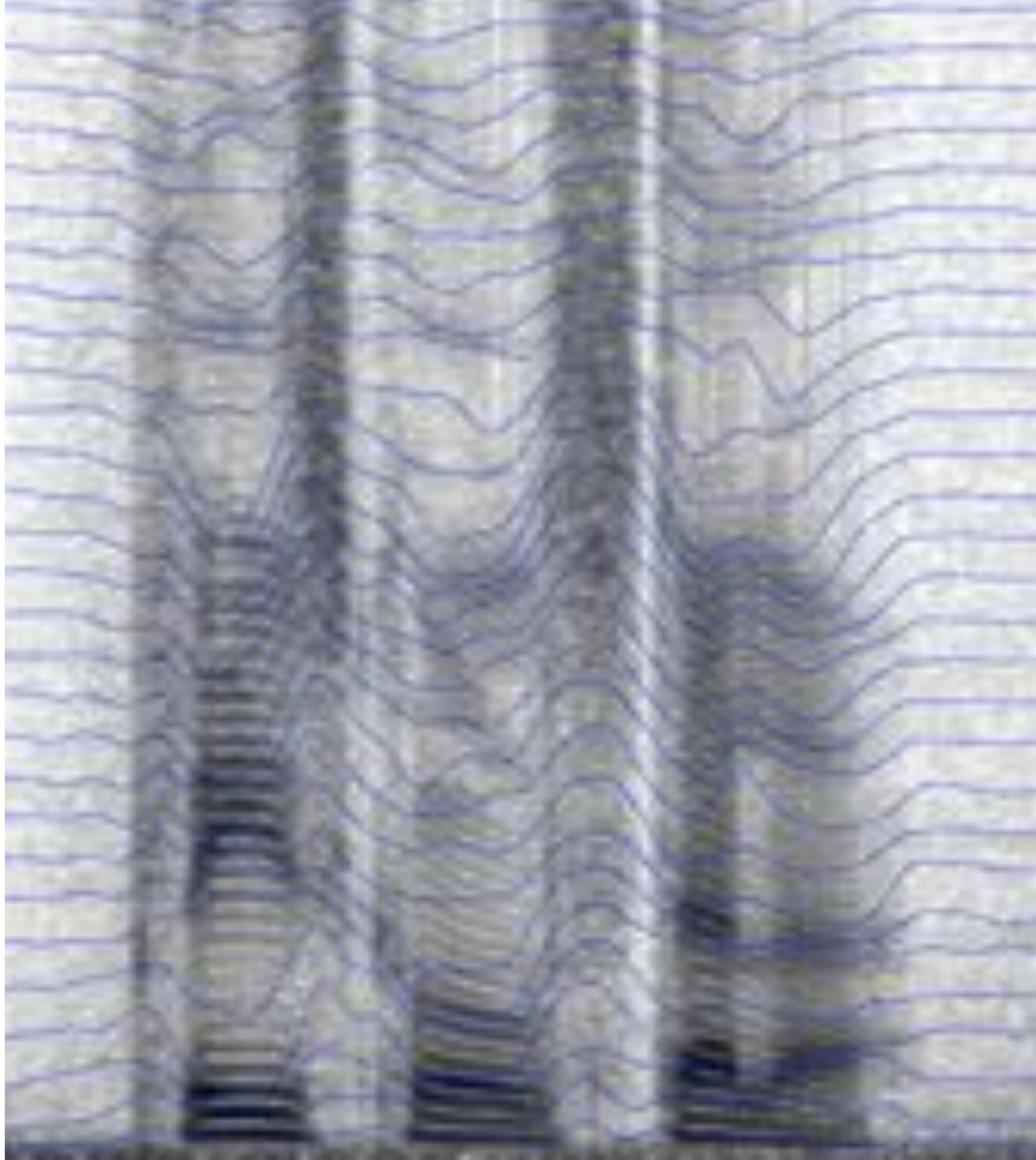
Speech Synthesis

Simon King
University of Edinburgh



Evaluation of speech synthesis

- why we need to evaluate
- when to evaluate
- which aspects to evaluate
- how to evaluate
- what to do with the outcome

What you should already know

- front end errors, such as
  - text normalisation
  - letter-to-sound
  - prosody
- unit selection errors
  - units from inappropriate contexts
  - audible joins
- acoustic phonetics & speech perception

What you should already know

- front end errors, such as
  - text normalisation
  - letter-to-sound
  - prosody

naturalness

- unit selection errors
  - units from inappropriate contexts
  - audible joins

intelligibility

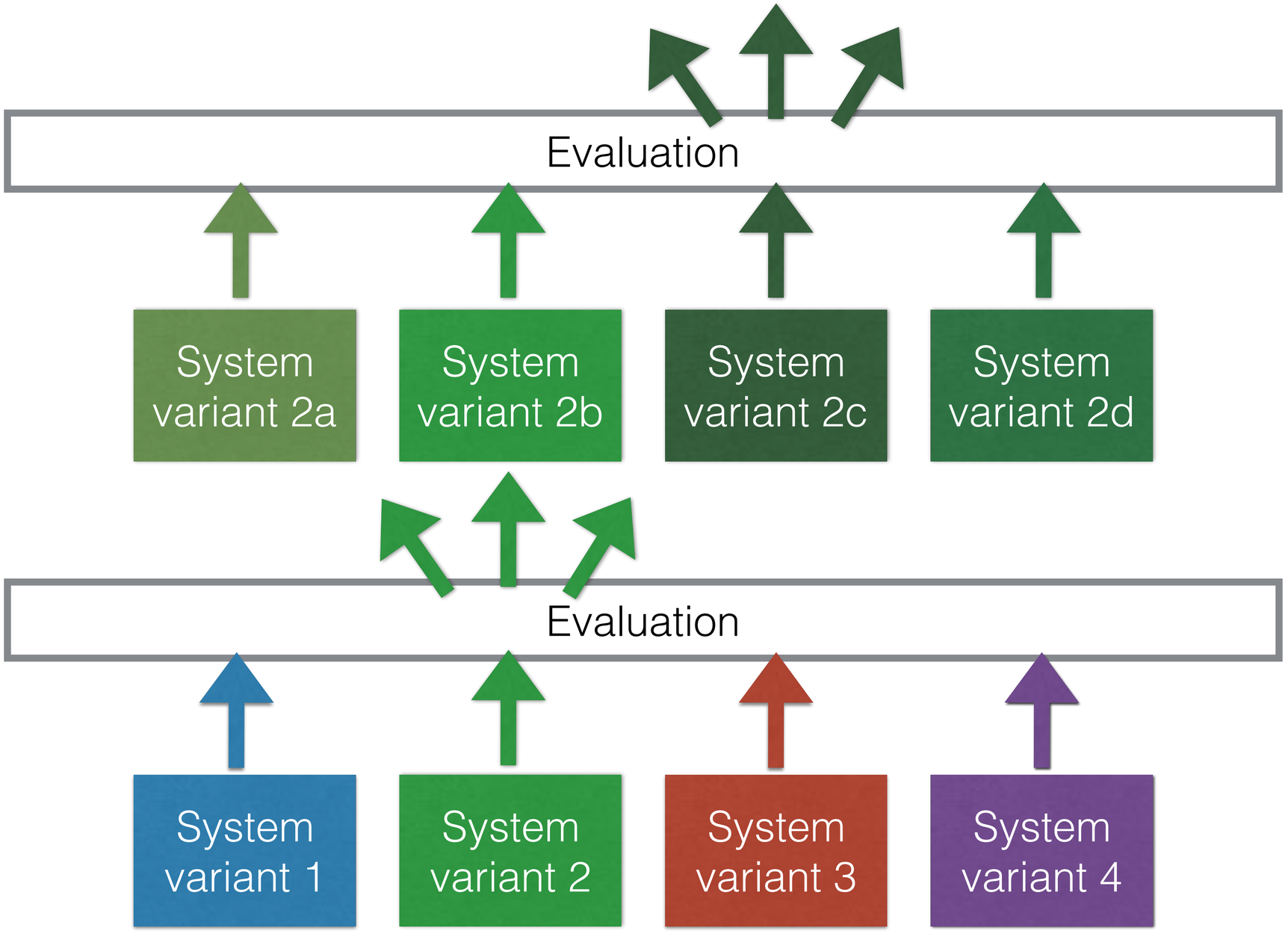
- acoustic phonetics & speech perception

