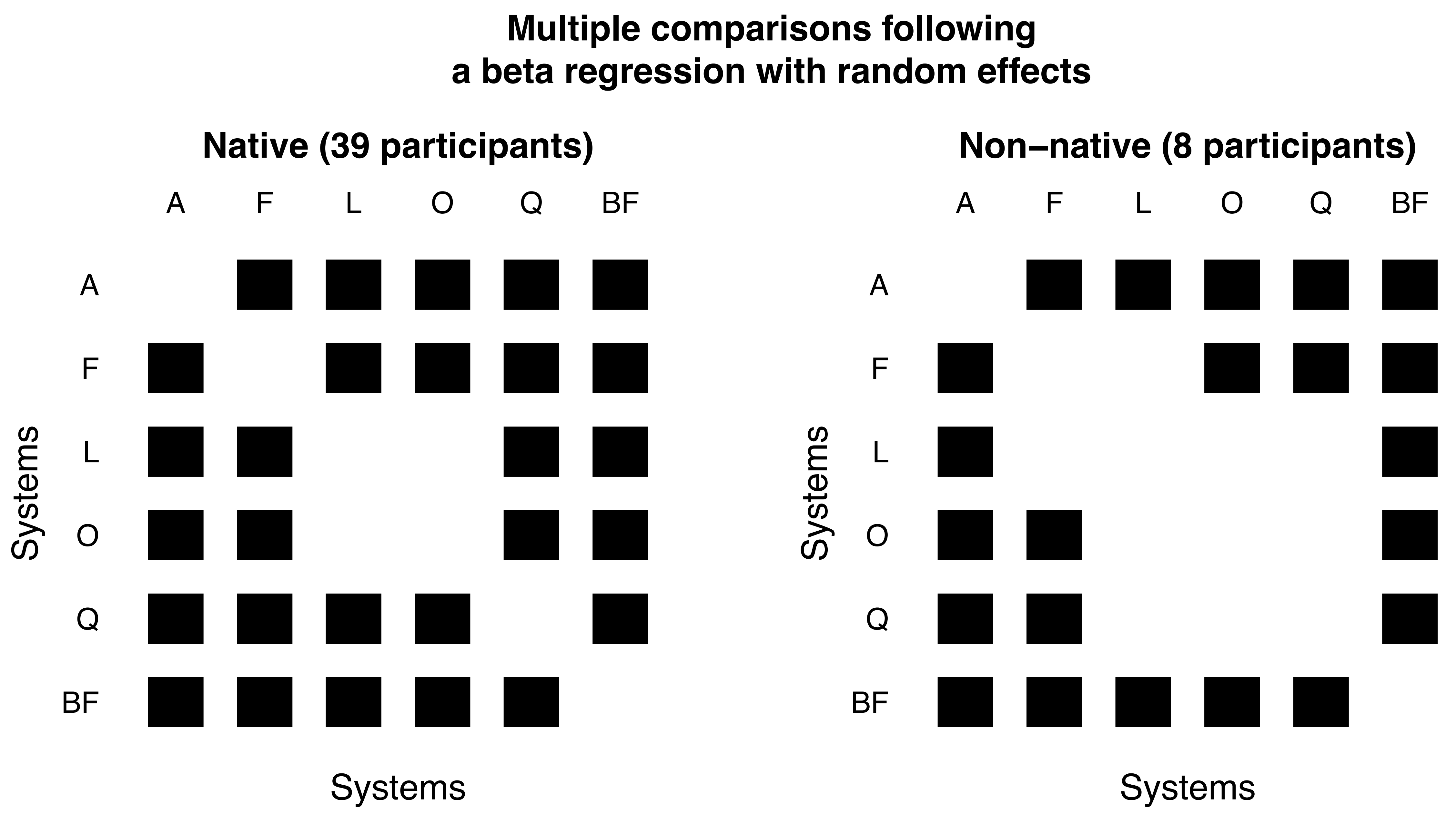
Speech quality | Spoke task

Effect of is native factor

Significant effect of is native factor

- Native listeners gave lower scores than non-natives
- Does affect the relative difference between systems
- Non-native listeners perceived less significant differences between systems than natives





