EPSRC
Engineering and Physical Sciences Research Council

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations

Mirjam Wester, Cassia Valentini-Botinhao and Gustav Eje Henter

CSTR, University of Edinburgh

Introduction

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Objective measures aren’t good enough at measuring the perceptual quality of synthetic speech

- Subjective listening tests remain the gold standard:
  - Mean Opinion Score (MOS) tests
  - Preference tests
  - ABX tests
  - Transcription tasks
  - MUSHRA tests

- Despite many listening test guidelines, contemporary evaluations are often very poor as they don’t take guidelines into account.

Our study

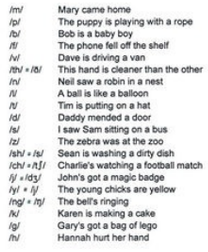
Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Common shortcomings in subjective evaluations from Interspeech 2014

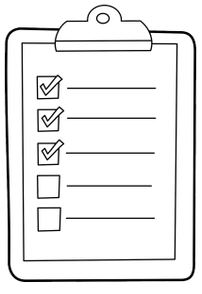
Using Blizzard 2013 data we show the importance of:

Sufficient participants



/m/	Mary came home
/p/	The puppy is playing with a rope
/b/	Bob is a baby boy
/f/	The phone fell off the shelf
/v/	Dave is driving a van
/th/ • /ð/	This hand is cleaner than the other
/n/	Neil saw a robin in a nest
/l/	A ball is like a balloon
/t/	Tim is putting on a hat
/d/	Daddy mended a door
/s/	I saw Sam sitting on a bus
/z/	The zebra was at the zoo
/sh/ • /s/	Sean is washing a dirty dish
/ch/ • /tʃ/	Charlie's watching a football match
/j/ • /dʒ/	John's got a magic badge
/y/ • /j/	The young chicks are yellow
/ng/ • /ŋ/	The bell's ringing
/k/	Karen is making a cake
/g/	Gary's got a bag of lego
/h/	Hannah hurt her hand

Sufficient test material



Checklist of elements that should be considered when designing a good listening test

Interspeech 2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Number of speech synthesis studies at Interspeech 2014 using a particular amount of listeners.

Number of listeners	Number of studies	
	Preference test	MOS

Interspeech 2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Number of speech synthesis studies at Interspeech 2014 using a particular amount of listeners.

Number of listeners	Number of studies	
	Preference test	MOS
1-10	10	8

Interspeech 2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Number of speech synthesis studies at Interspeech 2014 using a particular amount of listeners.

Number of listeners	Number of studies	
	Preference test	MOS
1-10	10	8
11-20	5	5

Interspeech 2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Number of speech synthesis studies at Interspeech 2014 using a particular amount of listeners.

Number of listeners	Number of studies	
	Preference test	MOS
1-10	10	8
11-20	5	5
21-30	0	1

Interspeech 2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Number of speech synthesis studies at Interspeech 2014 using a particular amount of listeners.

Number of listeners	Number of studies	
	Preference test	MOS
1-10	10	8
11-20	5	5
21-30	0	1
31-50	4	5

Interspeech 2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Number of speech synthesis studies at Interspeech 2014 using a particular amount of listeners.

Number of listeners	Number of studies	
	Preference test	MOS
1-10	10	8
11-20	5	5
21-30	0	1
31-50	4	5
>50	3	3

Interspeech 2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Number of speech synthesis studies at Interspeech 2014 using a particular amount of listeners.

Number of listeners	Number of studies	
	Preference test	MOS
1-10	10	8
11-20	5	5
21-30	0	1
31-50	4	5
>50	3	3
Not stated	2	0

Interspeech 2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Number of speech synthesis studies at Interspeech 2014 using a particular amount of listeners.

Number of listeners	Number of studies	
	Preference test	MOS
1-10	10	8
11-20	5	5
21-30	0	1
31-50	4	5
>50	3	3
Not stated	2	0
Total studies	24	22

Missing details IS-2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Missing details IS-2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- The **demographics** of listeners (native or non-native, age, accent, possible hearing impairments).

Missing details IS-2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- The **demographics** of listeners (native or non-native, age, accent, possible hearing impairments).
- The **language** of the synthesised speech.

Missing details IS-2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- The **demographics** of listeners (native or non-native, age, accent, possible hearing impairments).
- The **language** of the synthesised speech.
- The **domain** of the sentence material (training and test).