Things to think about before evaluating

- **Why** we need to evaluate
  - diagnostic test to guide future development
  - comparative test against another system, or a baseline
  - pass/fail test for a product release
- **When** to evaluate
  - individual components (during development) *-or-* the finished system ?
- **Which** aspects to evaluate
  - intelligibility, naturalness, speaker similarity, ….
- **How** to evaluate
  - listener task, test design, materials used, objective measures, ….
- **What** to do with the outcome

Evaluation of speech synthesis

- why we need to evaluate
- when to evaluate
- which aspects to evaluate
- how to evaluate
- what to do with the outcome


Evaluation
System variant 2a
System variant 2b
System variant 2c
System variant 2d
Evaluation
System variant 1
System variant 2
System variant 3
System variant 4

Evaluation of speech synthesis

- why we need to evaluate
- when to evaluate
- which aspects to evaluate
- how to evaluate
- what to do with the outcome

When to evaluate

- During development
  - isolated components - e.g., number expansion, POS tagging, LTS
  - components working within a complete system - e.g., waveform generator
- After building a complete system
  - **pass/fail** - does commercial product meet user or market requirements
  - **cross-system comparisons**
    - optionally, control certain components, such as
      - a common database (as in the Blizzard Challenge)
      - fixed annotation and label alignments
      - common front end

"Unit testing"

- Glass box
  - testing the **code**
  - finding **bugs**: e.g., handle any possible text input without crashing
  - **speed**: locate slowest parts of code and optimise them

- Black box
  - measuring the **performance** of an individual component
  - if output of component is not speech, then this might be done objectively for isolated components using **gold standard** data (in terms of accuracy, precision, F-score,…)

Does improving a component guarantee to improve whole system?

- Unfortunately not ! **Interactions** between components:
  - e.g., improved text normalisation now produces word sequences that are poorly represented in the database (which was normalised with the old component)
  - e.g., improved LTS produces phoneme sequences that are poorly represented in the database (which was force-aligned using phoneme sequences from the old component)

- In general, in **pipeline architectures**
  - output of an newly improved component is the input to a subsequent component, which might have been optimised using the older version

- And of course, in software engineering, fixing one bug may **reveal** other bugs

Evaluation of speech synthesis

- why we need to evaluate
- when to evaluate
- which aspects to evaluate
- how to evaluate
- what to do with the outcome

Which aspects to evaluate

- Synthetic speech
  - Quality (whatever that is)
  - Naturalness
  - Intelligibility, or perhaps comprehension
  - Speaker similarity (which sometimes matters, but not always)
  - … can you think of others?

Probably not a single dimension

- System performance
  - speed, memory, etc.

Intelligibility vs. comprehension

- Intelligibility
  - word accuracy of **sentence transcription**
  - assume main factor is **system**, not listener
- Comprehension
  - not as clear how to measure this
  - probably mainly influenced by intelligibility
  - *may* be more influenced by **listener** factors, including cognitive abilities such as as working memory

- Measuring **listening effort** would make sense, if we could do it…

image credit: http://www.metrovision.fr

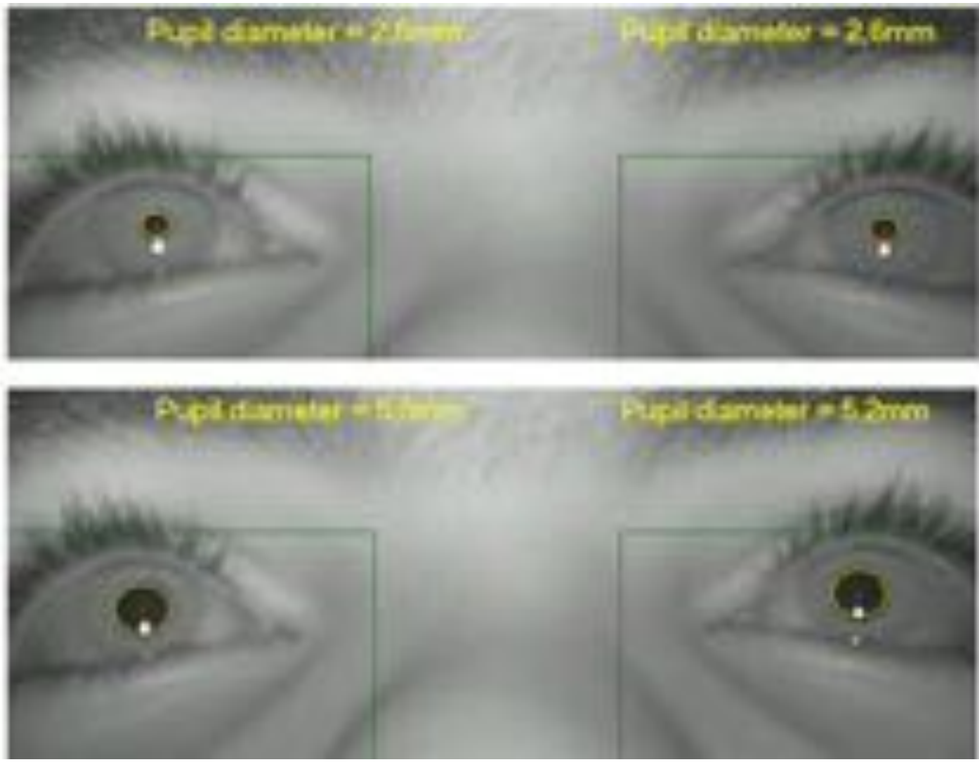


image credit: Universiteit Utrecht

Orientation

- So far we have
  - understood **why** we must evaluate
  - decided **when** to evaluate
  - listed some **aspects** of the system that we want to evaluate

