

Evaluating Speech Synthesis

Simon King

Centre for Speech Technology Research

University of Edinburgh

What are you going to learn?

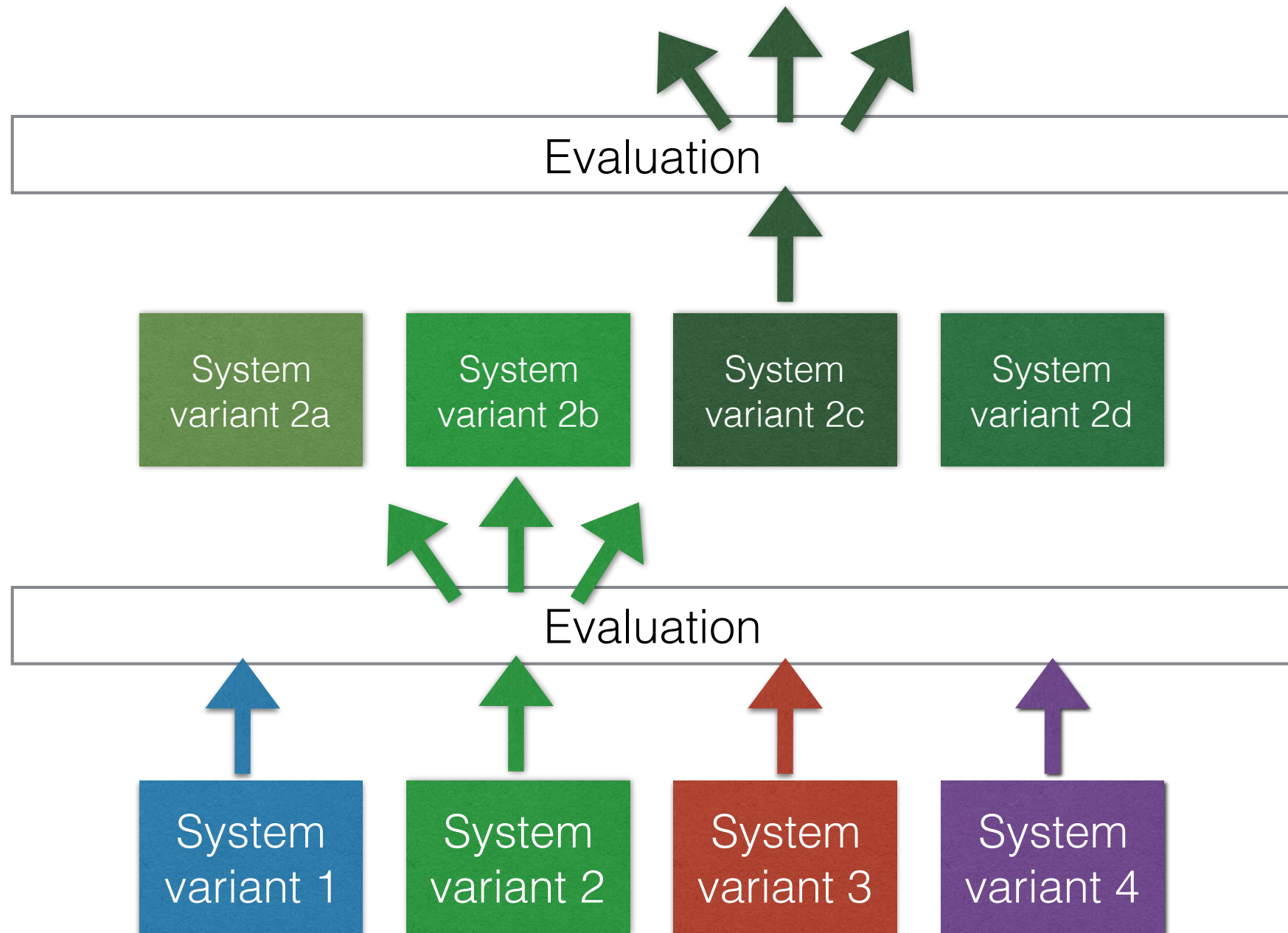
- **Why** evaluate?
 - **diagnostic** test to guide *future development*
 - **comparative** test against another system, or a baseline, *for publication*
 - **pass/fail** test *for a product*
- **What** to evaluate
 - whole system vs. components
- **Which** aspects of performance to evaluate
 - intelligibility, naturalness, speaker similarity,
- **How** to evaluate
 - listener task, test design, materials used, objective measures,

Evaluating Speech Synthesis

Why evaluate?

How to use evaluations to actually **improve** a system

- Typical pipeline (especially of the front end) architecture is not invertible
 - cannot ‘backpropagate’ listeners scores through the system
- Common methodology is “systematic trial and error”



Evaluating Speech Synthesis

What to evaluate

What to evaluate? Whole system vs. components

- Whole system
 - **pass/fail** - does commercial product meet user (or sales team!) requirements?
 - **cross-system** comparisons
 - optionally, control certain components
 - common database (see *Blizzard Challenge*)
 - fixed annotation and label alignments
 - common front end
- Components ('unit testing')
 - **isolated** component performance - e.g., POS tagger, LTS 'rules'
 - components **within a complete system** - e.g., waveform generator

Unit testing: does it predict whole system performance?

- restate: *does improving a component guarantee to improve whole system?*
- examples where this might not be the case:
 - text normalisation now produces word sequences that are poorly represented in the database (which was selected / normalised with the old component)
 - LTS produces phoneme sequences that are poorly represented in the database (which was aligned using phoneme sequences from the old component)
 - output of the improved component is used as input to the next component, which was optimised using the older version
- fixing bugs may reveal other bugs
- if all units perform 'perfectly' , would the whole system score 5/5 ?

Evaluating Speech Synthesis

Which aspects to evaluate

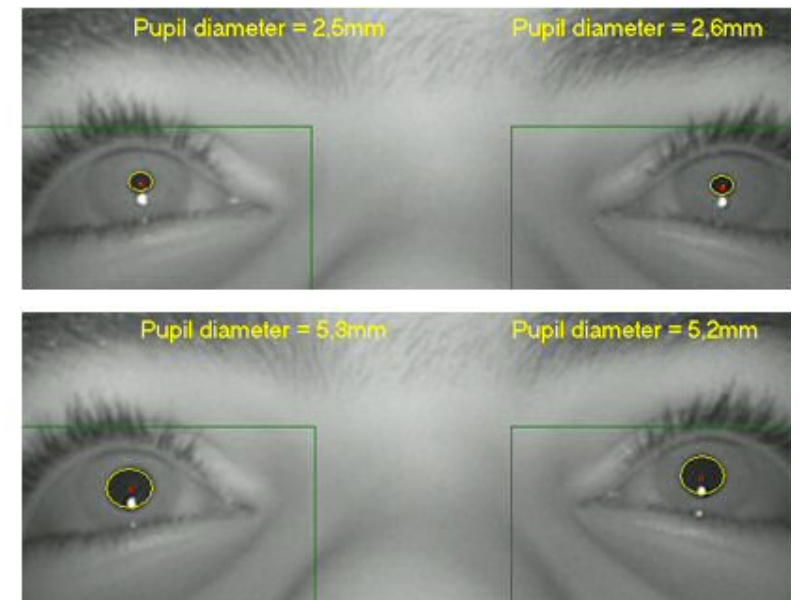
Which aspects of performance to evaluate?

- **Output quality**
 - Intelligibility
 - Comprehensibility
 - Naturalness
 - Speaker similarity
- **System performance**
 - speed
 - memory
- any more ?

Intelligibility vs. comprehension

image credit: Universiteit Utrecht

- Intelligibility
 - accuracy of word transcriptions
 - assume main factor is system, not listener
- Comprehension
 - not as clear how to measure this
 - probably mainly influenced by raw intelligibility
 - may be more influenced by listener factors, including cognitive abilities such as **working memory**
 - but ... measuring **listening effort** does make sense, if we can do it

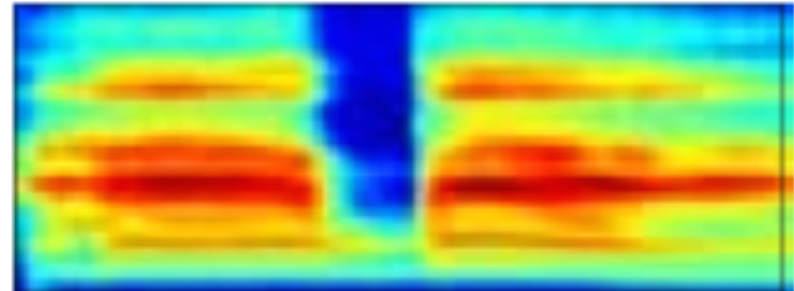


Evaluating Speech Synthesis

How to evaluate

How to evaluate?