Missing details IS-2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- The **demographics** of listeners (native or non-native, age, accent, possible hearing impairments).
- The **language** of the synthesised speech.
- The **domain** of the sentence material (training and test).
- The **number** of test samples (sentences, words, paragraphs).

Missing details IS-2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- The **demographics** of listeners (native or non-native, age, accent, possible hearing impairments).
- The **language** of the synthesised speech.
- The **domain** of the sentence material (training and test).
- The **number** of test samples (sentences, words, paragraphs).
- The specific **question** participants were asked to answer.

Missing details IS-2014

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- The **demographics** of listeners (native or non-native, age, accent, possible hearing impairments).
- The **language** of the synthesised speech.
- The **domain** of the sentence material (training and test).
- The **number** of test samples (sentences, words, paragraphs).
- The specific **question** participants were asked to answer.
- The **listening conditions** (headphones or speakers, listening booth or on the web).

Blizzard 2013

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- To illustrate importance of sentence coverage and number and type of listeners we re-analysed Blizzard 2013 data.
- Last English evaluation
- Focus on main task: MOS tests for naturalness and similarity (EH1)

listener type	number of listeners
EE (paid / lab / native)	50
ER (volunteers / not controlled)	92
ES (speech experts / not controlled)	52
All	194

Blizzard data details

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- 11 systems including natural speech
- 11 listener groups
  - 4-5 listeners per group for EE | 5-10 for ER | 3-5 for ES
- Each listener scores each system:
  - 5 times for naturalness
  - once for similarity
- MOS test was used for both naturalness and similarity

Re-analysis of Blizzard

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Progressively larger subsets of data

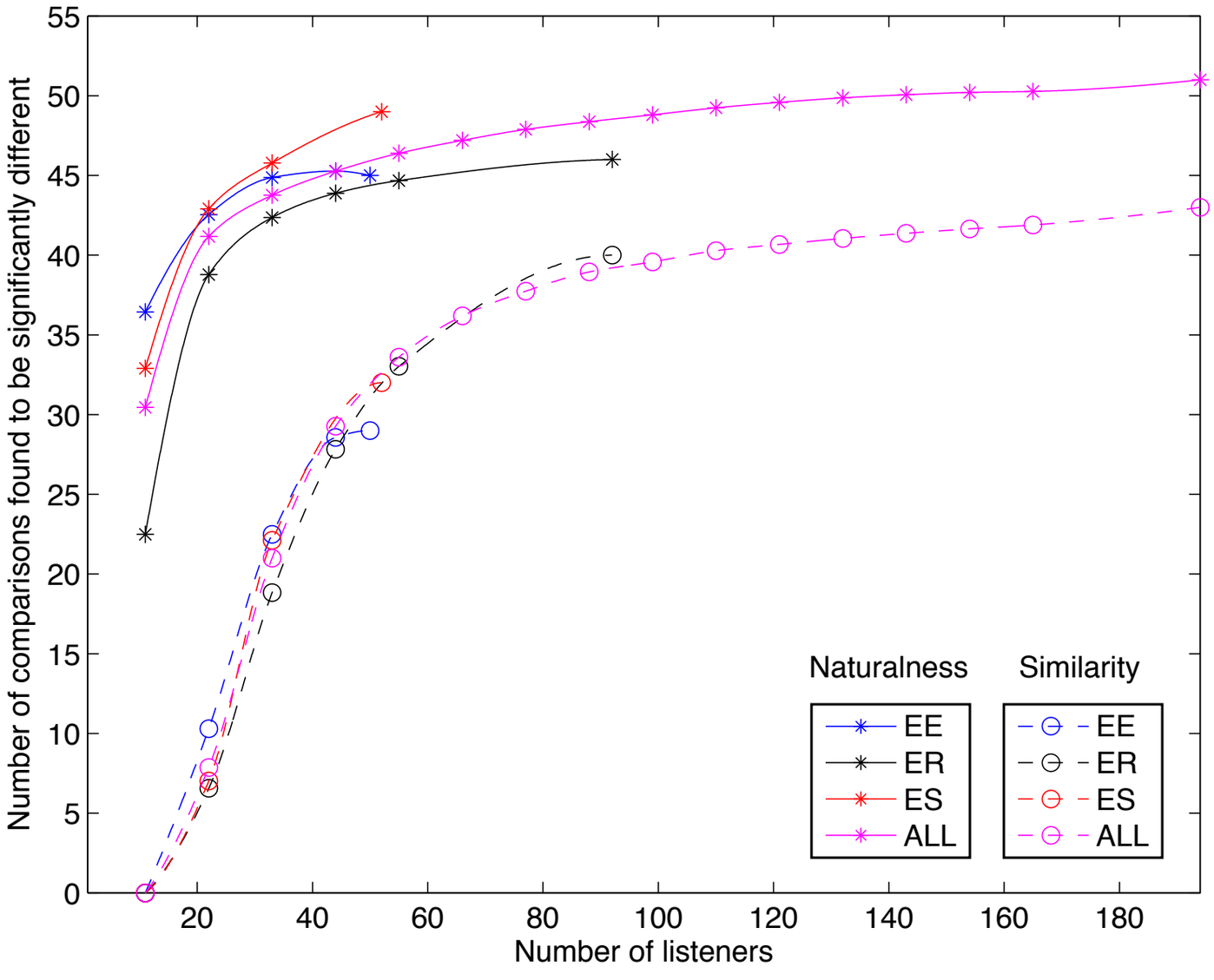
    1. number of significantly different system pairs