- We'll concentrate on evaluation of the output from a complete TTS system, in terms of **naturalness** and **intelligibility**

- Next
  - we need to know **how** to do that

Evaluation of speech synthesis

- why we need to evaluate
- when to evaluate
- which aspects to evaluate
- how to evaluate
- what to do with the outcome

Two distinct forms of evaluation for synthetic speech

- Subjective
  - ask listeners to perform some **task**
  - test **design**
  - **materials** used

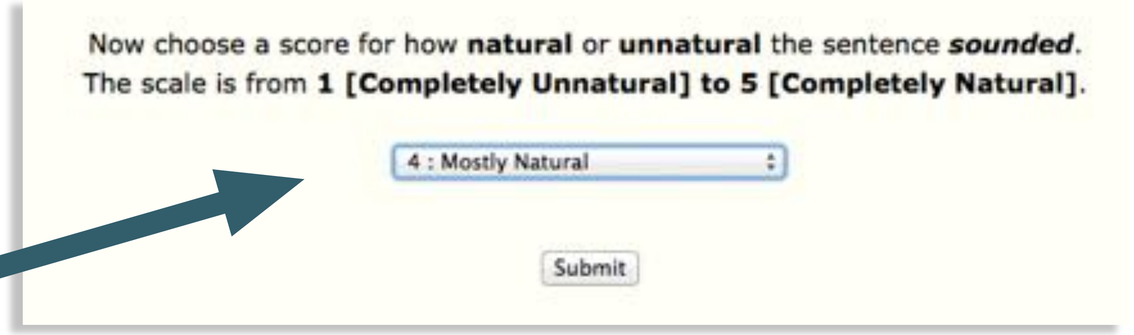
- Objective
  - simple **distances** to reference samples
  - or perhaps more sophisticated auditory **models**

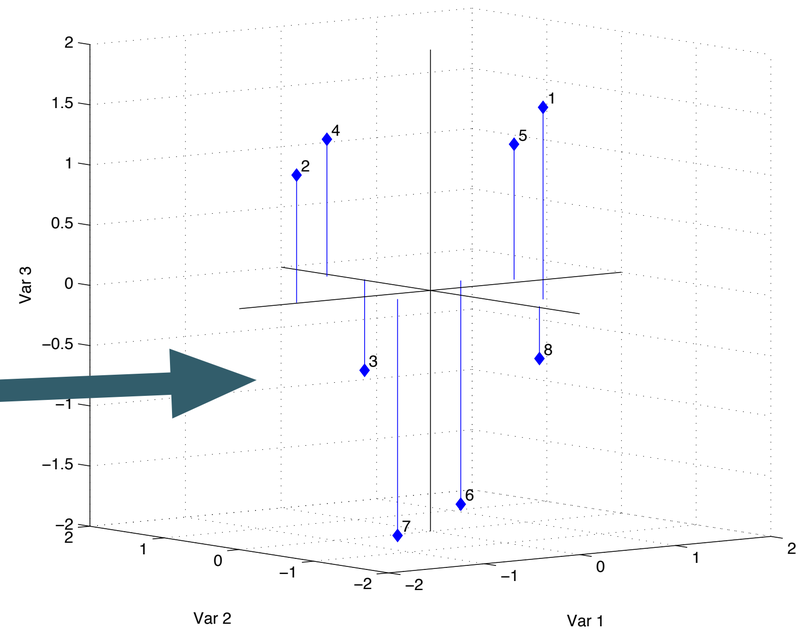
Evaluation of speech synthesis

- why we need to evaluate
- when to evaluate
- which aspects to evaluate
- how to evaluate **subjectively**
- what to do with the outcome

Listener task

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





Test design

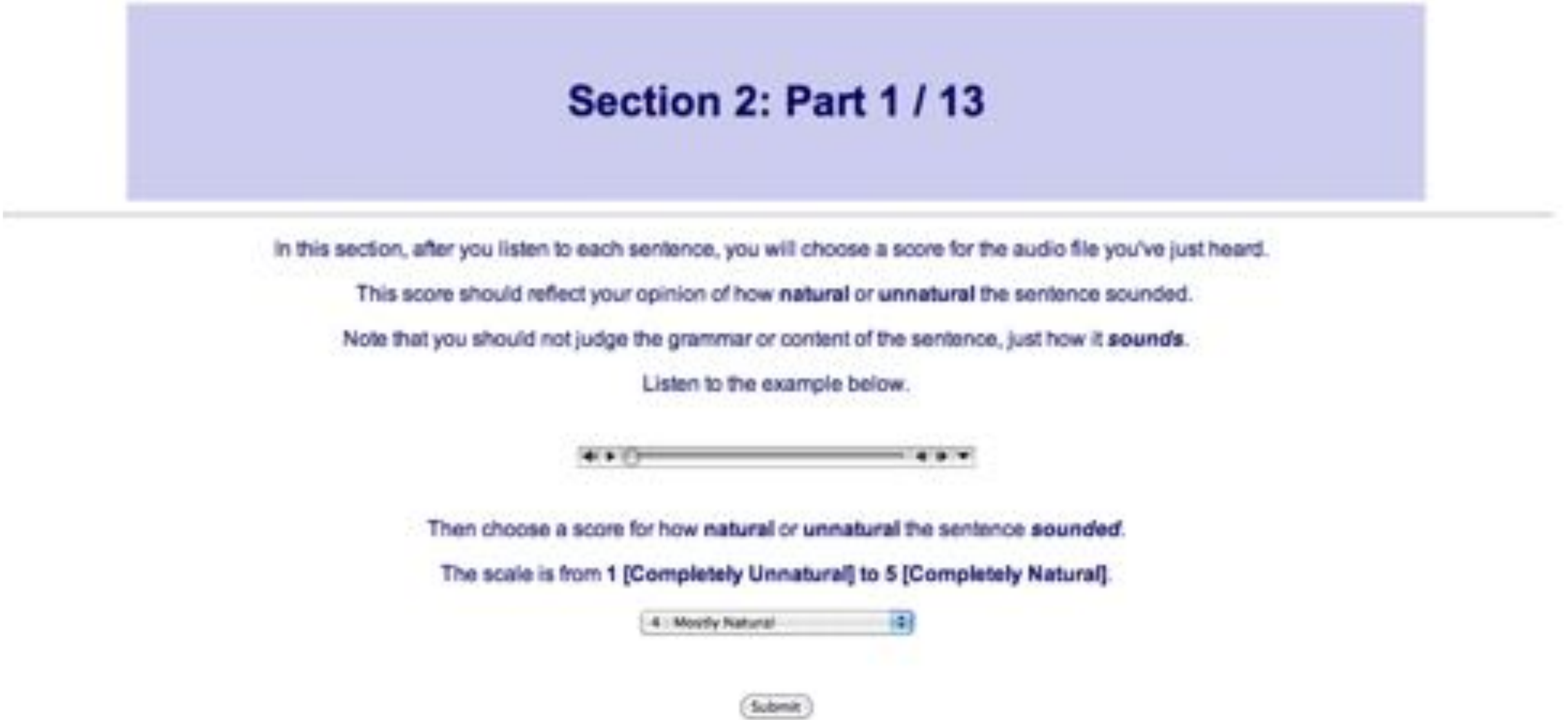
- **absolute vs. relative** judgements
  - do we need to include reference stimuli?
- **interface**
  - presenting stimuli to listeners
  - obtaining their response
- test / sample **size**
  - number of listeners, test duration per listener, number of stimuli per listener and in total
- the listeners ("**subjects**")
  - type of listener, how to recruit them, quality control of their responses

