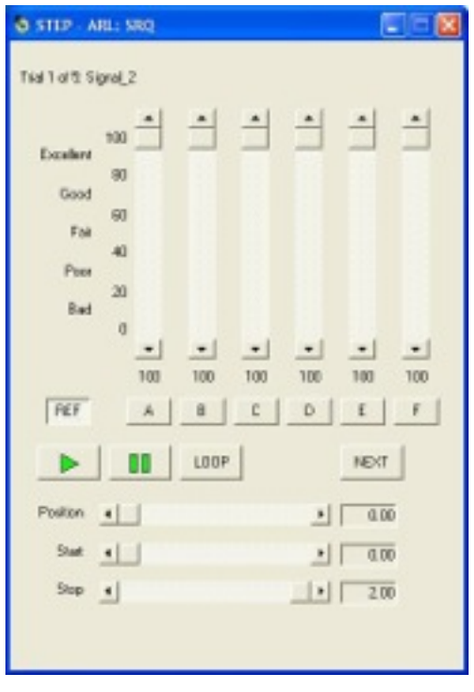
Subjective Evaluation of TTS

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

/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



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

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

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/ + /ʒ/	Sean is washing a dirty dish
/ch/ + /ʧ/	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

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- See hand-out

- Take home message:
  - *Think before you test!*
  - Don’t treat subjective evaluation as a fishing expedition.
    - Carry out pilot tests to learn how long experiments take, if they are feasible and doable for subjects.
    - Conform to good scientific practice by deciding your hypothesis and how many subjects are needed *before* running the experiment, and correct for multiple comparisons in the statistical analysis.
    - Report on the design of your experiment and motivate the choices made – just showing an MOS plot is not enough information.

Checklist for successful testing

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- What test to use? MOS, MUSHRA, preference, intelligibility, and same/different judgements all fit different situations.
- Which question(s) to ask? Be aware that the question you ask may influence the answer you get.
- The terms you use may be interpreted differently by listeners, e.g., what does “quality” or “naturalness” actually mean?
- What type of listeners? Native vs. non-native? Speech experts vs. naïve listeners? Age, gender, hearing impairments? Different listener groups can lead to different results.
- Is a reference needed? Consider giving a reference or adding training material, particularly for intonation evaluation. Also consider the case for including other anchors.
- How many listeners to use? Our Blizzard analyses showed a minimum of 30 paid listeners.
- How many datapoints are needed? Our Blizzard analyses suggest a minimum of 150 points per MOS value.
- Is the task suitable for human listeners? Take into consideration listener boredom, fatigue, and memory constraints, as well as cognitive load.
- Can you use crowdsourcing? The biggest concern here is how to ensure the quality of the test-takers.
- How is the experiment going to be conducted? With headphones or speakers, over the web or in a listening booth?
- Is the evaluation material unbiased and free of training data?
- Report the design of experiment and motivate choices made.