Effects of speech expertise and is native factors

Similar behaviours on the MOS scores as on the MUSHRA scores

- Speech expertise
  - Significant but **small** effects on the results
  - Slightly better scores given by speech experts
  - Similar pairwise differences between systems for experts and non-experts

- Is native
  - Significant and **important** effects on the results
  - Lower scores given by native listeners
  - Much less pairwise differences perceived by non-native listeners

➡ Importance of having native listeners, even for non-intelligibility tests (speech quality, speaker similarity)

Conclusion

Conclusions

- Systems: all DNN
  - Acoustic model: 11 FastSpeech-like or non-attentive Tacotron-like design ; 7 VAE conditioned by text
  - Vocoder: 15 GAN-based models
- Speech quality evaluation
  - Some systems are indistinguishable from natural speech in MOS conditions (not MUSHRA)
  - All acoustic model types performed well
- Speaker similarity evaluation
  - We have discussed the validity of the protocol
  - High similarity scores for some systems for both tasks
  - Speaker adaptation task (Spoke task) with 2h of training data is handled as well as the Hub task (51h of data)
- Intelligibility evaluation
  - Excellent scores on SUS
  - Use of LLM promising for homographs

Future directions

- Current architecture are now becoming very competitive for the synthesis of high-quality **isolated sentences** in terms of speech quality, speaker similarity and intelligibility

➡ More challenging tasks

- Less data
- Speech synthesis in context

➡ More challenging evaluations

- Quality, similarity and intelligibility on specific events (not globally anymore)
- New dimensions: expressivity, comprehensibility, capturing attention ability, etc.
- Adapted to a specific use case: is the communication task successful?