Test design: absolute vs relative judgements

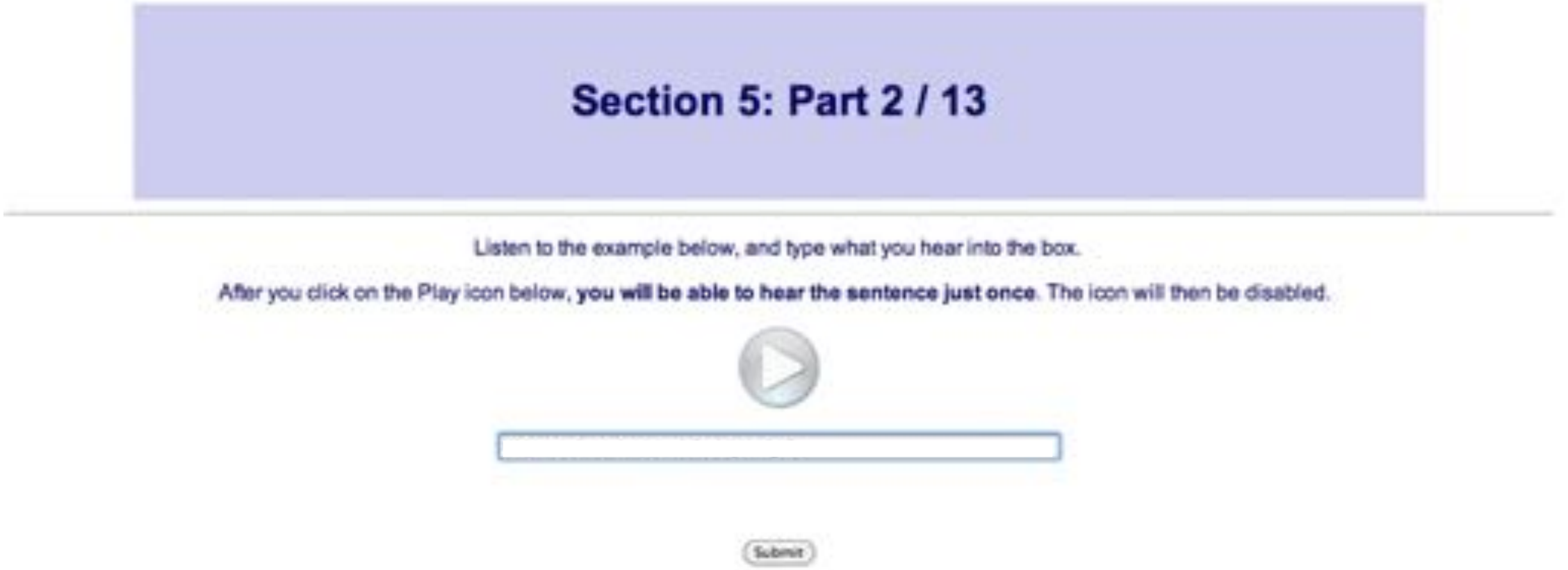
- Absolute - in other words, listeners rate a **single**, isolated stimulus
  - Mean Opinion Score (MOS)
    - note: "absolute" does not necessarily imply "repeatable" or "comparable"
  - type-in tests for intelligibility

- Relative - listeners compare **multiple** stimuli
  - pairwise "which is most natural?"
    - forced choice, or allow a 3rd "equally natural" option
  - **more than two** stimuli, optionally including **references** (lower and/or upper)
    - rating (e.g., multiple MOS), ranking , sorting