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

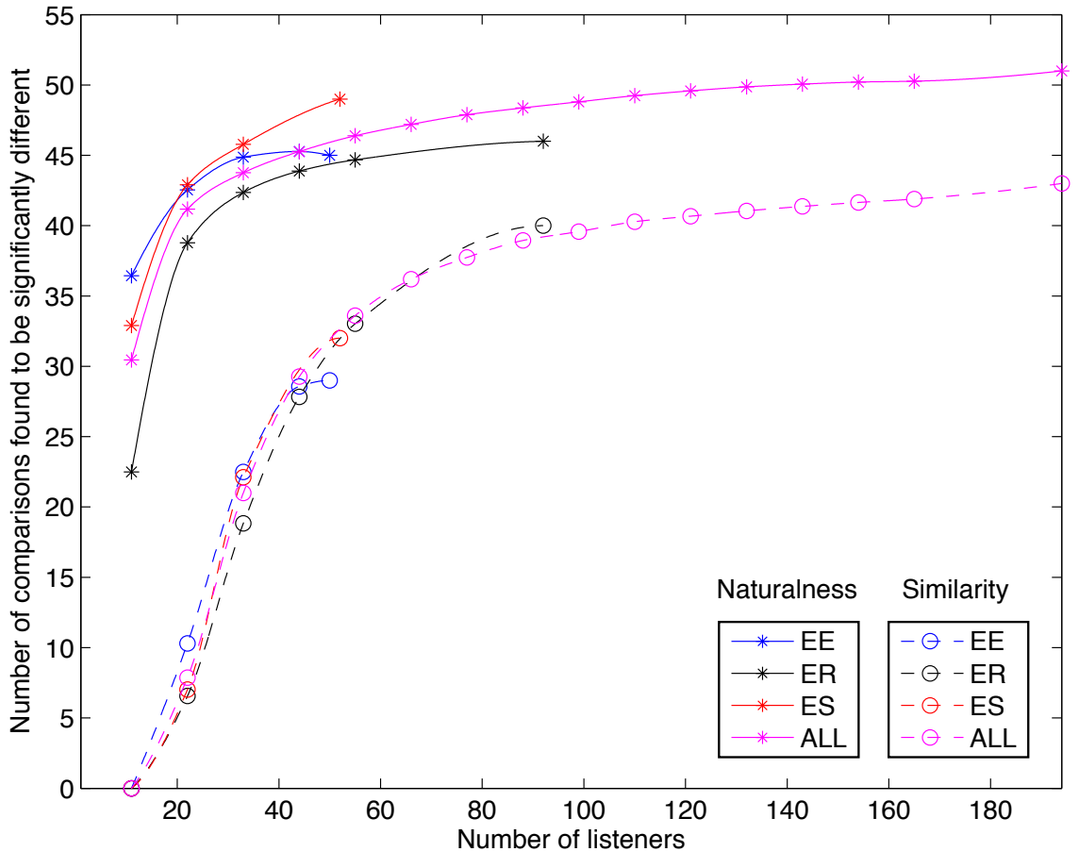
- The **demographics** of listeners (native or non-native, age, accent, possible hearing impairments).
- The **language** of the synthesised speech.
- The **domain** of the sentence material (training and test).
- The **number** of test samples (sentences, words, paragraphs).
- The specific **question**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

Participants (I)

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Number of system pairs found to be significantly different when gradually increasing the number of listeners included in the analysis.