- subjective measures
 - listener **task**
 - test **design**
 - test sample size
 - **materials** used
- objective measures
 - simple distances to reference samples
 - sophisticated auditory models



Listener task: what do we ask them to do?

- a simple, obvious task
 - “choose the version you prefer”
 - 5 point scales
 - “type in the words you heard”
- training the listeners
 - to pay attention to specific aspects of speech, e.g., prosody
- or give them a simpler task
 - and perform a more sophisticated analysis of the outcome
 - e.g., pairwise task followed by multi-dimensional scaling analysis

Now choose a score for how **natural** or **unnatural** the sentence **sounded**.
The scale is from **1 [Completely Unnatural]** to **5 [Completely Natural]**.

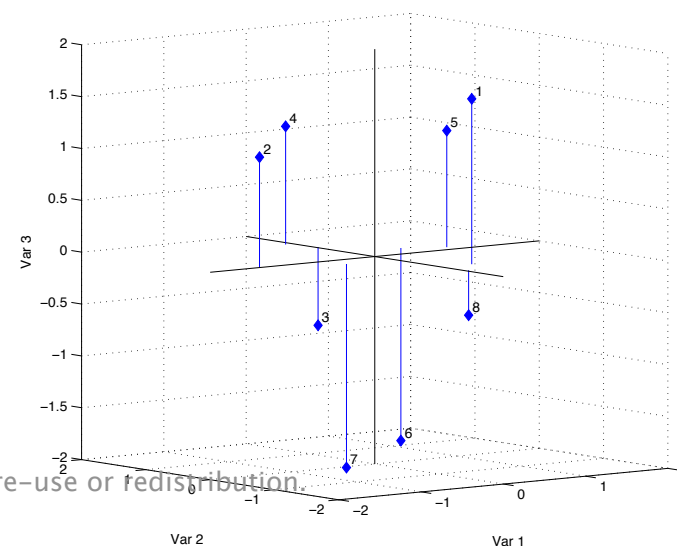
4 : Mostly Natural

Submit

Listen to the audio file by clicking on the image below, and type what you hear into the text box.



Submit



Test design

- **absolute vs. relative** judgements, making comparisons across different tests
 - do we need to include reference stimuli?
- **interface**
 - presenting stimuli to listeners
 - obtaining their response
- test **size**
 - duration per listener, number of test stimuli per listener and in total
- the **listeners**
 - type of listener, how to recruit them, quality control

Test design: absolute vs relative judgements

- within test
 - across stimuli
 - across listeners
- across test
- including reference stimuli
 - inclusion of natural speech, i.e., whole sentences
 - unit selection database contains natural speech, but synthetic output will not be rated 5/5
- how do we establish a **lower** bound? do we need one?

Test design: interface / presentation

- single stimulus
 - suitable for type-in test, e.g., using SUS
- pairs of stimuli
 - typically used in forced-choice “which do you prefer?” tests
- multiple stimuli
 - e.g., MUSHRA
- example web-based interface
 - <http://groups.inf.ed.ac.uk/blizzard/blizzard2013/english/register-es.html>

Section 2: Part 1 / 13

In this section, after you listen to each sentence, you will choose a score for the audio file you've just heard.

This score should reflect your opinion of how **natural** or **unnatural** the sentence sounded.

Note that you should not judge the grammar or content of the sentence, just how it **sounds**.

Listen to the example below.



Then choose a score for how **natural** or **unnatural** the sentence **sounded**.

The scale is from 1 [Completely Unnatural] to 5 [Completely Natural].

Submit

Report problems to [blizzard](#)

Section 5: Part 2 / 13

Listen to the example below, and type what you hear into the box.

After you click on the Play icon below, **you will be able to hear the sentence just once**. The icon will then be disabled.



Submit

Report problems to [blizzard](#)

ten to a short passage from an audio book, and you will give your opinion about various aspects of the voice you just heard.



ponse for each question below. Your score will be represented by a slider. For example, the midpoint in the overall quality slider should be used to best possible quality.

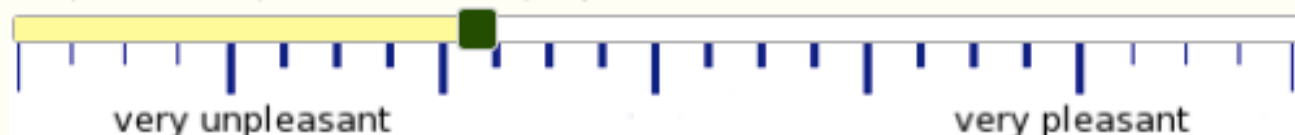
Overall impression

How do you rate the overall quality of the voice that read this passage?



Pleasantness

How pleasant did you find the voice you just heard?



Speech pauses

How did the pauses between words and sentences affect your listening to the passage?



Word stress

What did you think of the way words in the passage were stressed?



Intonation

What did you think of the "melody" of the voice reading this passage?



Add audio examples throughout

Test design: low literacy listeners (e.g., children)

- use pictures
 - “point at the cat”
- experimenter is present
 - and enters the subject’s responses

Test design: size of the test

- How many stimuli do we need?
 - **per listener** - depends on their patience (or amount of pay)
 - we try never to exceed 45 minutes test duration (when paying cash)
 - **in total** (see *Significance Testing*)
- Simple design
 - all listeners hear the same thing
 - may randomise order per listener
- More complex design
 - listeners do not all hear the same thing (see also *Blizzard Challenge*)
 - need to carefully balance materials in terms of listeners/systems/sentences

Test design: multi-listener designs

- too many stimuli to play to a single listener?
- form listeners into groups
 - as a group, they hear all stimuli
- can use a Latin Square to balance the design
 - good to also try to balance **ordering**

0	1	2	3	4
1	2	3	4	0
2	3	4	0	1
3	4	0	1	2
4	0	1	2	3

0	4	3	1	2
2	1	0	3	4
4	3	2	0	1
1	0	4	2	3
3	2	1	4	0

Test design: the listeners

- What **type of listener** do we want?
 - expert vs naive
 - native speakers
 - special skills (phoneticians? musicians?)
- **Recruitment**
 - local vs remote (e.g., AMT)
 - volunteers vs. paid
- **Quality control**
 - building this into the test - ‘*gold*’ items
 - building this into the analysis of the test outcome - *outlier removal*

Materials: how to design them

- two potentially opposing requirements
 - expected usage (domain) of the system
 - **goals of the evaluation** and the type of analysis we plan to do
- e.g., for intelligibility testing we might choose between:
 - isolated words
 - can narrow down range of possible errors listener can make
 - can design around minimal pairs (e.g., DRT, MRT)
 - but need to play them in a carrier sentence
 - full sentences
 - errors will be more variable & harder to predict, so harder to analyse
 - more natural task for the listener, perhaps closer to target domain

Materials: intelligibility

- ‘normal’ material - e.g., sentences from a newspaper
 - ceiling effect, due to interference from semantics (predictability)
- SUS - “*The unsure steaks overcame the zippy rudder*”
 - not representative of actual system usage
- DRT / MRT - “*Now we will say cold again.*” “*Now we will say gold again.*”
 - specific to individual phonemes - a **diagnostic unit test**

ANSI/ASA S3.2-2009 (R2014)

Method for Measuring the Intelligibility of Speech over Communication Systems

The scope of this standard includes the measurement of the intelligibility of speech over entire communication systems and the evaluation of the contributions of elements of speech communication systems. The scope also includes evaluation of the factors that affect the intelligibility of speech.



Price: \$120.00

ADD TO CART

Materials: intelligibility

semantically unpredictable sentence (SUS)

- read the journal paper: DOI 10.1016/0167-6393(96)00026-X
- SUS are *not* random sequences of words
 - if we did that (and ignored syntax), would be very hard for listener to process
- How can we use 'normal' sentences, but **avoid the ceiling effect**?