    2. rank correlation between the ranking given by the current data subset and the ranking obtained when considering all participants for the test in question

Participants (I)

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

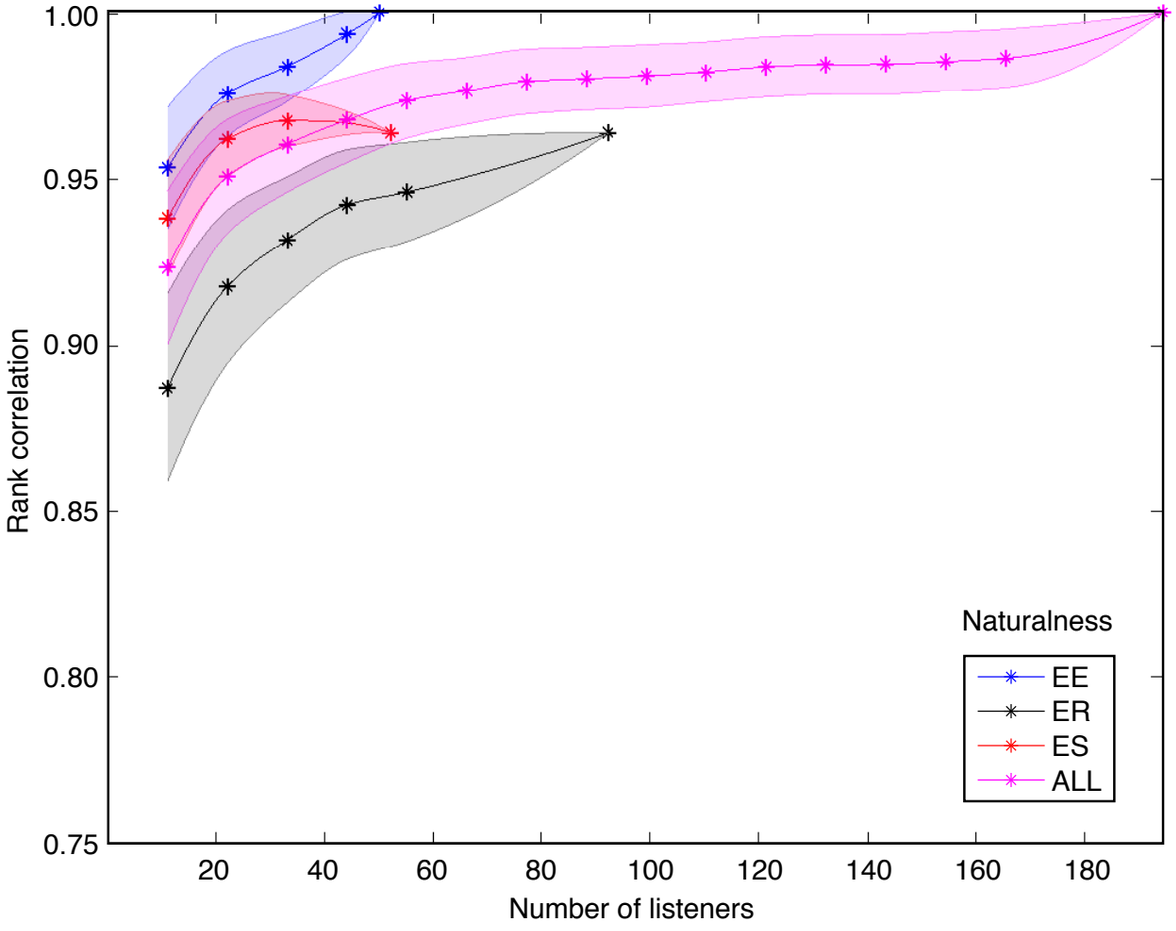


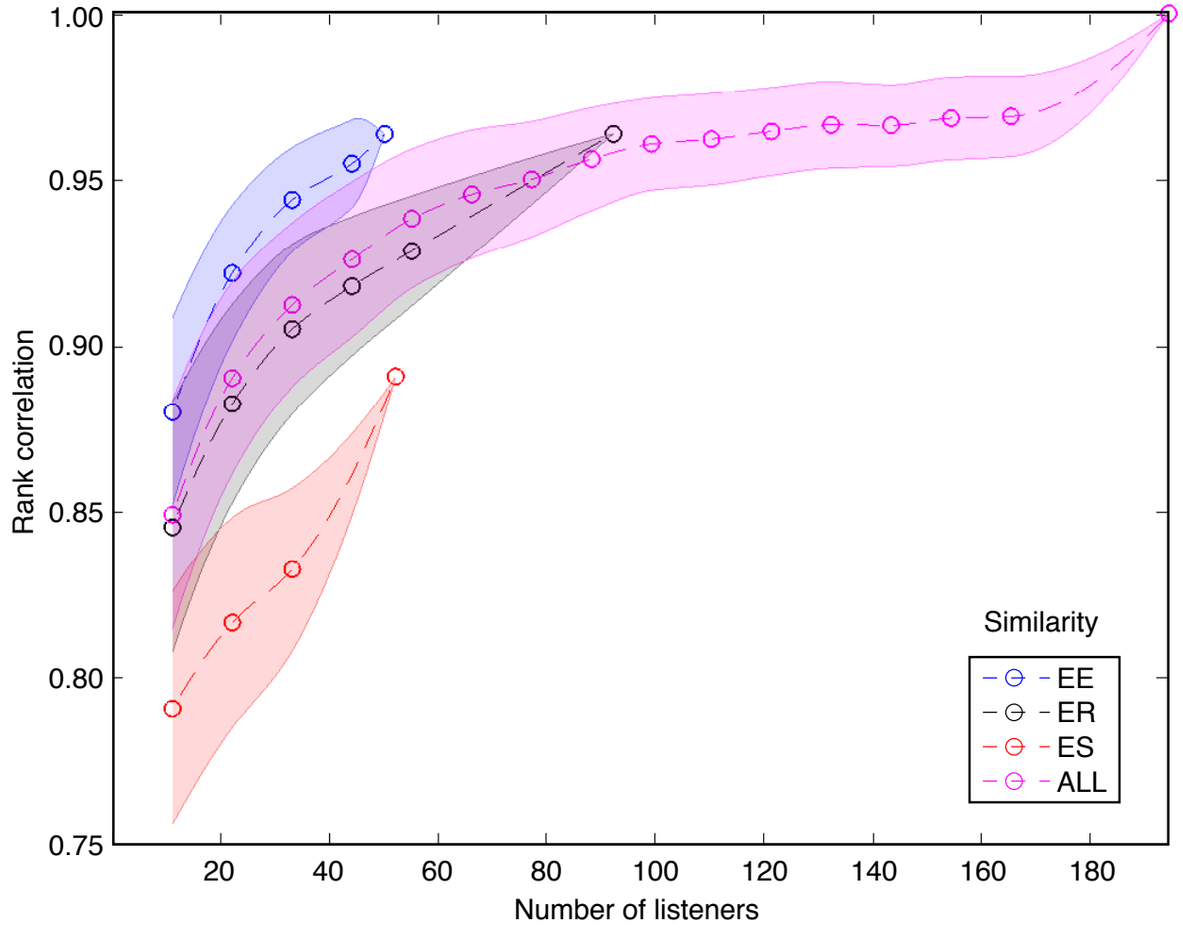
- Blizzard similarity tests overall resulted in fewer significant differences than the naturalness evaluation.