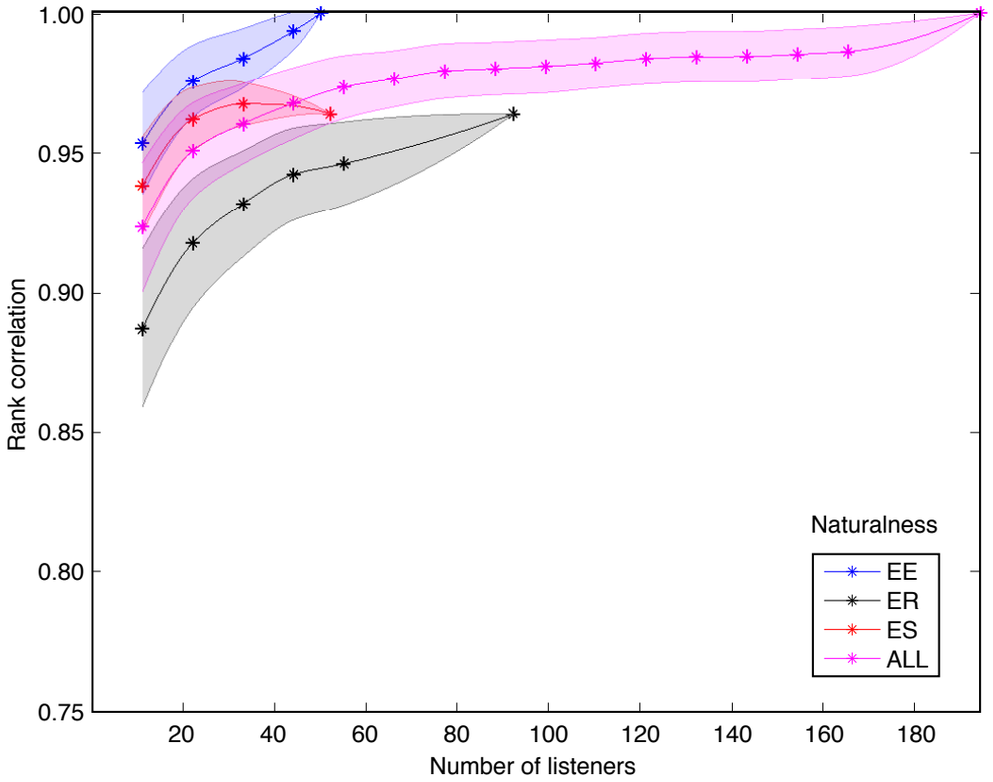


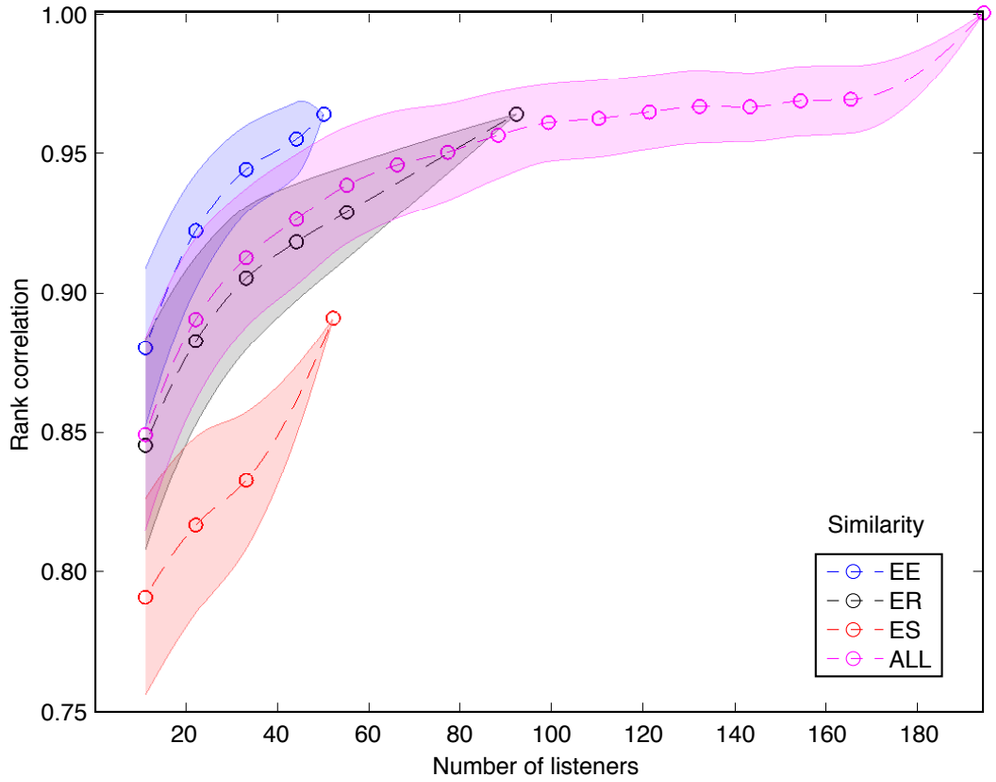
- Blizzard similarity tests overall resulted in fewer significant differences than the naturalness evaluation.

Participants (II)

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Rank correlation between system rankings (sub-sampled data) and the final ranking.



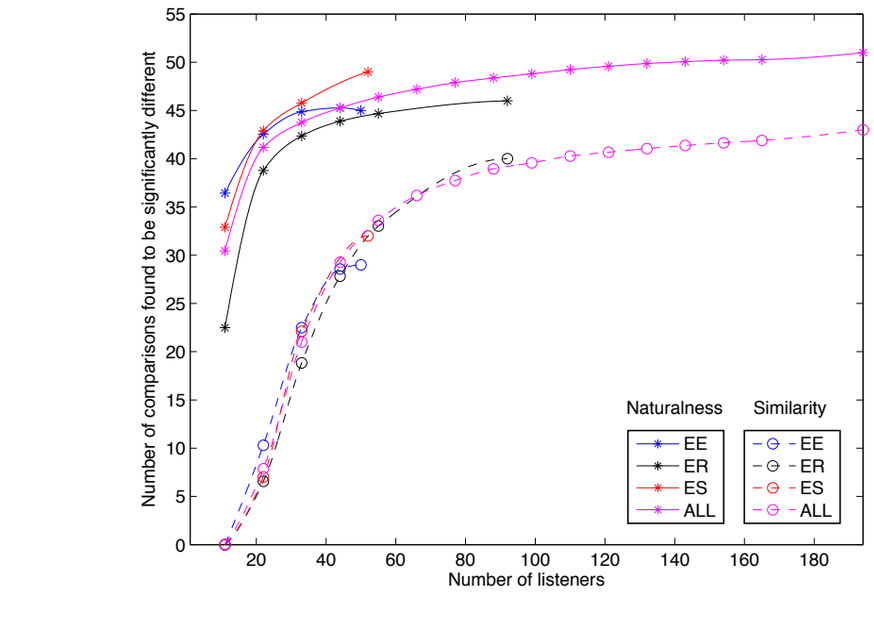


- 30 paid participants (EE) sufficient for strong correlation (>0.98) naturalness.