-
-
-
-
-
-



Materials: naturalness

- where does the text come from?
 - **randomly** selected
 - what domain?
 - newspapers
 - novels
 - **carefully** designed
 - Harvard (IEEE) sentences - phonetically balanced

APPENDIX C

1965 Revised List of Phonetically Balanced Sentences (Harvard Sentences)

List 1

1. The birch canoe slid on the smooth planks.
2. Glue the sheet to the dark blue background.
3. It's easy to tell the depth of a well.
4. These days a chicken leg is a rare dish.
5. Rice is often served in round bowls.
6. The juice of lemons makes fine punch.
7. The box was thrown beside the parked truck.
8. The hogs were fed chopped corn and garbage.
9. Four hours of steady work faced us.
10. A large size in stockings is hard to sell.

List 2

1. The boy was there when the sun rose.
2. A rod is used to catch pink salmon.
3. The source of the huge river is the clear spring.
4. Kick the ball straight and follow through.

IEEE RECOMMENDED PRACTICE FOR SPEECH QUALITY MEASUREMENTS

Materials: prosody

- does the use of a natural reference make sense here?
- will the choice of text influence listener judgements?
- removing the effects of the text
 - low-pass filtered speech
 - delexicalised speech
 - use of 'neutral' text

Materials: Blizzard Challenge

- **news** - from the Glasgow Herald newspaper
 - *“He was taken to the Western Infirmary and later released.”*
- **novel** - from out-of-copyright novels (similar to the ARCTIC corpus)
 - *“It was a blow in the face to Sheldon.”*
- **SUS** - semantically unpredictable sentences
 - *“The fire turned as the capital point.”*

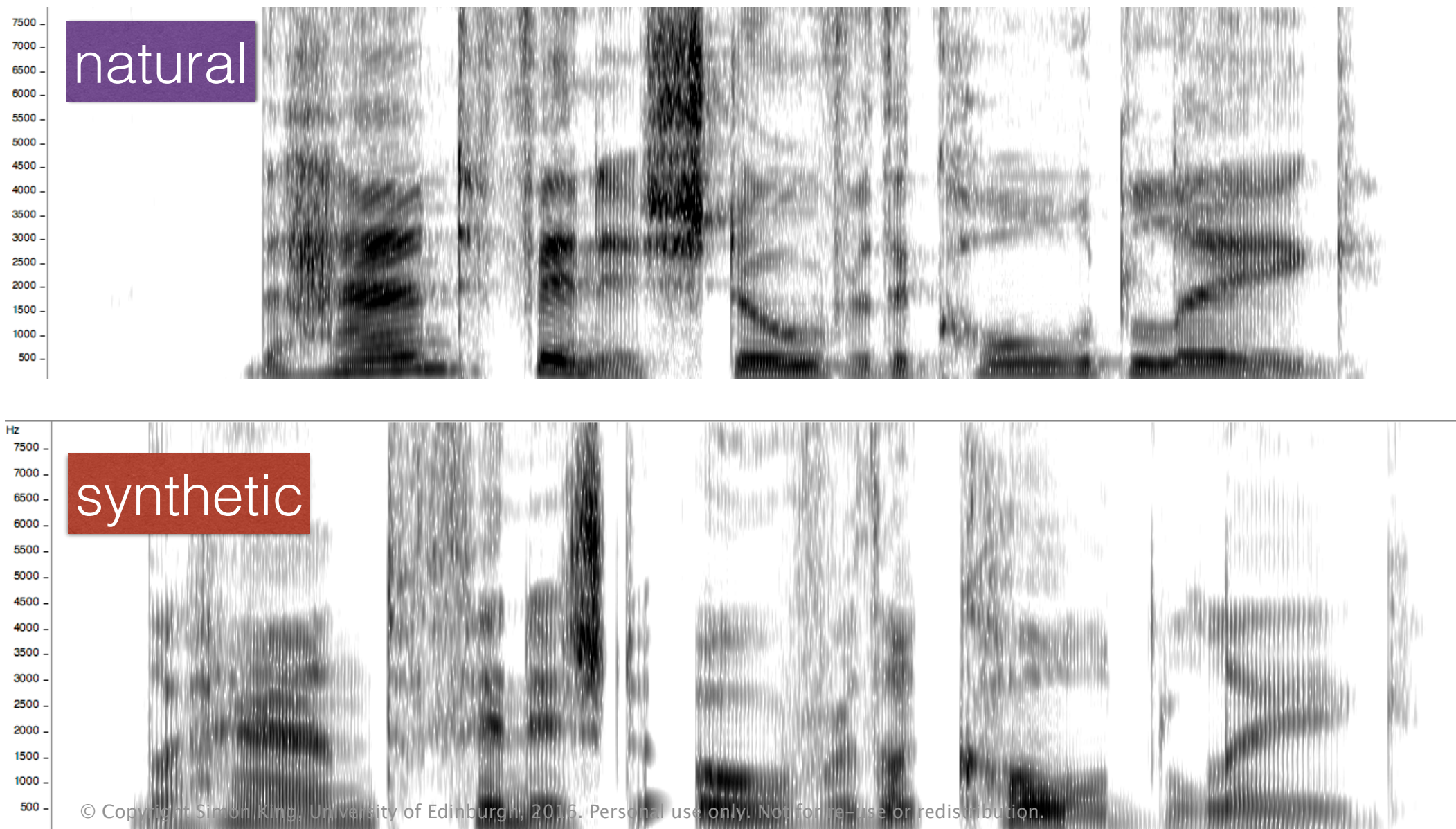
What about objective methods (i.e., no listeners) ?

- **Subjective** methods
 - play examples to listeners, obtain response
 - slow, laborious, expensive
 - generally thought reliable and useful
- **Objective** methods
 - computational, perhaps involving measuring distance between synthetic and natural versions of the same utterance
 - fast, automatic, cheap
 - not guaranteed to correlate with listeners' scores