To be discussed further together at the end of the day

Acknowledgements

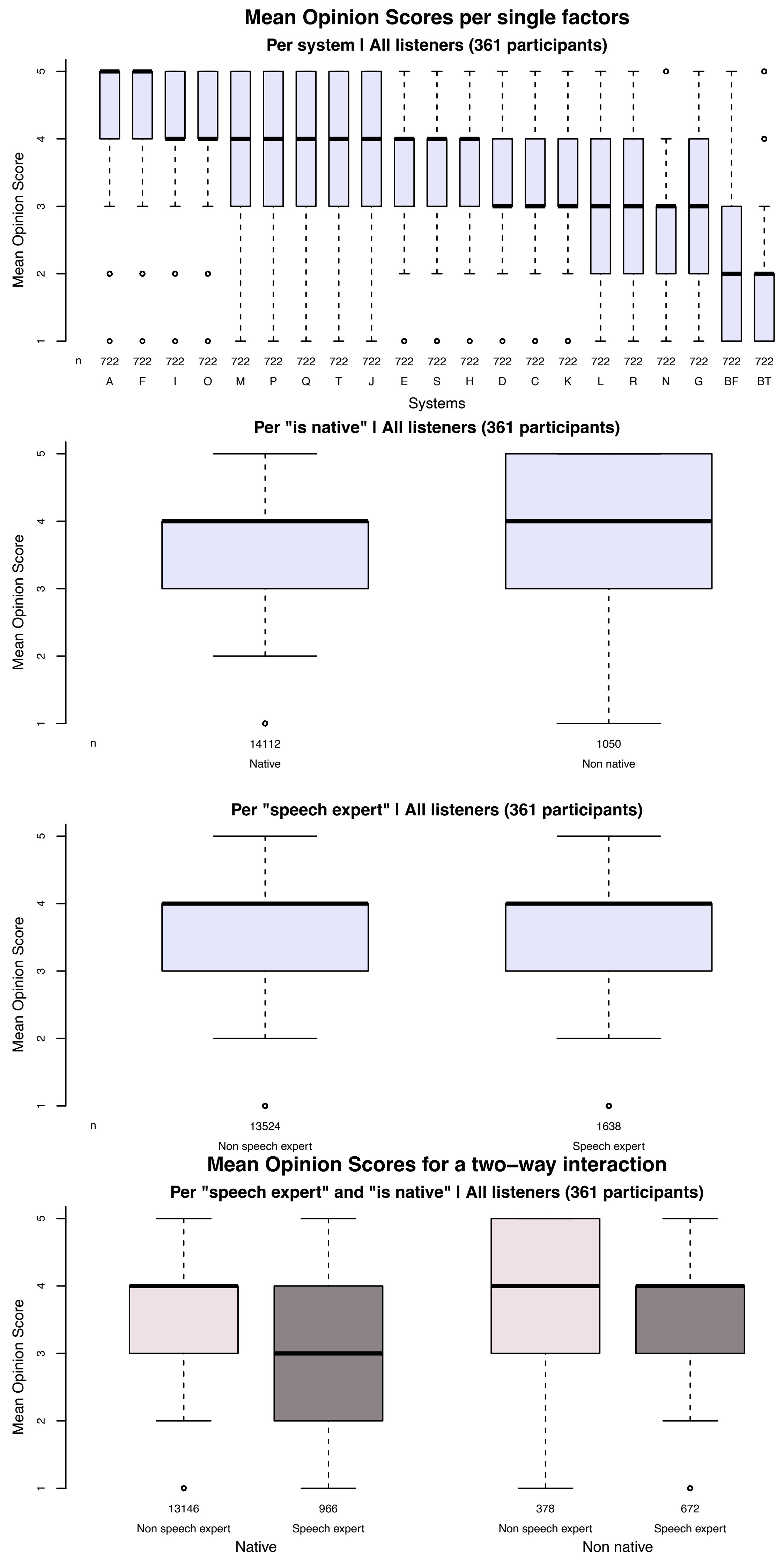
- Organisation
  - **Simon King:** advice in the challenge organisation
- Benchmark systems
  - **Brooke Stephenson:** built and trained both Tacotron and FastSpeech benchmarks
- Evaluation
  - **Brooke Stephenson:** experimental design; development and running
  - **Silvain Gerber:** statistical analysis
  - **Gérard Bailly**: generation of SUS and homographs sentences; recording of homographs reference audios; computation of WER; advice in the evaluation
- Corpus
  - **Aurélie Derbier:** voice for the FS1 task
  - **Romain Legrand, Frédéric Elisei:** recording of Aurélie Derbier
  - **Gérard Bailly:** creation of the Blizzard train and test sets (full annotation)
- Technical
  - **Sébastien Le Maguer:** post-processing of the submitted data
  - **Martin Lenglet:** web development for the online listening tests
  - People involved in previous Blizzard Challenges: initial test design; initial scripts to produce statistics and graphs
- Scientific
  - **Olivier Perrotin, Gérard Bailly, Damien Lolive, Nicolas Obin, Simon King:** scientific committee for the Blizzard challenge tasks definition
- And last, but never least, our usual thanks to all participants and listeners