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)

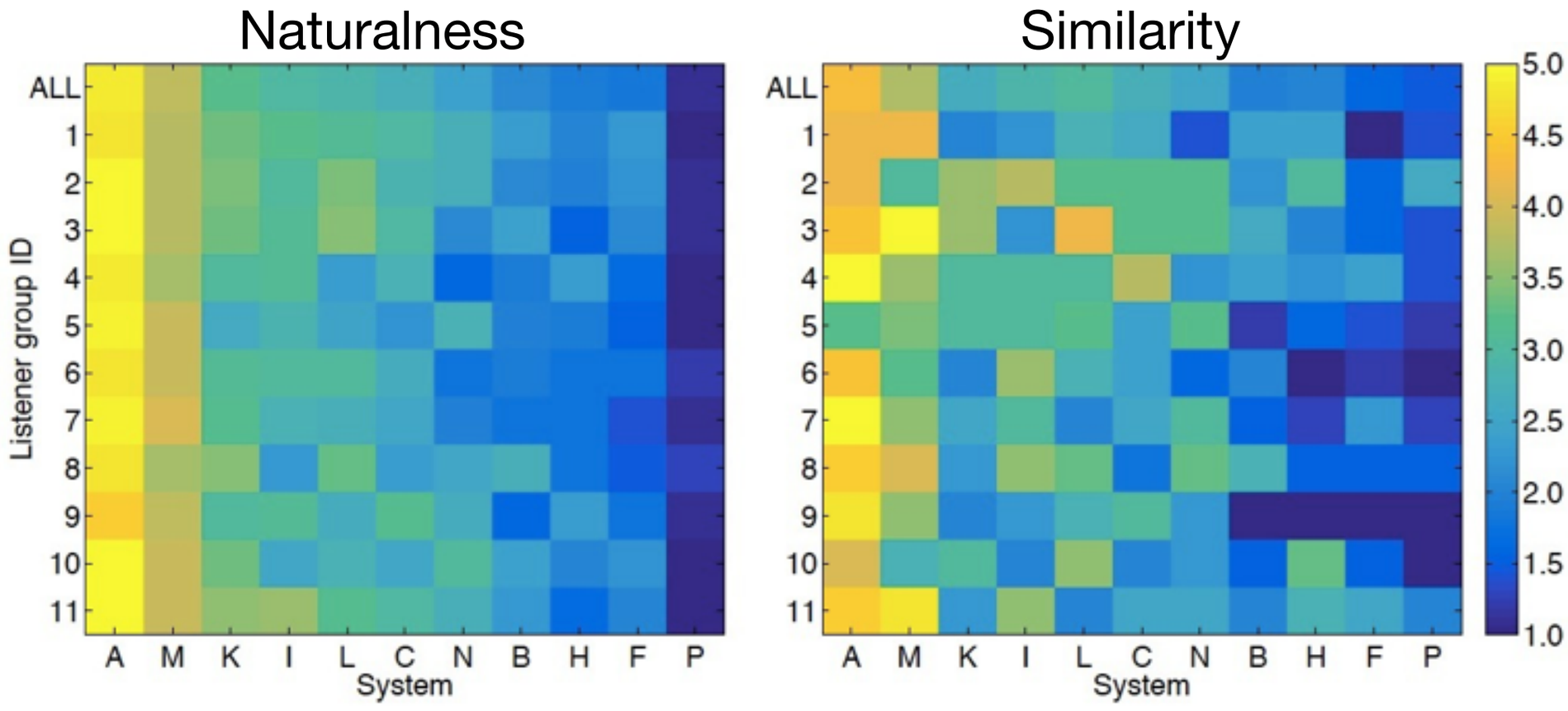
/m/ Mary came home
/p/ The puppy is playing with a rope
/b/ Bob is a baby boy
/f/ The phone fell off the shelf
/v/ Dave is driving a van
/th/ + /ð/ This hand is cleaner than the other
/n/ Neil saw a robin in a nest
/t/ A ball is like a balloon
/t/ Tim is putting on a hat
/d/ Daddy mended a door
/s/ I saw Sam sitting on a bus
/z/ The zebra was at the zoo
/sh/ + /s/ Sean is washing a dirty dish
/ch/ +/t∫/ 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

- To illustrate the importance of covering all system-sentence combinations we computed the average score of each system for each listener group (wherein everyone scored the same stimuli).