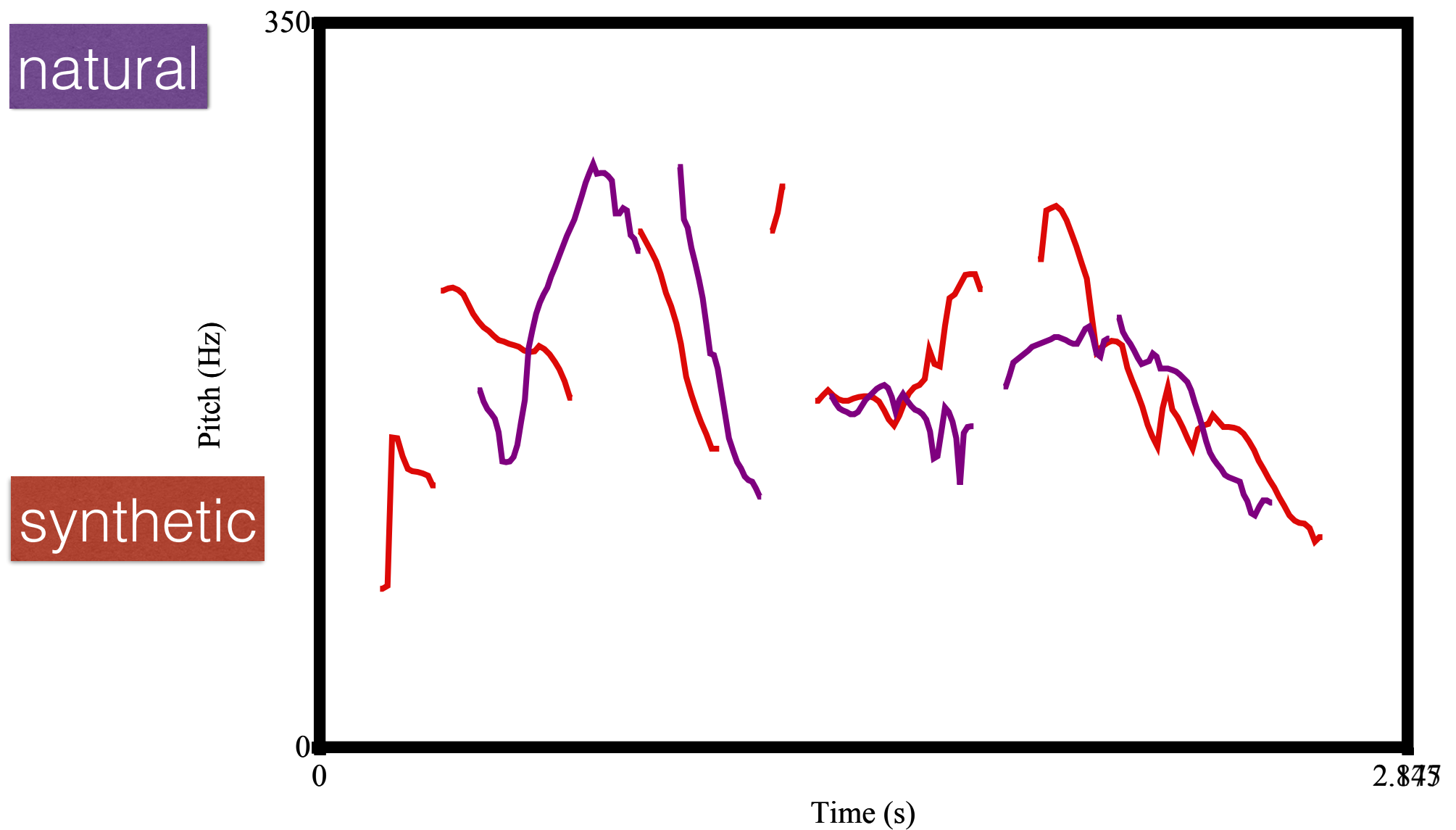
Simple objective measures

- Typically compare **acoustic properties** to natural reference samples
- Assumes that natural version is the ‘gold standard’
 - **time-align** natural and synthetic
 - then perform **frame-by-frame** comparison
- Cannot account for natural variation (but could use multiple natural examples)
- Based only on properties of the signal
 - spectral envelope: **Mel-cepstral distortion** (MCD)
 - F0 contour: **Root Mean Square Error of F0** (RMSE F0); correlation
- *Is MCD a reasonable thing to measure for unit selection synthetic speech?*

Mel-cepstral distortion



RMSE for F0



Complex objective measures: naturalness

- From the field of telecommunications
 - standardised objective measures of **speech signal quality**
 - e.g., PESQ (P.862) ; POLQA (P.863)
 - originally designed to test speech transmission over phone lines
 - PESQ is based on a weighted combination of many properties of speech, such as the higher-order statistical properties of various spectral coefficients
- PESQ does **not** well predict perceived naturalness of synthetic speech
- Modified version by Hinterleitner et al:
 - weights are tuned on previous perceptual evaluations of synthetic speech
 - reasonable predictions of naturalness for unseen samples - i.e., moderate correlation with listener scores

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Methods for objective and subjective assessment of
quality

Perceptual evaluation of speech quality (PESQ): An
objective method for end-to-end speech quality
assessment of narrow-band telephone networks
and speech codecs

Complex objective measures: intelligibility

- Originally intended to measure intelligibility of natural speech **in noise**
 - essentially measuring the predicted **audibility** of the speech, above the noise
- Simplistic measures based on the spectrum
- More complex methods, typically employing an auditory model
 - can capture effects such as auditory frequency scales, frequency masking, temporal masking, ...
 - on natural speech: good correlations with subjective intelligibility (i.e., obtained from listening tests)
 - also make reasonable predictions for synthetic speech
 - but not applicable to speech in **clean** conditions
- Can we use an ASR system?

Test sample size, significance and magnitude of effect

- Significance testing (e.g., paired t-test)
 - measures repeatability / consistency
 - says **nothing about the magnitude** of the effect!
 - very small effects can still be significant
 - e.g., all listeners thought system B was a tiny bit better than system A
 - large effects *tend* to be significant more often, but not guaranteed
 - e.g., half the listeners thought system B was a much better than system A, and half the listeners thought it was a tiny bit worse
- Magnitude of the effect
 - whether the size of the effect is meaningful is a **judgement call** and depends on the goal of the evaluation

Which type of test to use?

Type of test

What is being evaluated?

	<i>MOS</i>	<i>Task-based performance</i>	<i>Forced choice</i>
Naturalness	Yes	?	Yes
Similarity to target speaker	Yes	Maybe	Yes
Intelligibility	No !	Yes	Only for DRT/MRT
Non-specific	Maybe	Maybe	Yes

Evaluating Speech Synthesis

Case study: the Blizzard Challenge

Evaluation case study: The Blizzard Challenge

- Annual evaluation of speech synthesis systems in which participating teams build a voice for their system using a common data set
- A large online listening test is used to evaluate the systems
- Goal:
 - understand and compare research techniques
- Method:
 - build voices on a common dataset
 - evaluate them in a single listening test
- The “hub” task is to take the released speech data, build synthetic voices, and synthesize a prescribed set of test sentences.
 - There are usually also several optional “spoke” tasks

Typical timeline

- Jan/Feb 2009 Participants register for this year's Challenge
- Feb 2 2009 Databases released
- Apr 6 2009 Test sentences released
- Apr 12 2009 Deadline for submitting synthesized speech
- Apr 20 2009 Evaluation system goes live
- Jun 15 2009 End of Evaluation
- Jun 19 2009 Deadline for returning the participant questionnaire
- Jun 26 2009 Results distributed to teams
- Jul 24 2009 Workshop papers due
- Sep 4 2009 Presentation of results at a workshop

Benchmark systems

- NATURAL Natural speech from the same speaker as the corpus
- FESTIVAL The Festival unit-selection benchmark system
- HTS2005 A speaker-dependent HMM-based benchmark system
- HTS2007 A speaker-adaptive HMM-based benchmark system