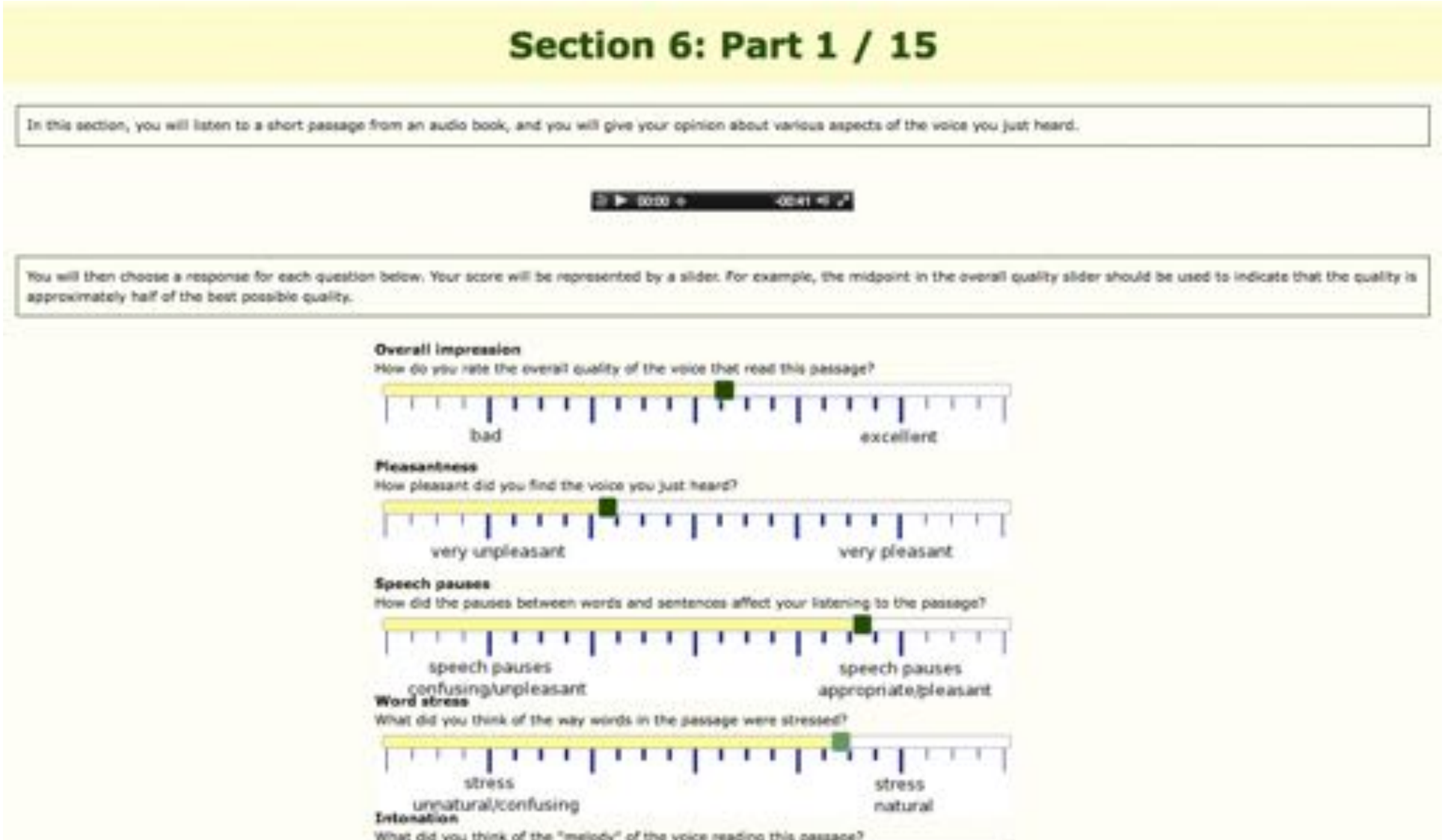
Test design: interface for Mean Opinion Score



Test design: interface for a type-in test



Test design: interface for multiple dimensions



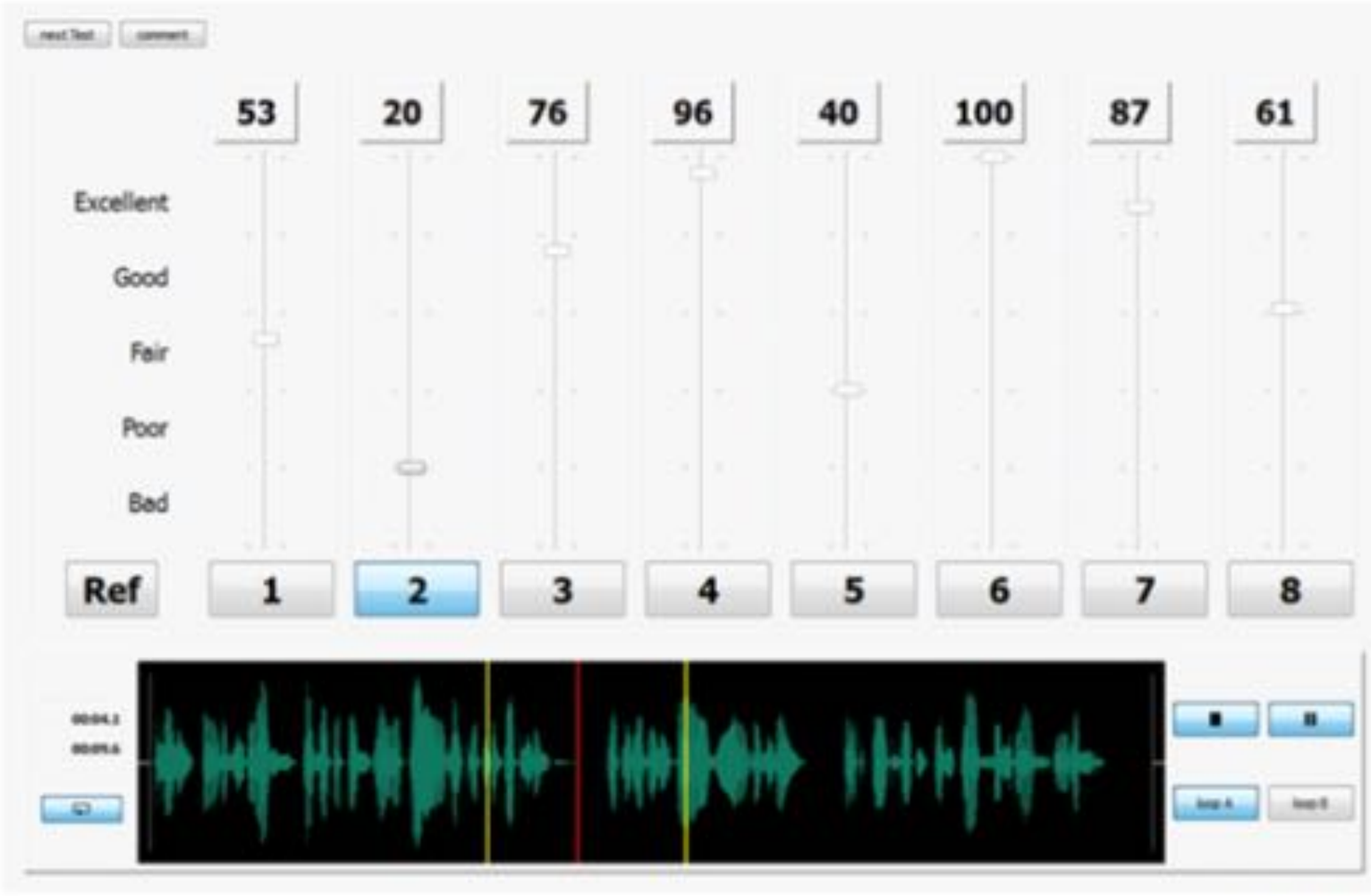
Test design: MUSHRA

BS.1534 : Method for the subjective assessment of intermediate quality levels of coding systems

Recommendation BS.1534-3 (10/2015) of the International Telecommunication Union

FIGURE 2

Example of a computer display used for a MUSHRA test



BS.1534-02

Test design: size

- Sample size determines statistical power - needed to discover **significant** differences
  - but this is not a course on statistics !