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

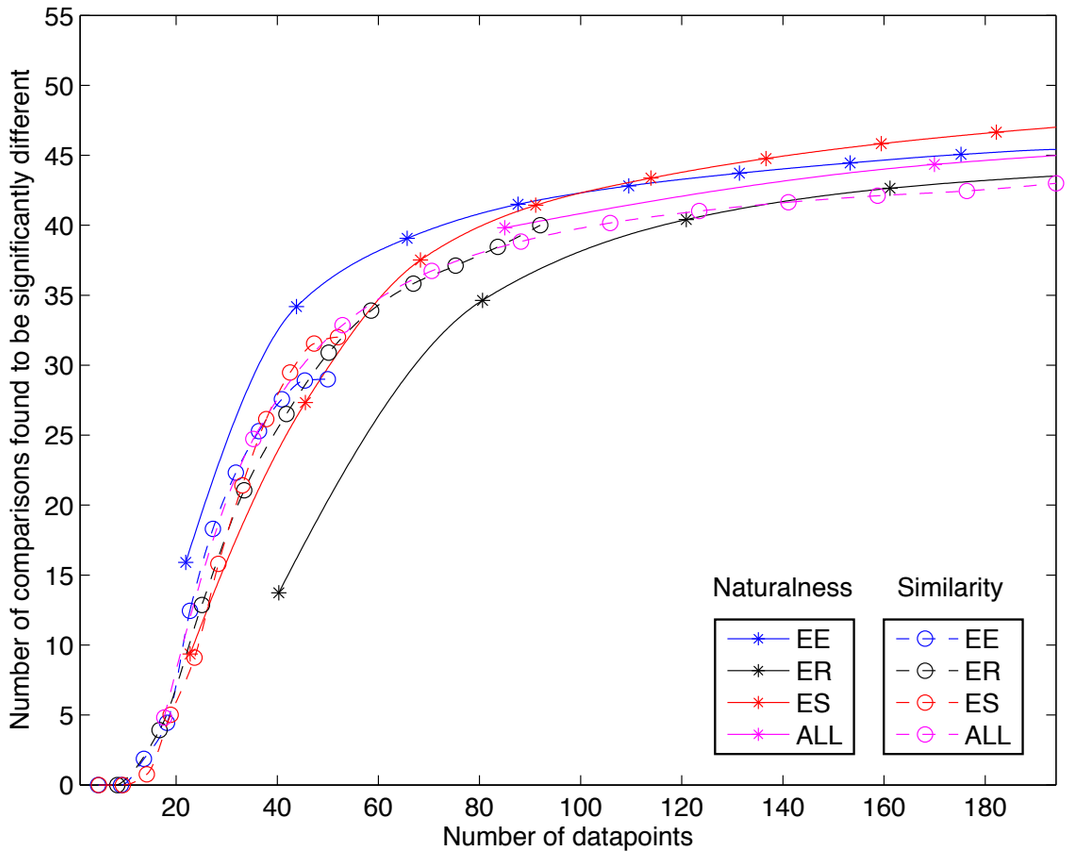
/m/ Mary came home
/p/ The puppy is playing with a rope
/b/ Bob is a baby boy
/f/ The phone fell off the shelf
/v/ Dave is driving a van
/th/ + /ð/ This hand is cleaner than the other
/n/ Neil saw a robin in a nest
/t/ A ball is like a balloon
/t/ Tim is putting on a hat
/d/ Daddy mended a door
/s/ I saw Sam sitting on a bus
/z/ The zebra was at the zoo
/sh/ + /s/ Sean is washing a dirty dish
/ch/ +/t∫/ 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)

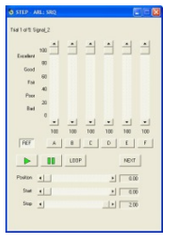
Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Number of significantly different pairs in MOS tests, as a function of the task and the number of datapoints.



- 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.