Participants (II)

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



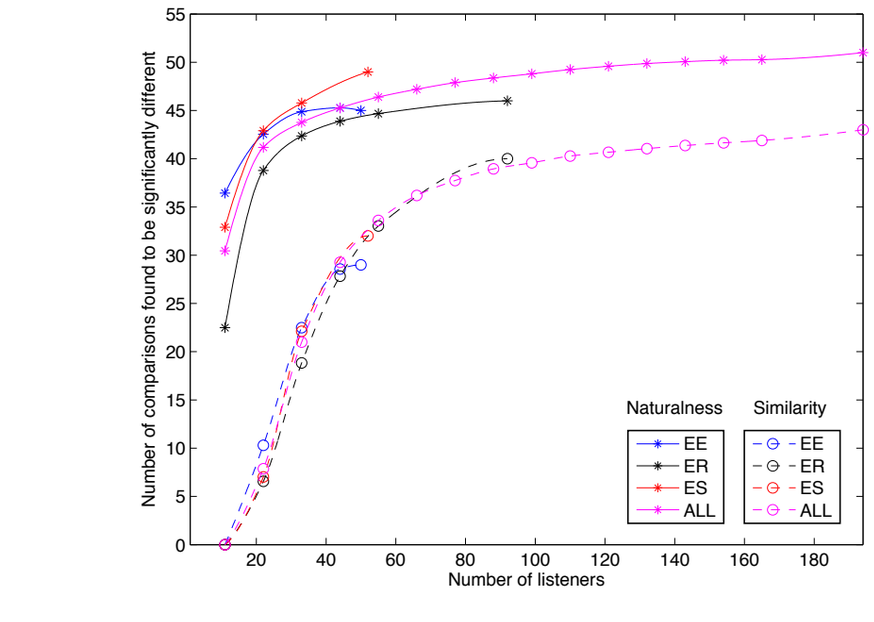


- Naturalness: 30 paid participants (EE) sufficient for strong correlation (>0.98).
- Similarity: results never quite reach stability

Participants (III)

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



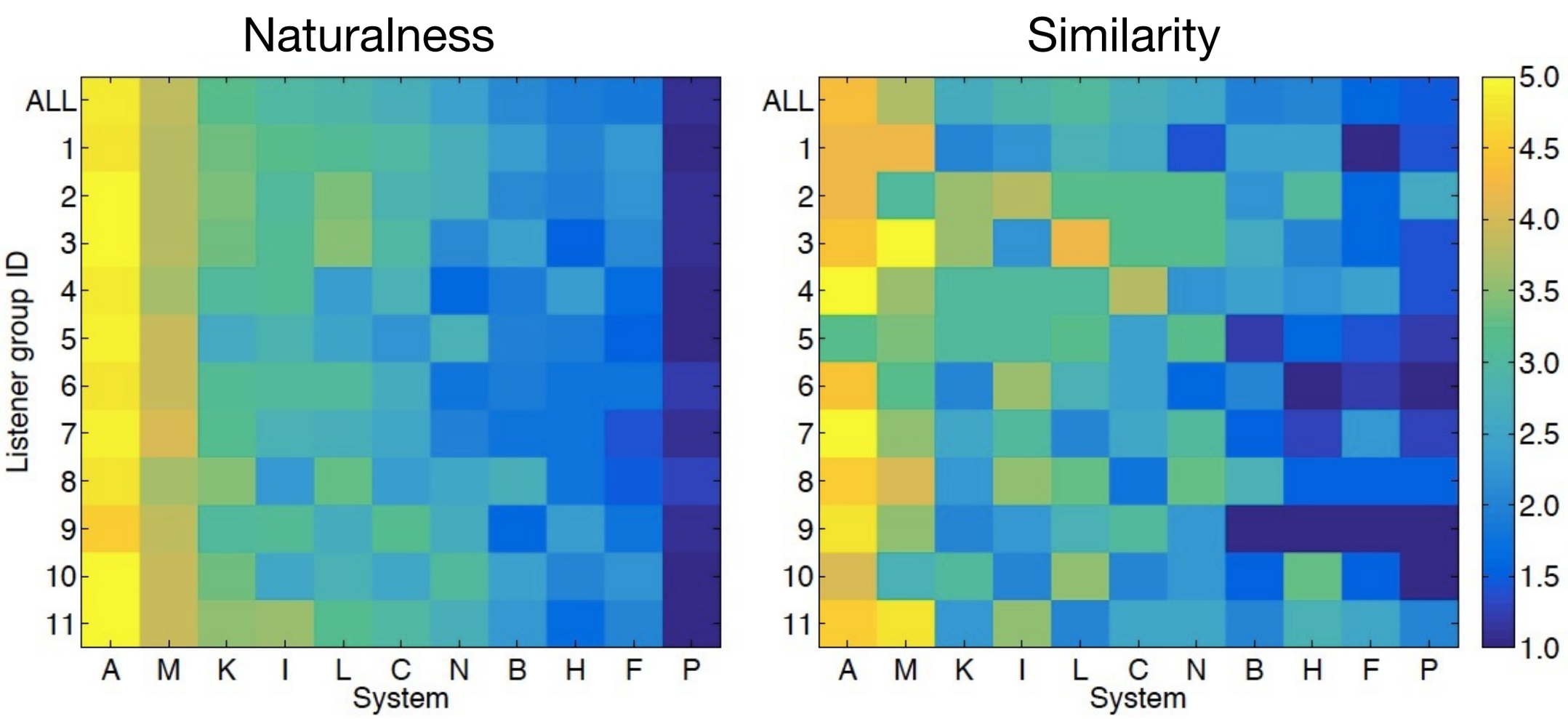
- EE (Paid listeners) correlate best with full-data rankings
- ER (Volunteers) consistently give low rank correlations and least number of significant pairs for a given number of listeners
- ES (Expert listeners) identify a large number of significant differences in naturalness, but their rank correlation with the overall full data picture was either close to average (naturalness) or the lowest observed (similarity)