- Rules of thumb
  - maximum test duration 45 minutes, for paid listeners in a controlled environment
    - much less for, say, online (remote) listeners
  - at least 20 listeners, and preferably more
  - as many different sentences as possible, to mitigate the effects of any atypical ones

Test design: *within* vs. *between* subjects designs

- Simplest design is **within subjects**
  - all listeners hear exactly the same stimuli, possibly in (randomly) different orders
- But, what if there are **too many** systems to fit into a 45 minute test?
- Or, what if there might be **priming** or **ordering** effects? e.g., in an intelligibility test
  - must have *every* sentence synthesised by *every* system, to be fair
  - but cannot repeat the same sentence to any individual subject: they will **remember** it

	Sentence A	Sentence B	Sentence C	Sentence D	Sentence E
Subject I	0	4	3	1	2
Subject II	2	1	0	3	4
Subject III	4	3	2	0	1
Subject IV	1	0	4	2	3
Subject V	3	2	1	4	0

Orientation

- We're currently talking about subjective evaluation

- So far we have covered
  - the listeners' task
  - the test interface
  - using a between subjects design, if necessary

- Next, we must decide what materials to use

Materials

- 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
    - might even design around minimal pairs (e.g., DRT, MRT)
  - full sentences
    - errors will be more variable & harder to analyse
    - much more natural task for the listener, perhaps closer to target domain

Materials: intelligibility

- ‘normal’ material - e.g., sentences from a newspaper
  - tend to get a ceiling effect, due to interference from semantics (predictability)
- **Semantically Unpredictable Sentences** (SUS)
  - e.g., “The unsure steaks overcame the zippy rudder”
  - not representative of actual system usage, but avoids ceiling effect

- Diagnostic Rhyme Test (DRT) or **Modified Rhyme Test** (MRT) uses minimal pairs
  - e.g., “Now we will say cold again.” “Now we will say gold again.”
  - specific to individual phonemes - a diagnostic unit test
  - very time consuming and therefore rarely used

Materials: intelligibility - examples of Semantically Unpredictable Sentences

How shall a milk force the umbrella?

Drop a dry floor off the mail.

The egg that knelt earns the tables.

Why should meats toss wars?

A beef that posed opens the coats.

The lips rolled the apartment that wins.

The cars wept outside a rich capital.

The fast meals viewed a wood.

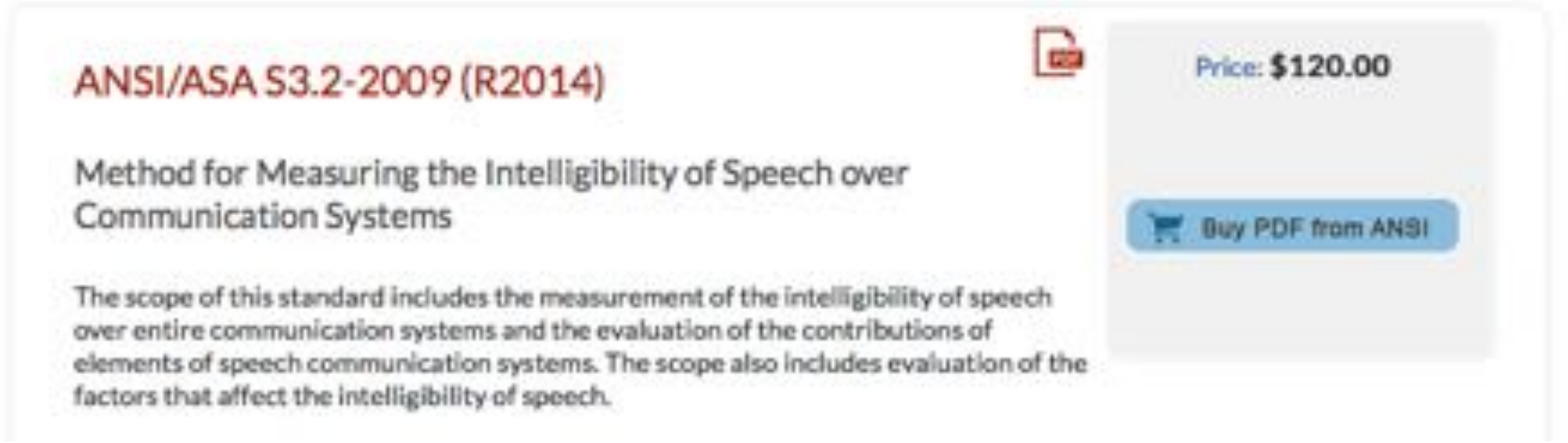
The balloons dreamed under a brown hair.

Spare the pages and the honest bodies.

A gun fears a fine sarcasm.

A smart place comes inside the telephones.

The brown cars brought a ship.



from the American National Standards Institute

Materials: intelligibility - other ways to avoid a ceiling effect

- Add noise
- Induce additional cognitive load with another task in parallel
- … can you think of any more?

Materials: naturalness

- "Randomly" selected text
  - what domain?
    - newspapers?
    - novels?
- Carefully designed text
  - e.g., Harvard (IEEE) sentences
    - in phonetically balanced lists

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

Evaluation of speech synthesis

- why we need to evaluate
- when to evaluate
- which aspects to evaluate
- how to evaluate **objectively**
- what to do with the outcome

Two distinct forms of evaluation for synthetic speech

- Subjective
  - ask listeners to perform some **task**
  - test **design**
  - **materials** used

- Objective
  - simple **distances** to reference samples
  - or perhaps more sophisticated auditory **models**