2008 systems

Natural speech

Festival benchmark

HTS benchmark

IIIT

INESC-ID

CASIA

VUB

AHOLAB

SUCLAST

USTC

CSTR/Cereproc

UPC

CMU

mXac

I²R

Nokia

DFKI

TUD

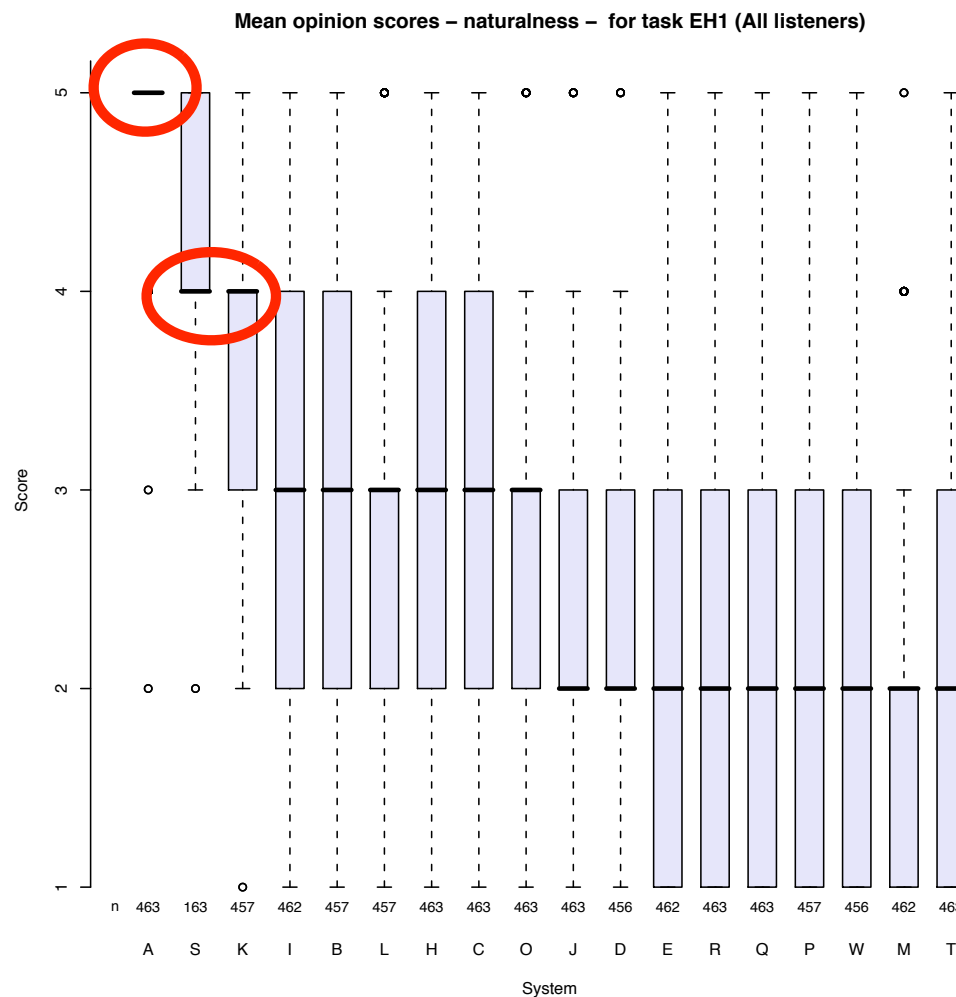
IBM

NICT/ATR

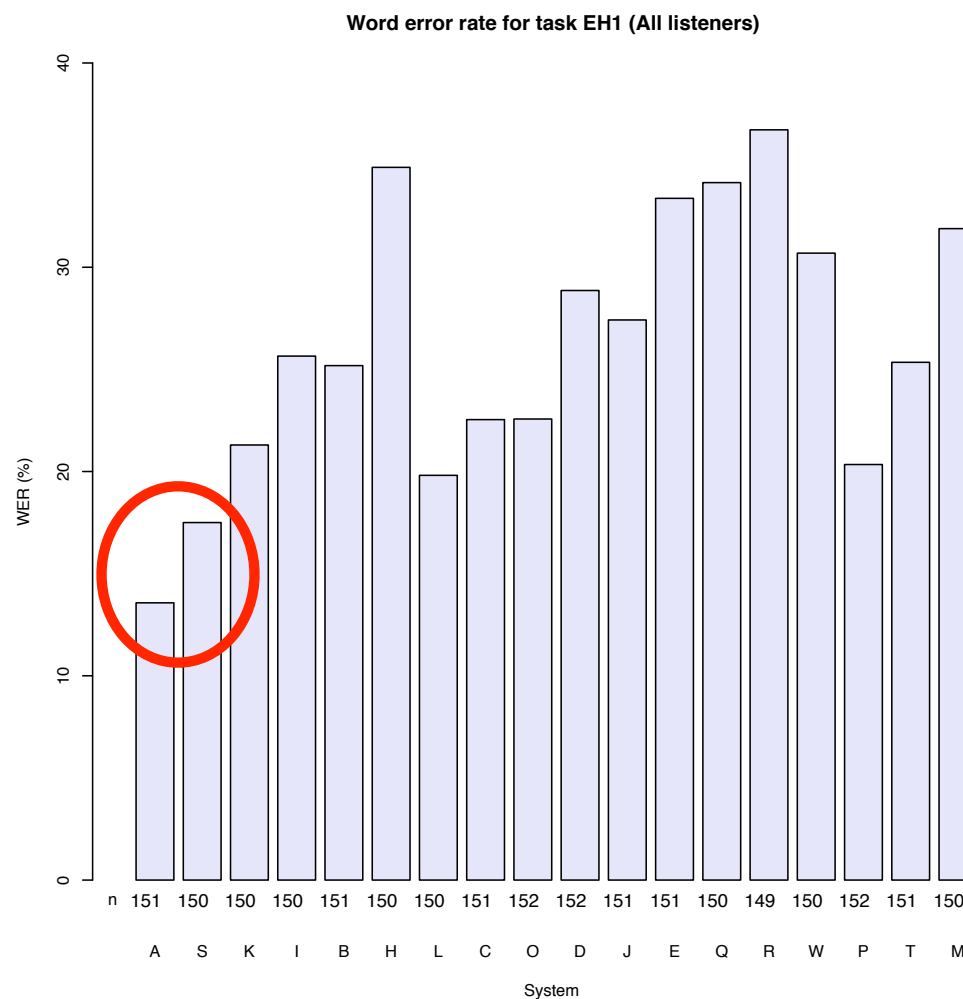
Toshiba

HTS

2009 - Results for EH1: MOS



2009 - Results for EH1: WER



2009 - Summary of results for EH1

- Natural speech is significantly more natural and more similar to the original speaker than any synthesiser
- Systems S and K are both significantly more natural and more similar to the original speaker than all other synthesisers
- System S is as intelligible as natural speech
- But there is no significant difference in intelligibility between system S and a number of other systems (B,C,K,L,O,P)
 - so we cannot state that system S is more intelligible than other systems

Evaluating Speech Synthesis

Calibration

Calibration

- Comparing across different listening tests
- Anchoring one (or both) ends of a subjective scale
- Selecting appropriate materials for SUS

Calibration: cross-test comparisons

- MOS responses are uncalibrated !
 - score for one system depends on what it is being compared against
- Blizzard Challenge
 - include a couple of benchmark systems in every test, as a crude form of calibration
 - can at least say which systems are “better than Festival”, for example
- Be *very* suspicious of papers that report a MOS score as an absolute value
 - “Our system has a MOS of 3.7 and therefore is good” !!

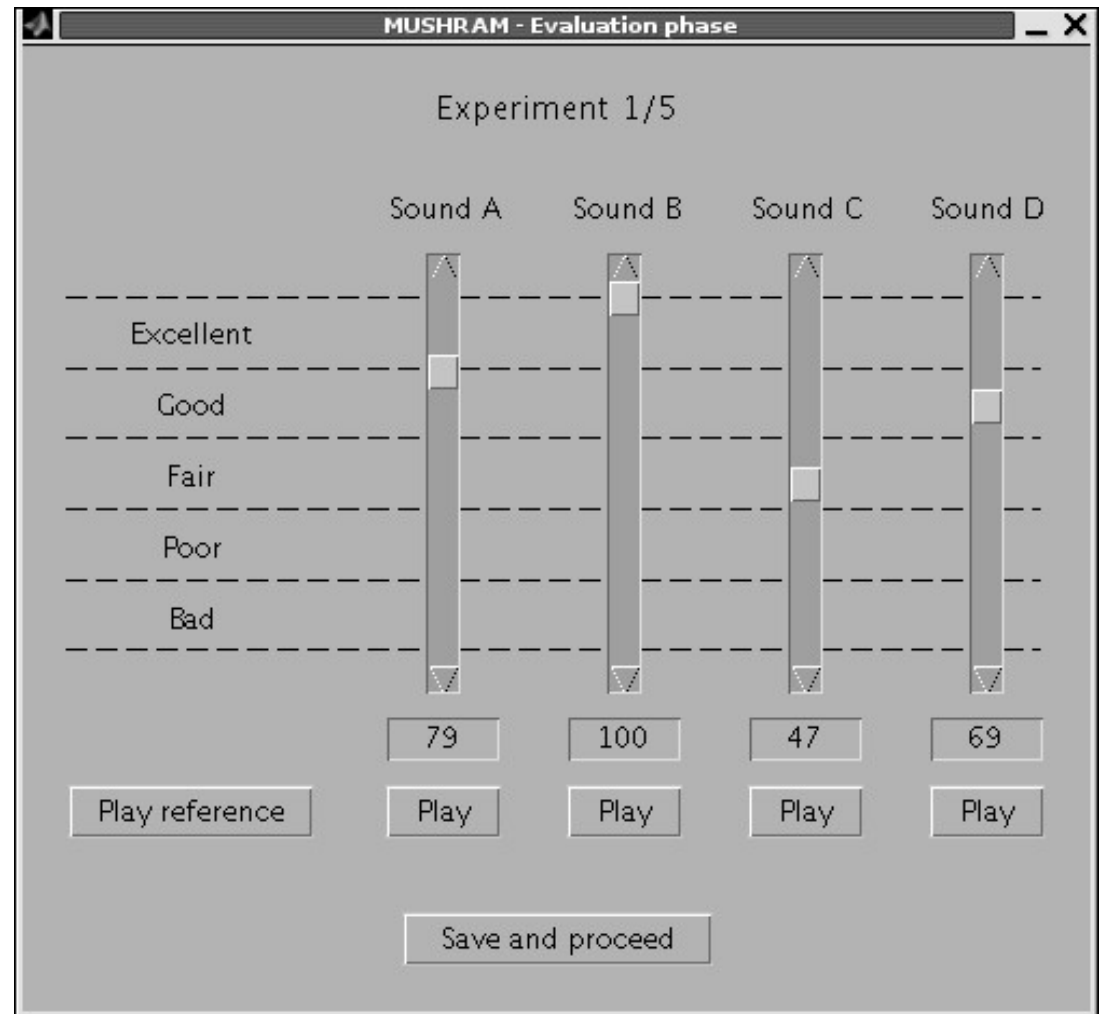
Calibration: by providing a reference

- Natural speech
 - as an **explicit and separate reference**, labelled as such to listeners
 - e.g., for speaker similarity in Blizzard Challenge
 - as '**just another system**', listeners are not aware of this fact
 - e.g., for naturalness and intelligibility in Blizzard Challenge