Q&A

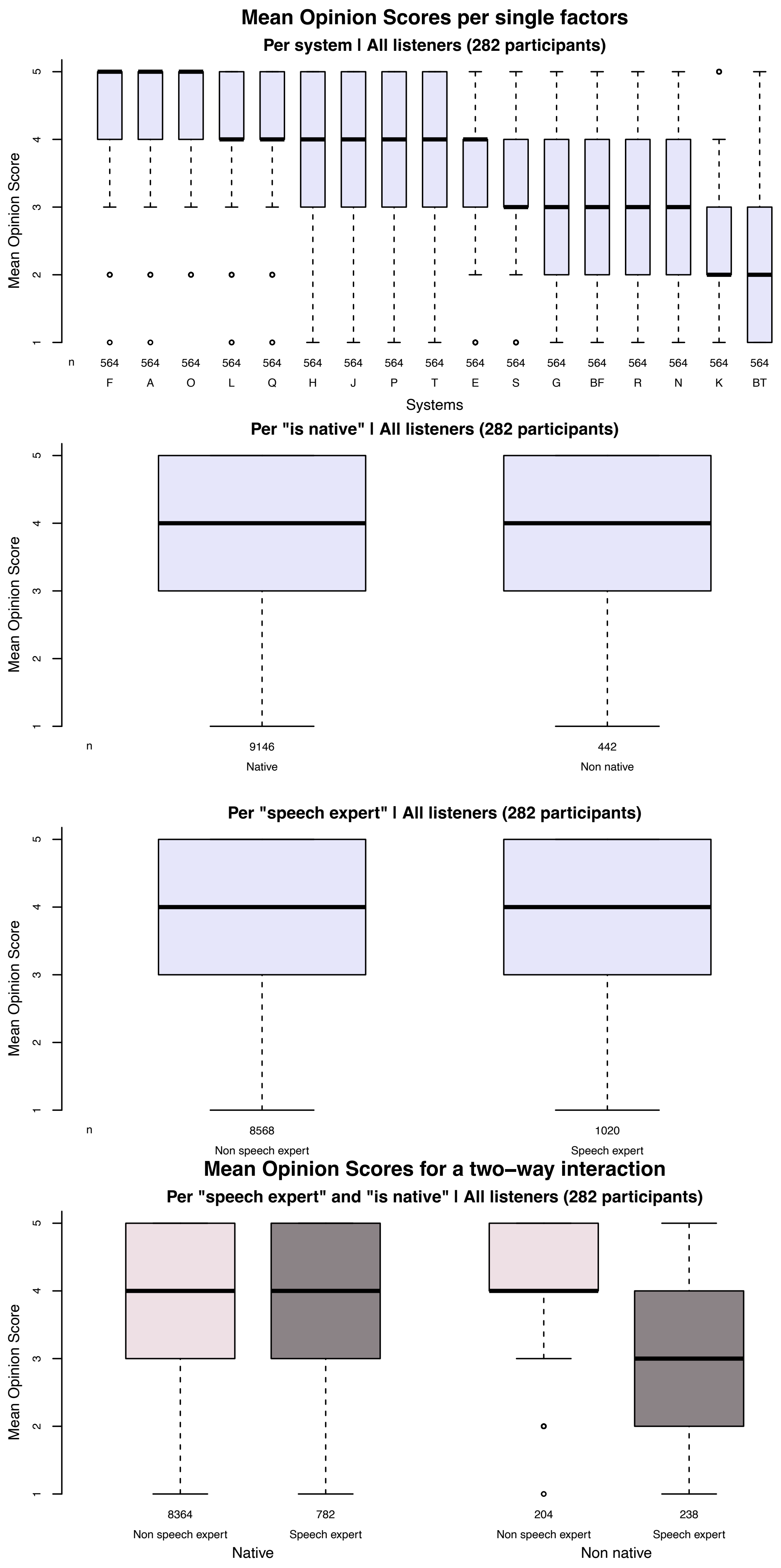
Appendices

Effect of speech expertise and is native

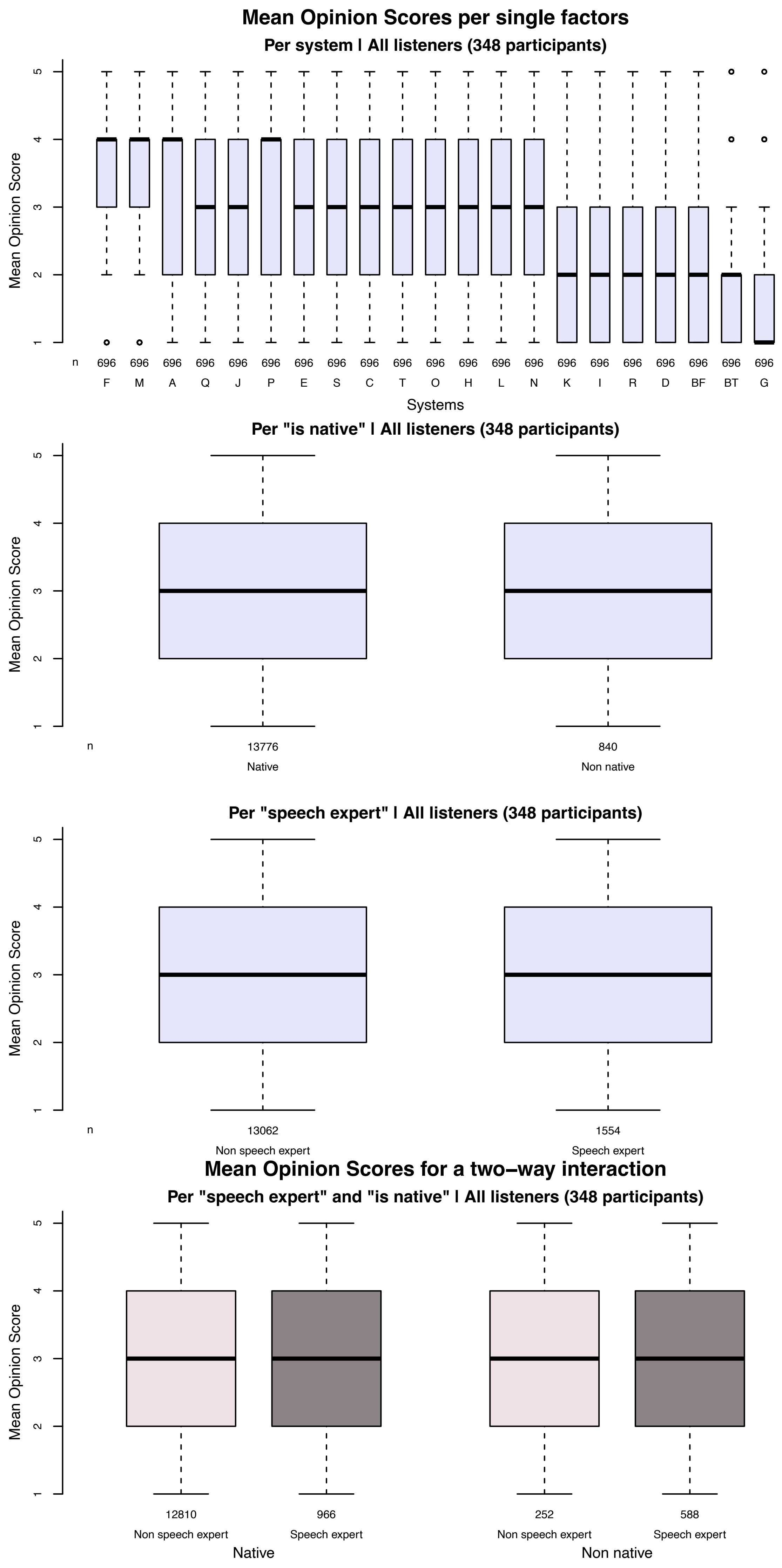
Quality | Hub task