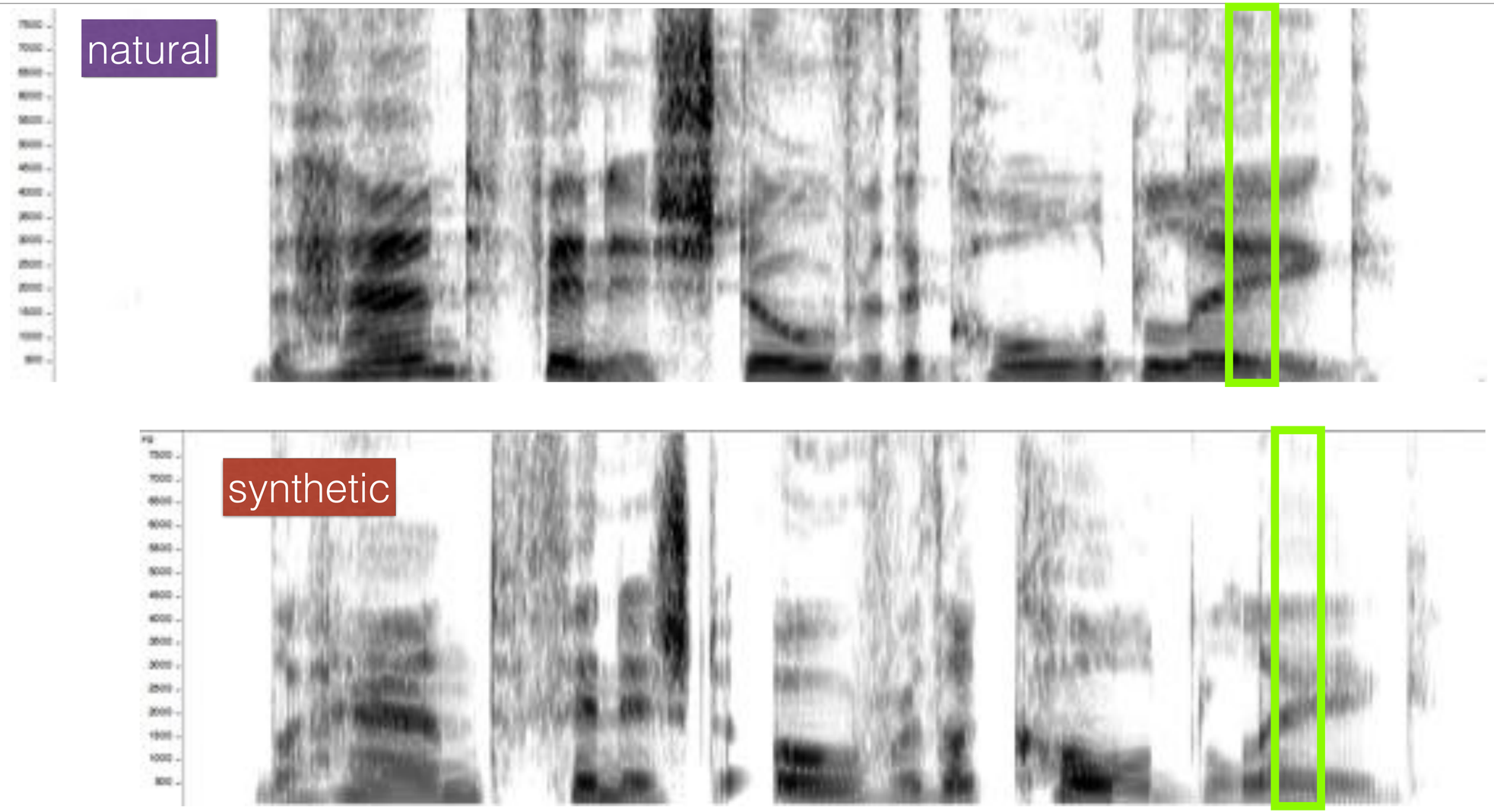
Simple objective measures

- Compare acoustic properties to a natural reference sample
  - assumes that natural version is the 'gold standard'
- **Time-align** natural and synthetic
  - perform frame-by-frame comparison, sum up **local** differences
- Does not account for natural variation (could use multiple natural examples)

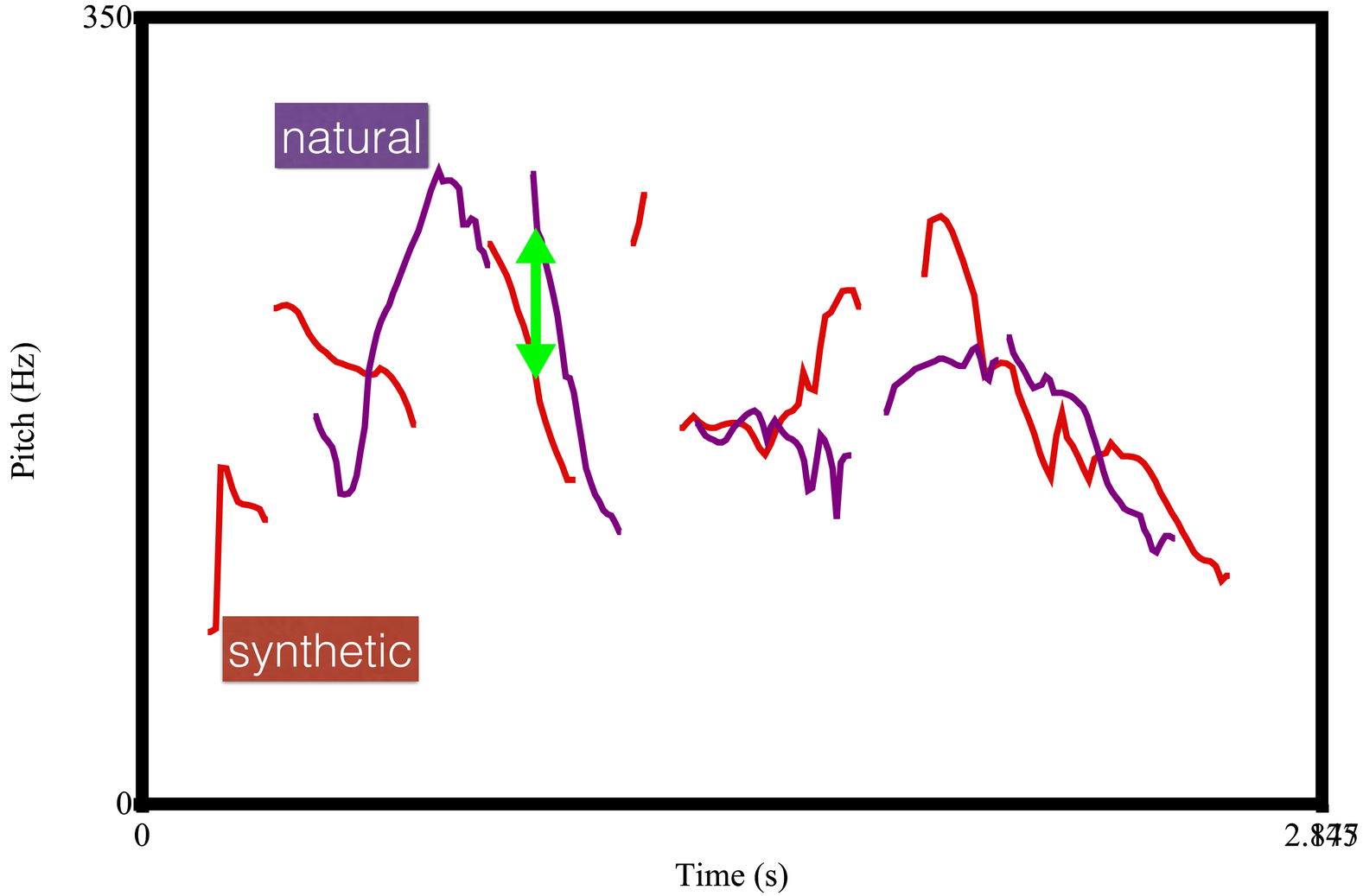
- Based only on properties of the signal
  - **spectral envelope**: Mel-Cepstral Distortion (MCD)
  - **F0 contour**: Root Mean Square Error of F0 (RMSE F0) and/or correlation