Calibration: by anchoring both ends of a scale

- the MUSHRA paradigm

- *MU*lti Stimulus test with Hidden Reference and Anchor
- hidden reference (e.g., natural speech)
- anchor (e.g., low-pass filtered natural speech)

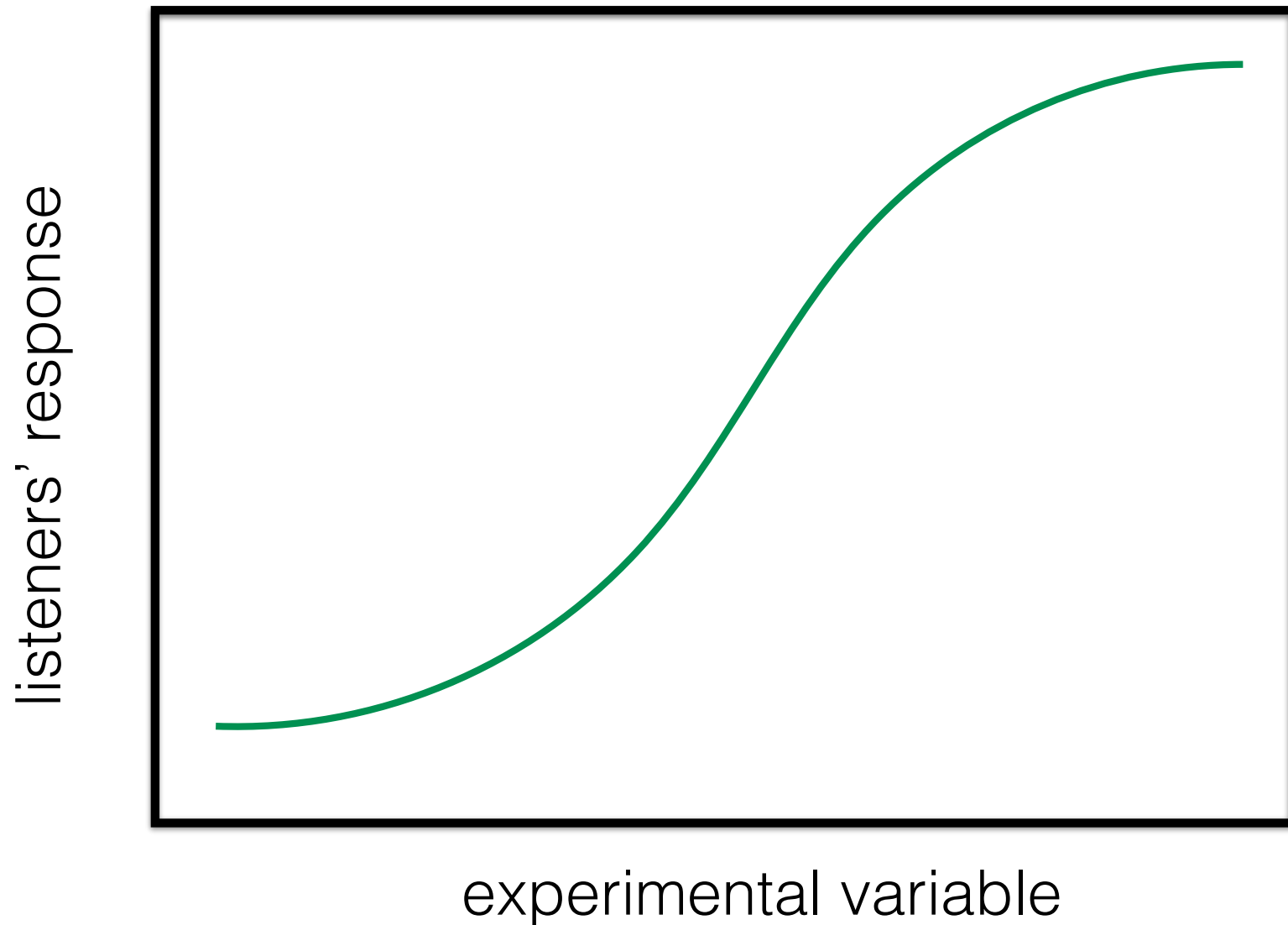


<http://c4dm.eecs.qmul.ac.uk/downloads/#mushram>

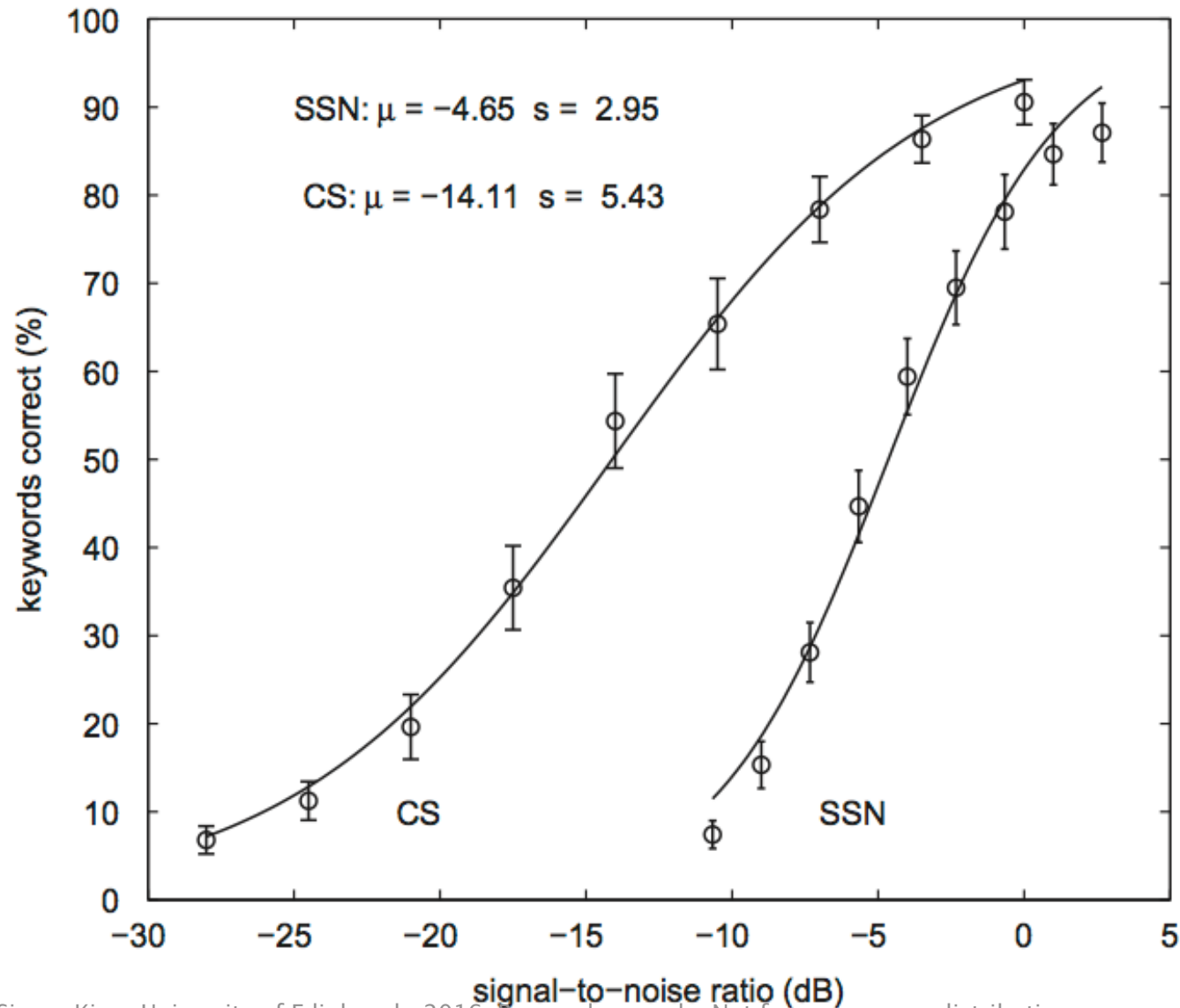
Calibration: by choosing appropriate materials

- Intelligibility tests
- Materials too **hard**?
 - Almost all responses will be wrong
- Materials too **easy**?
 - Almost all responses will be incorrect
- Will not be able to discriminate between systems being compared
 - All will appear to have **similar** intelligibility
- How can we know if the materials are too hard / too easy ?
 - run a **pre-test**