/m/ Mary came home
/p/ The puppy is playing with a rope
/b/ Bob is a baby boy
/f/ The phone fell off the shelf
/v/ Dave is driving a van
/th/ + /ð/ This hand is cleaner than the other
/n/ Neil saw a robin in a nest
/l/ A ball is like a balloon
/t/ Tim is putting on a hat
/d/ Daddy mended a door
/s/ I saw Sam sitting on a bus
/z/ The zebra was at the zoo
/sh/ + /s/ Sean is washing a dirty dish
/ch/ +/tʃ/ Charlie's watching a football match
/j/ +/dʒ/ John's got a magic badge
/y/ + /j/ The young chicks are yellow
/ng/ + /ŋ/ The bell's ringing
/k/ Karen is making a cake
/g/ Gary's got a bag of lego
/h/ Hannah hurt her hand

Data Coverage (I)

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



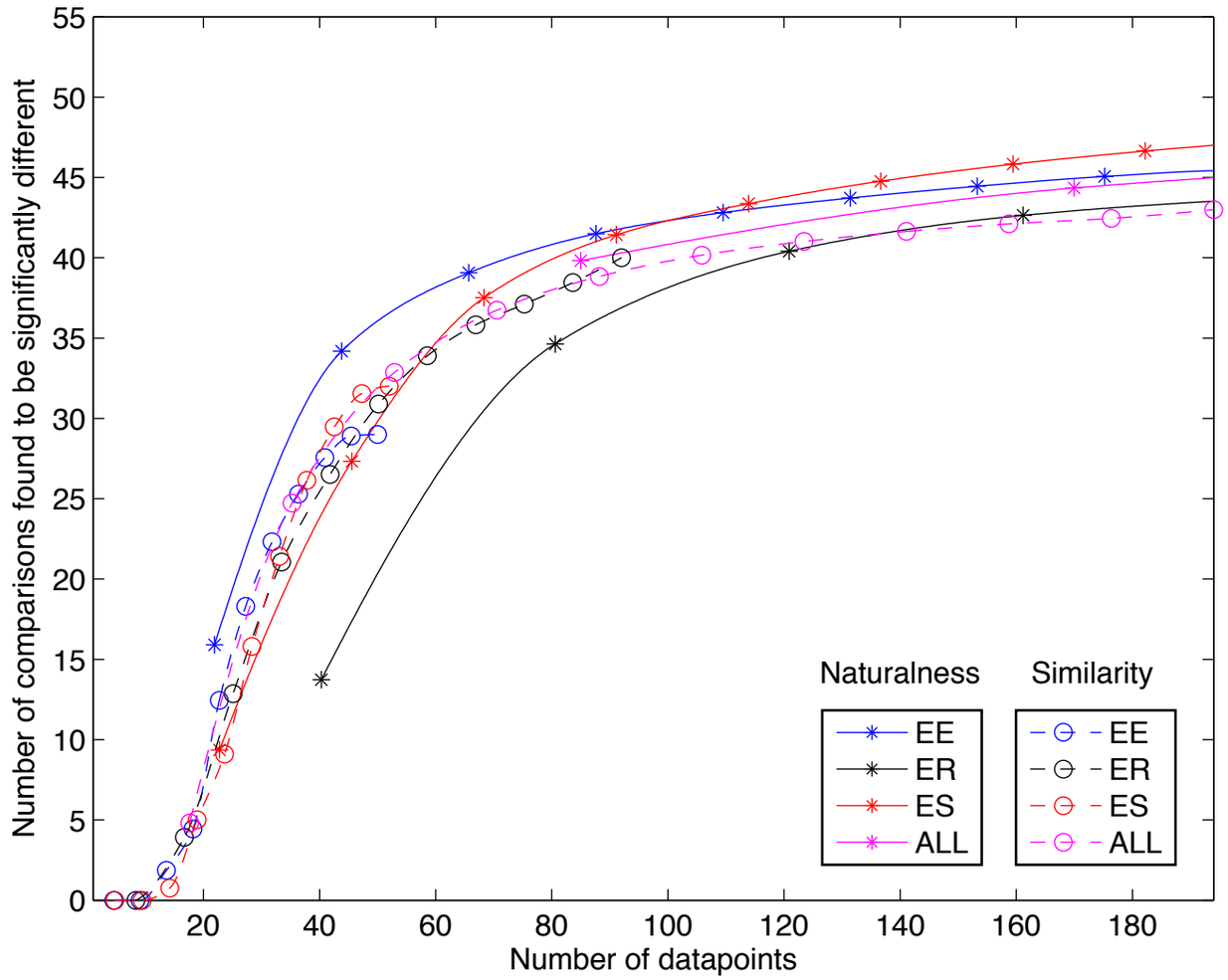
- Judgments change substantially between listener groups, particularly for the similarity scores.

/m/ Mary came home
/p/ The puppy is playing with a rope
/b/ Bob is a baby boy
/f/ The phone fell off the shelf
/v/ Dave is driving a van
/th/ + /ð/ This hand is cleaner than the other
/n/ Neil saw a robin in a nest
/l/ A ball is like a balloon
/t/ Tim is putting on a hat
/d/ Daddy mended a door
/s/ I saw Sam sitting on a bus
/z/ The zebra was at the zoo
/sh/ + /s/ Sean is washing a dirty dish
/ch/ +/tʃ/ Charlie's watching a football match
/j/ +/dʒ/ John's got a magic badge
/y/ + /j/ The young chicks are yellow
/ng/ + /ŋ/ The bell's ringing
/k/ Karen is making a cake
/g/ Gary's got a bag of lego
/h/ Hannah hurt her hand

Data Coverage (II)

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



- The big gap between naturalness and similarity tasks in previous figures can largely be explained by the difference in the number of scores collected per listener.

Summary

Natural Speech Technology

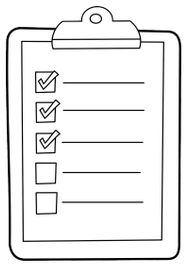
Edinburgh – Cambridge – Sheffield

- Blizzard analyses suggest that at least 30 listeners are needed for reliable results.

- Each listener should listen to several examples of each system evaluated.
  - 150 judgements per MOS should probably be a minimum.

- Types of listeners: paid participants, online volunteers and expert listeners
  - paid: above numbers apply
  - online: more data and more listeners
  - experts: their preferences differ from those of the general public

- Thought and design of experiments is paramount



Conclusion

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Take home message:
  - *Think before you test!*
- Report on the design of your experiment and motivate the choices made
- See the checklist in the paper for inspiration