  which do not correlate perfectly with human perception

Simple objective measures: Mel-Cepstral Distortion (MCD)



Simple objective measures: Root Mean Square Error (RMSE) of F0



Complex objective measures

- Borrowed from the field of telecommunications
  - originally designed for distorted natural speech (not the same as synthetic speech!)
  - e.g., PESQ (P.862) or POLQA (P.863)

- PESQ is based on a weighted combination of differences in many properties of speech, such as the higher-order statistical properties of various spectral coefficients
  - does **not** well predict perceived naturalness of synthetic speech
  - modified versions do work for synthetic speech (refer to readings for this module)

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.862

Amendment 1

(03/2003)

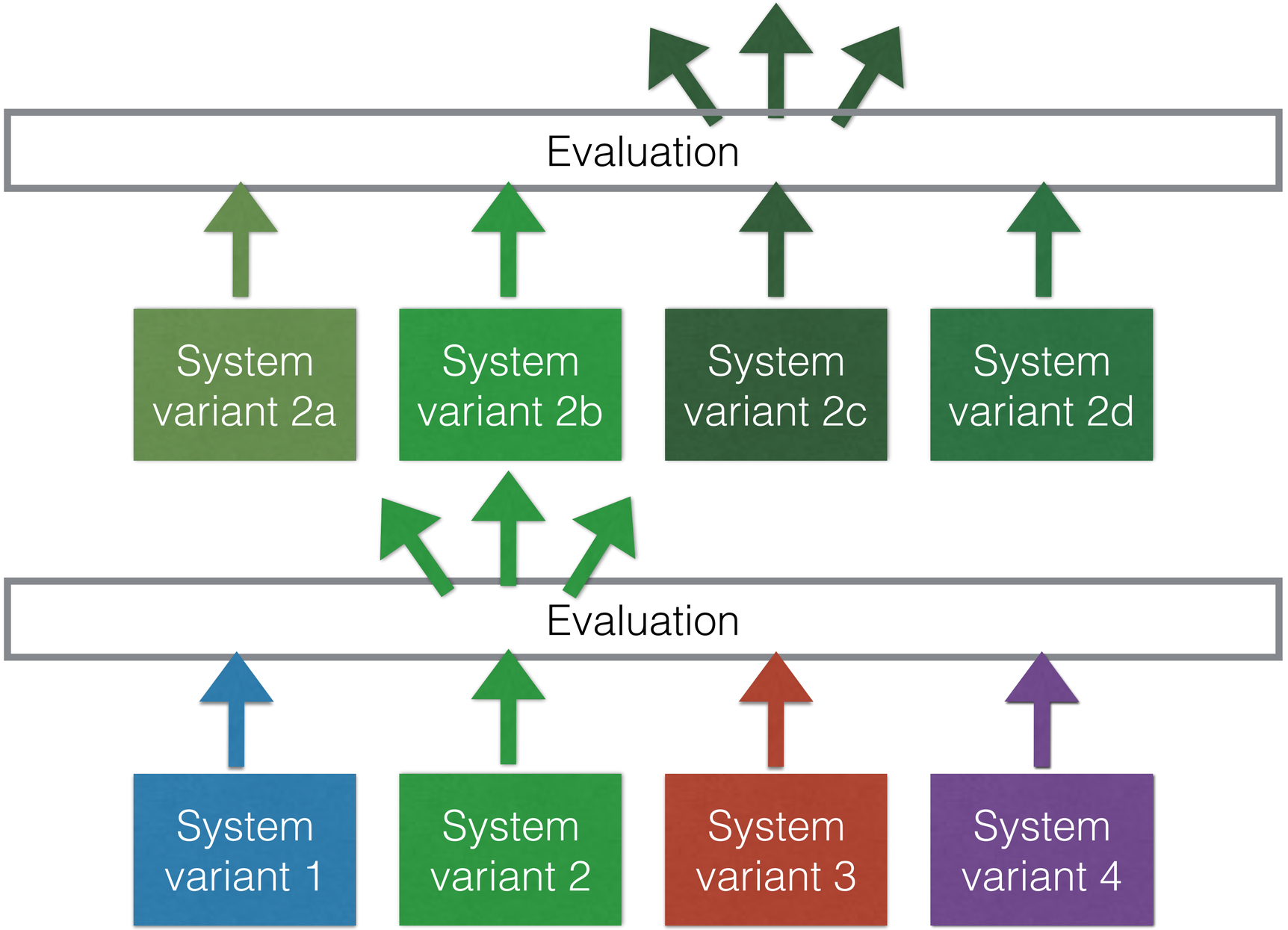
SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

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

Evaluation of speech synthesis

- why we need to evaluate
- when to evaluate
- which aspects to evaluate
- how to evaluate
- what to do with the outcome


Evaluation
System variant 2a
System variant 2b
System variant 2c
System variant 2d
Evaluation
System variant 1
System variant 2
System variant 3
System variant 4

Type of test

What is being evaluated?

	MOS or MUSHRA	Performance on a task	Forced choice
Naturalness	Yes	?	Yes
Similarity to target speaker	Yes	?	Yes
Intelligibility	No !	Yes	Only for DRT/ MRT
Non-specific preference	Maybe	?	Yes

What next?

- This concludes the first part of the course:
  - we know how to build a unit selection voice,
  - and we know how to evaluate it.

- The next part of the course covers statistical parametric methods for speech synthesis

What next?

- This concludes the first part of the course:
  - we know how to build a unit selection voice,
  - and we know how to evaluate it.

Go and put these into practice, in the "build your own unit selection voice" exercise !

- The next part of the course covers statistical parametric methods for speech synthesis