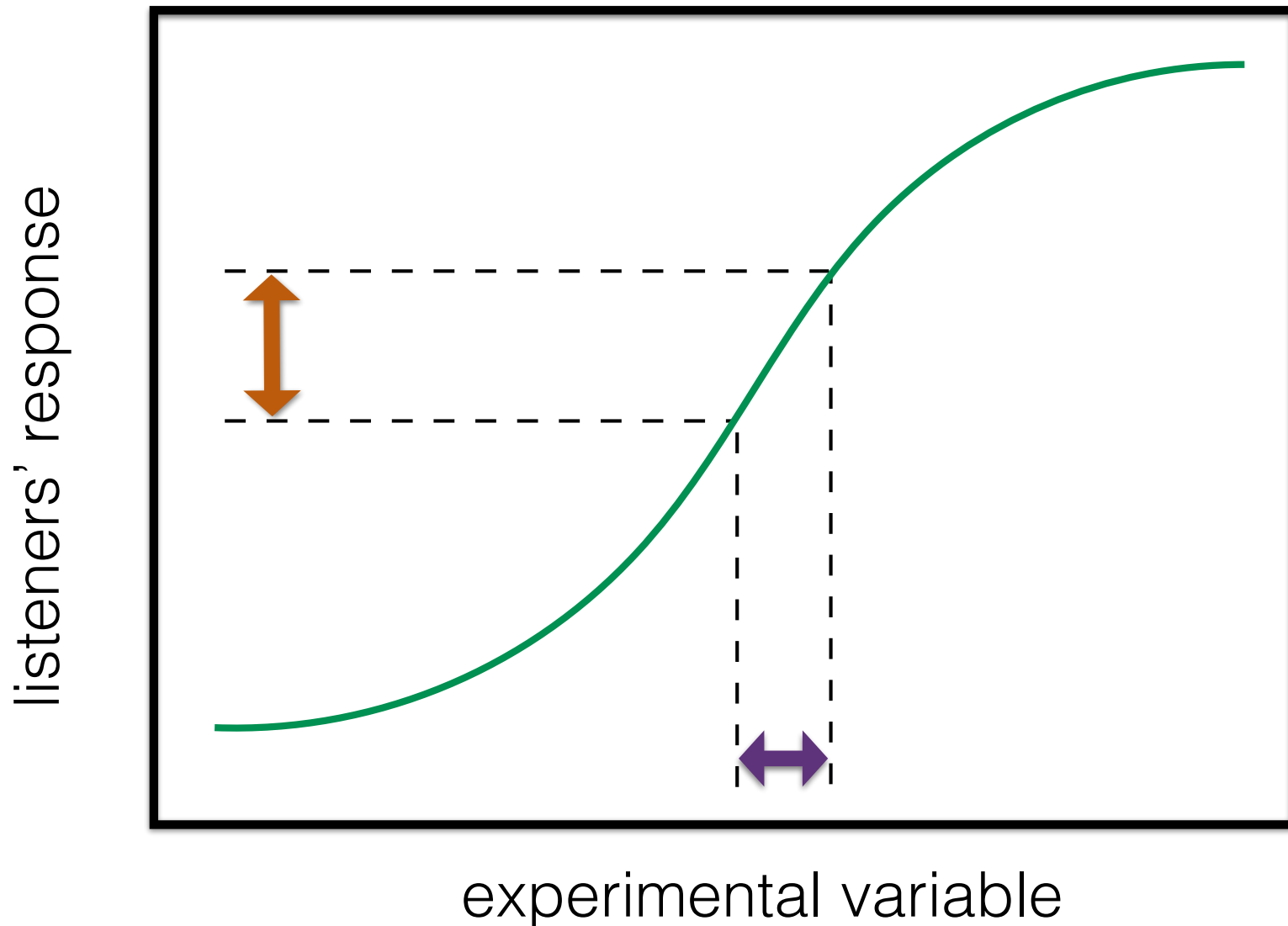
Calibration: using a psychometric function



Calibration: using a psychometric curve



Using a psychometric function to express results in more useful units



Summary

- Main form of evaluation is subjective testing
- Objective measures useful in limited cases
- Aspects of synthesis usually evaluated are
 - naturalness, intelligibility, and sometimes speaker similarity
- Would like to do diagnostic evaluation
 - discover what listeners are attending to and why they make certain judgements
- Statistics are important, especially significance testing
- Some conventions established (MOS naturalness, SUS intelligibility)
 - but not entirely satisfactory
 - little discussion in the literature (e.g., < 1% of Taylor's book devoted to the subject)

Evaluating Speech Synthesis

Blizzard Challenge audio samples

Evaluating Speech Synthesis

Blizzard Challenge trends

Rank of the Festival benchmark system (naturalness)