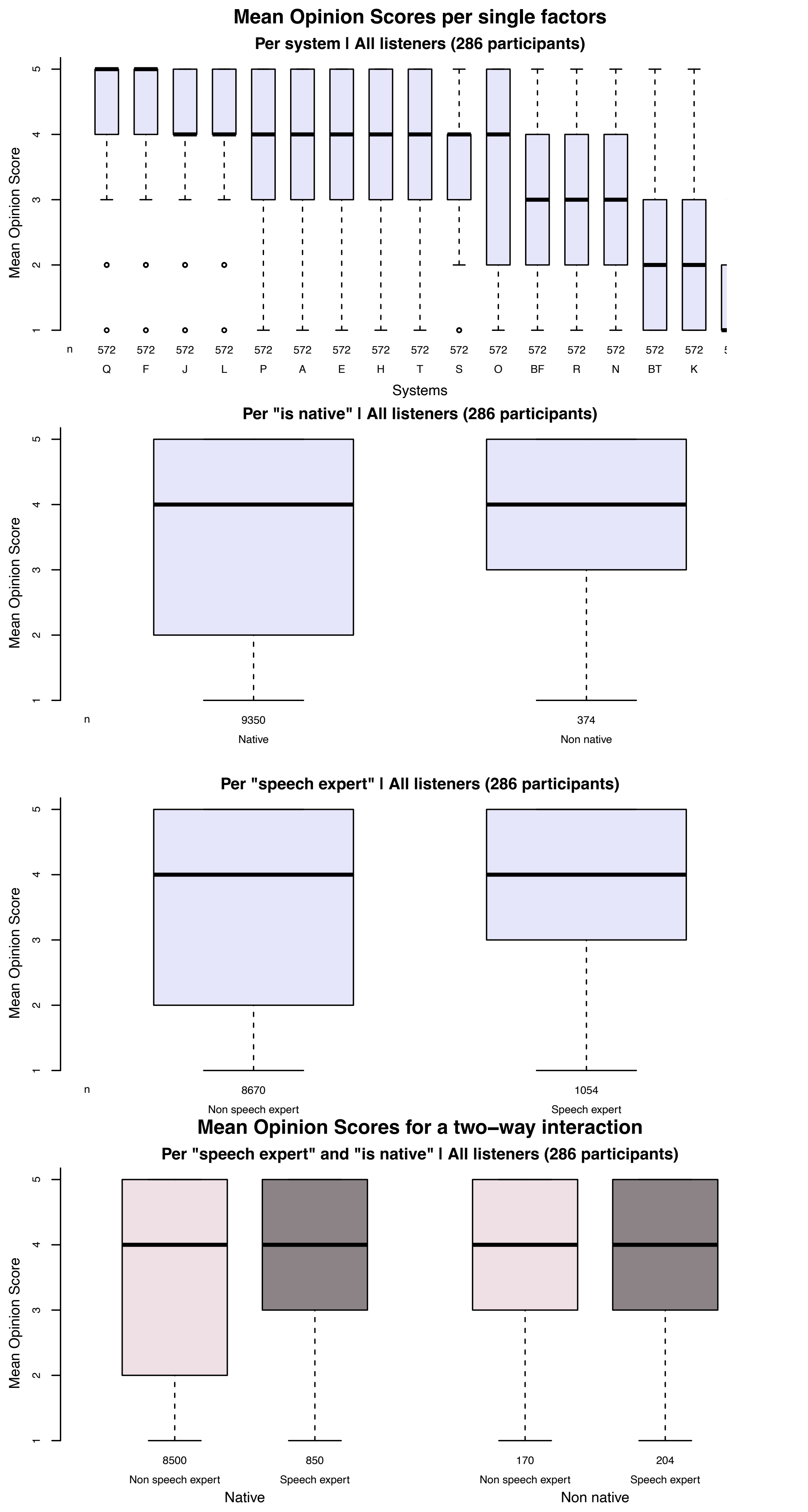
Quality | Spoke task



Similarity | Hub task



Similarity | Spoke task



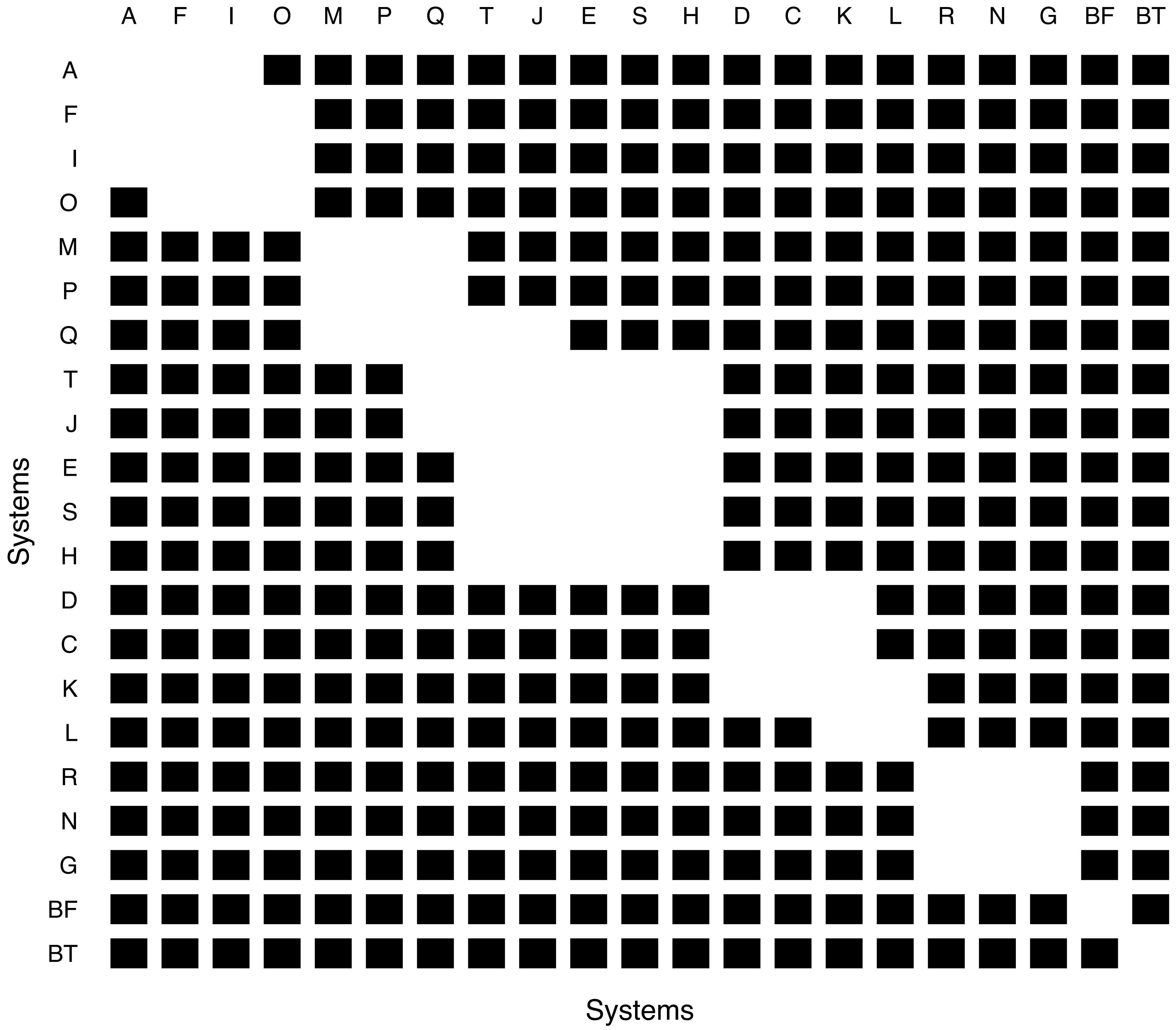
Speech quality | Hub task

Multiple comparisons vs. Wilcoxon

Significant differences in MOS scores between systems, indicated by solid black boxes ($p < 0.01$)

All listeners (361 participants)

Multiple comparisons following an ordinal regression with random effects



Pairwise comparison following a Wilcoxon test and Bonferroni correction