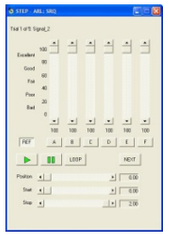


MUSHRA (I)

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) combines elements of MOS and preference tests
- Potential advantages of MUSHRA over MOS
  - sliding scale 0-100 makes it possible to rate very small differences
  - pairwise comparisons (same sentence)
  - hidden reference (natural speech)

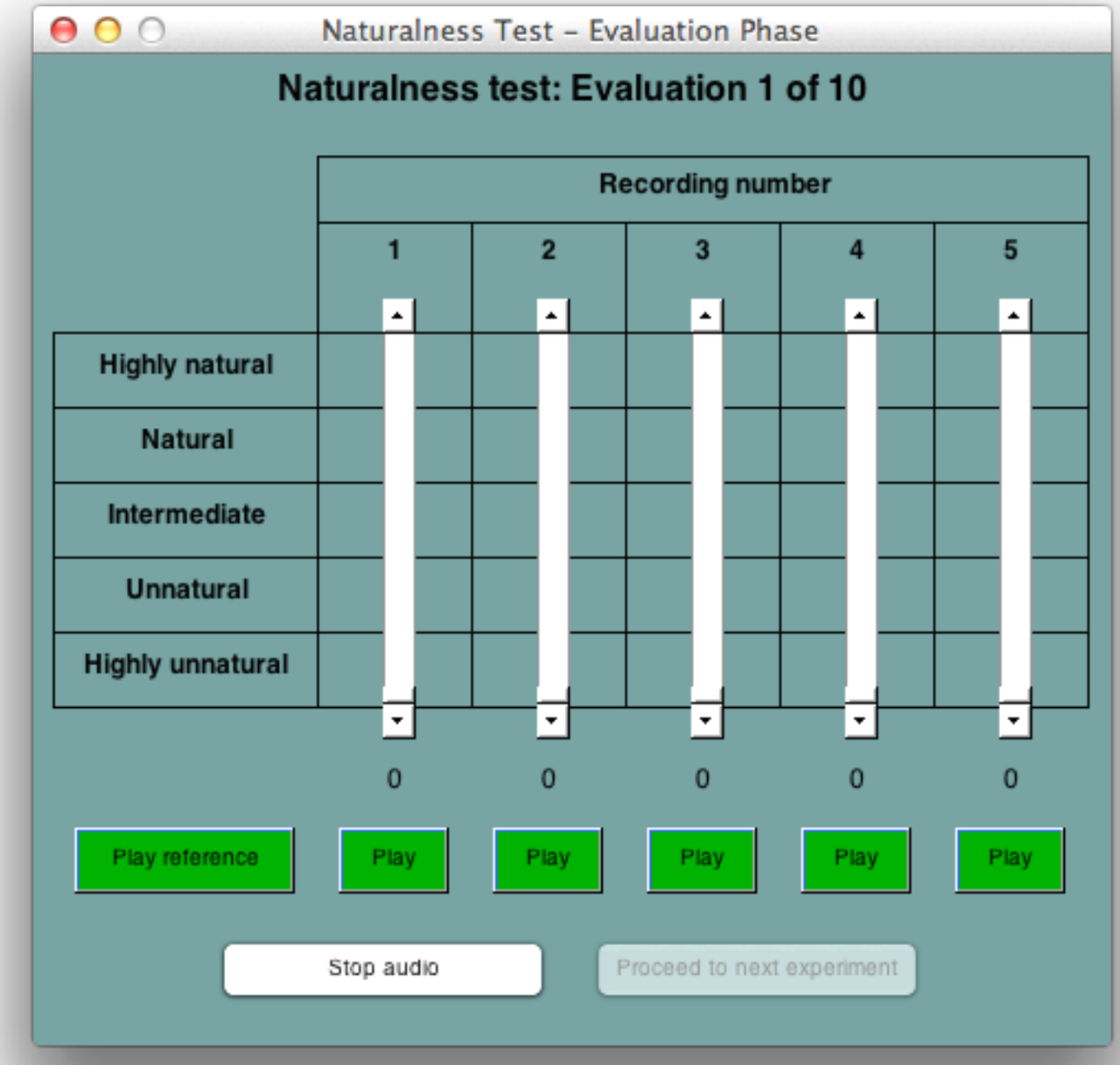


MUSHRA (II)

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Example



Summary

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Blizzard analyses shows 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: pay attention to the checklist!