- 2008 - 4th of 20 = **0.2**

- 2009 - 4th of 17 = **0.2**


- 2010 - 7th of 17 = **0.4**

- 2011 - 6th of 12 = **0.5**

- 2012 - 4th of 10 = **0.4**

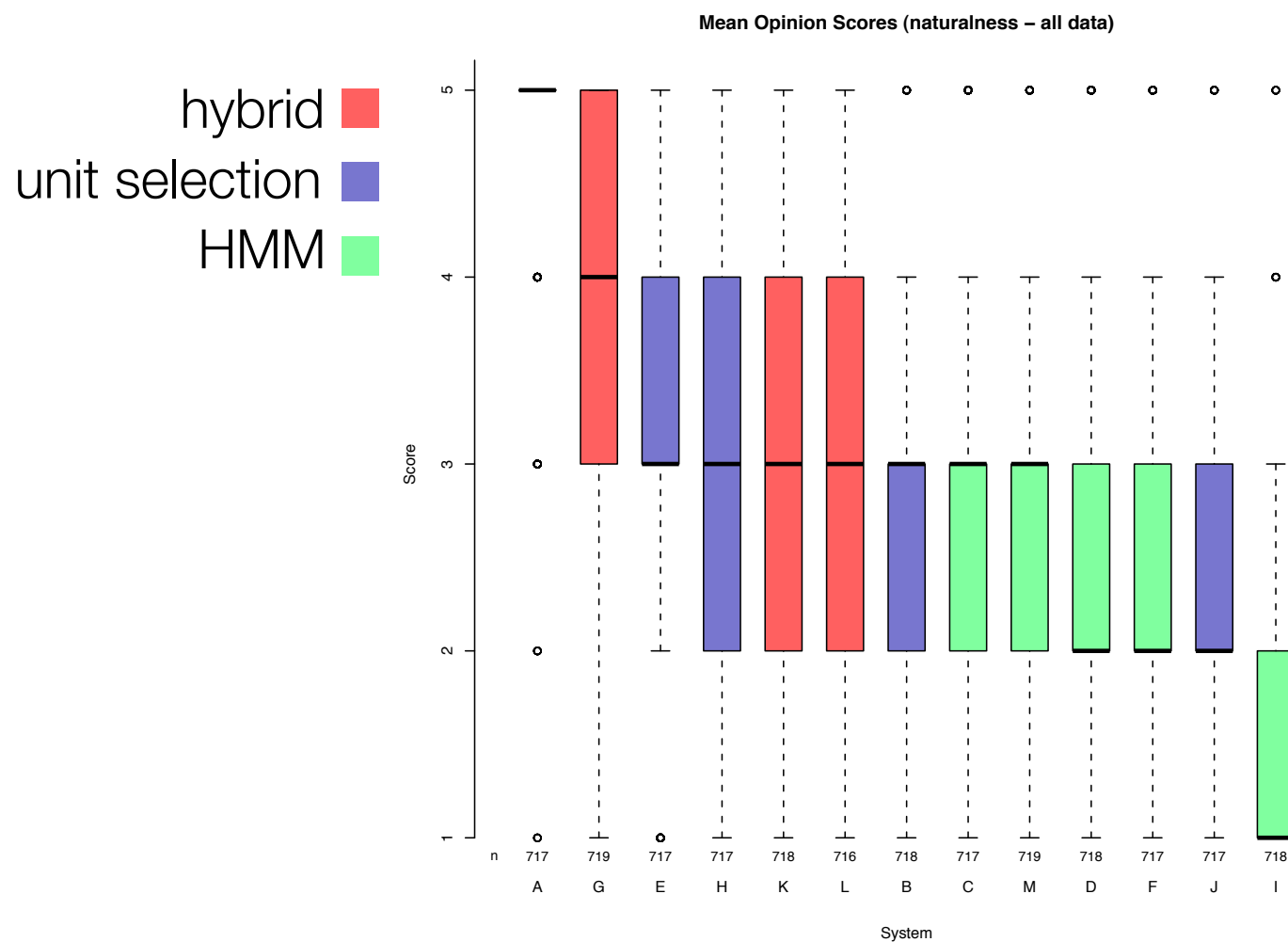
- 2013 - 7th of 10 = **0.7**

- 2016 - 4th of 16 = **0.25 !**

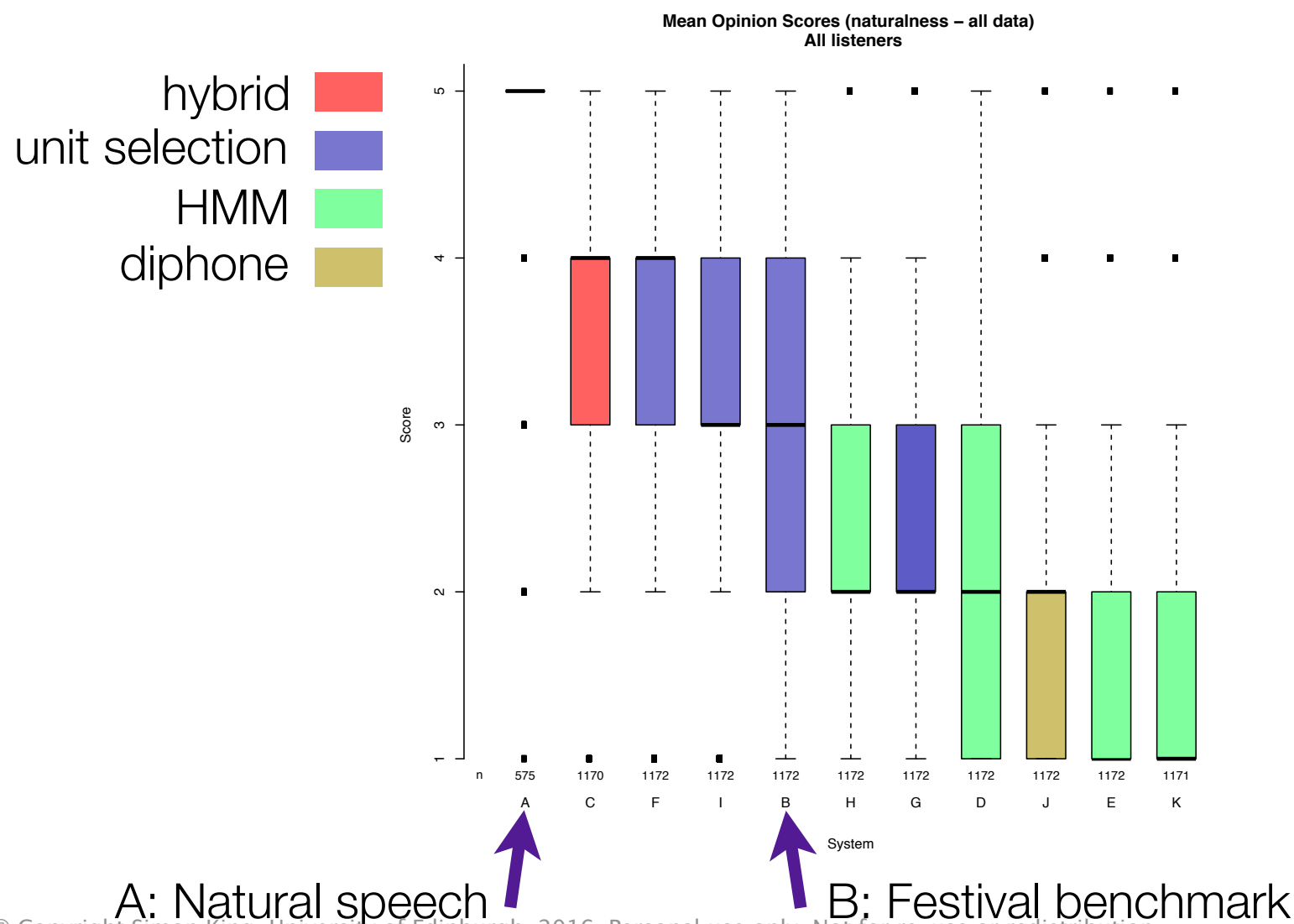


Festival benchmark
dropping down the ranking
over time

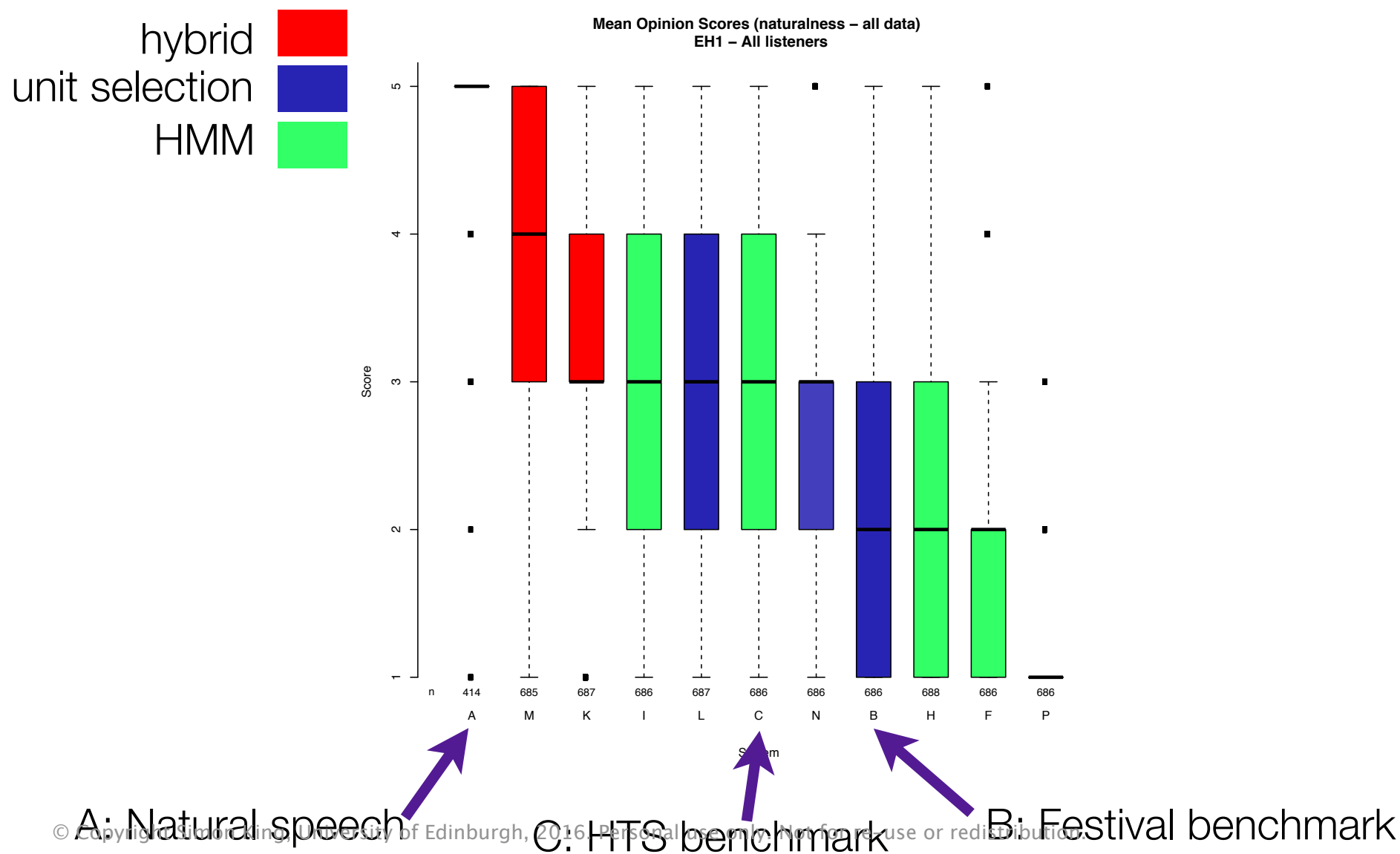
Blizzard 2011 - naturalness



Blizzard 2012 - naturalness



Blizzard 2013 - naturalness



Blizzard 2013 - intelligibility (Word Error Rate)

