Statistical Parametric Speech Synthesis - from regression trees to DNNs

- Class slides

What we'll cover today

- **Quick recap**

- Discussion points and exercises on DNN-based TTS

- Lab report, experiments, and write-up
  - marking sheet with Q&A

What is a simple feedforward neural network?

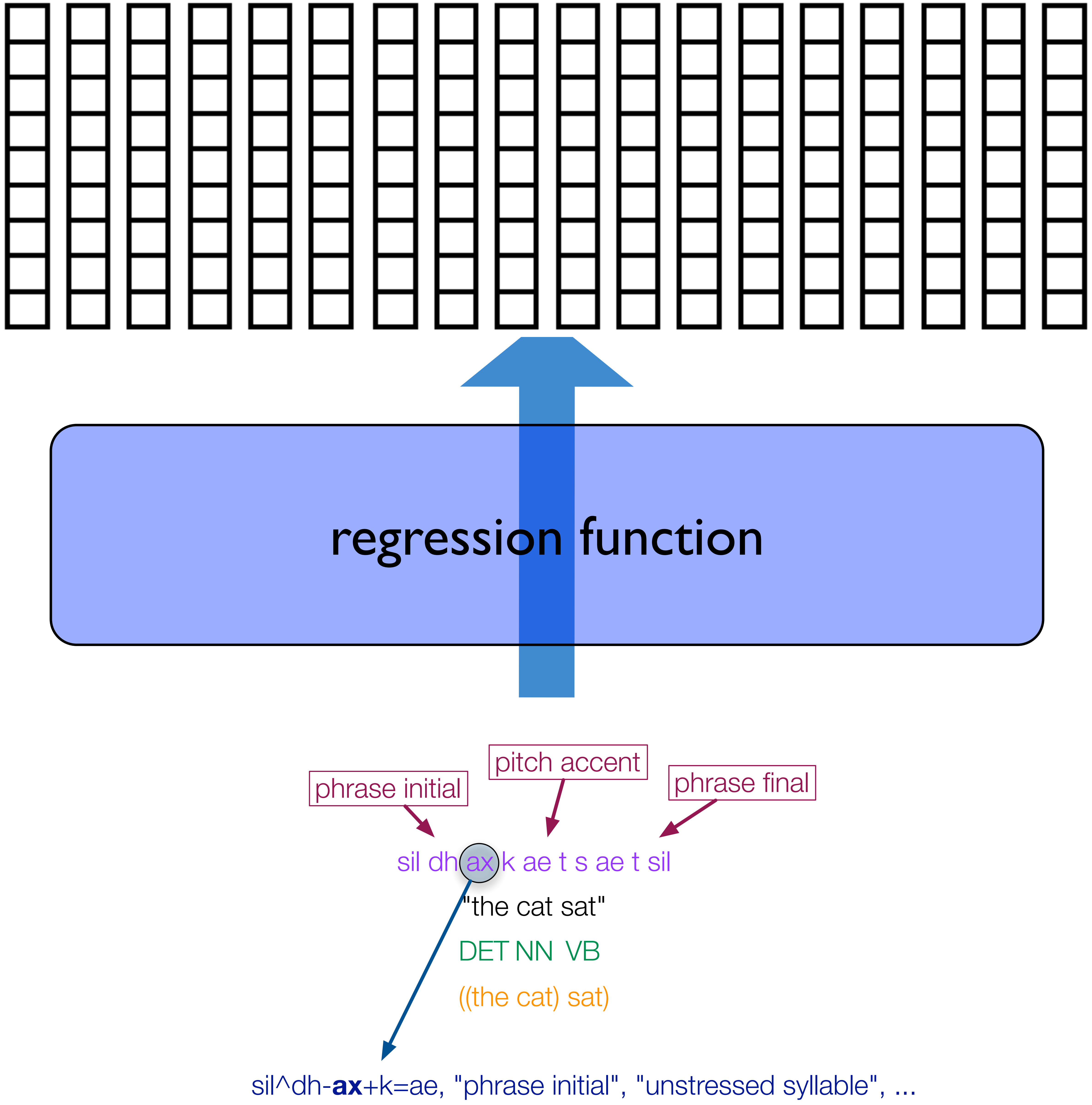
- input/output representations

- the anatomy of a unit (or more rarely now "neuron")
  - incoming weights, activation, activation function, output

- combining multiple units into a layer
- stacking layers to make a network

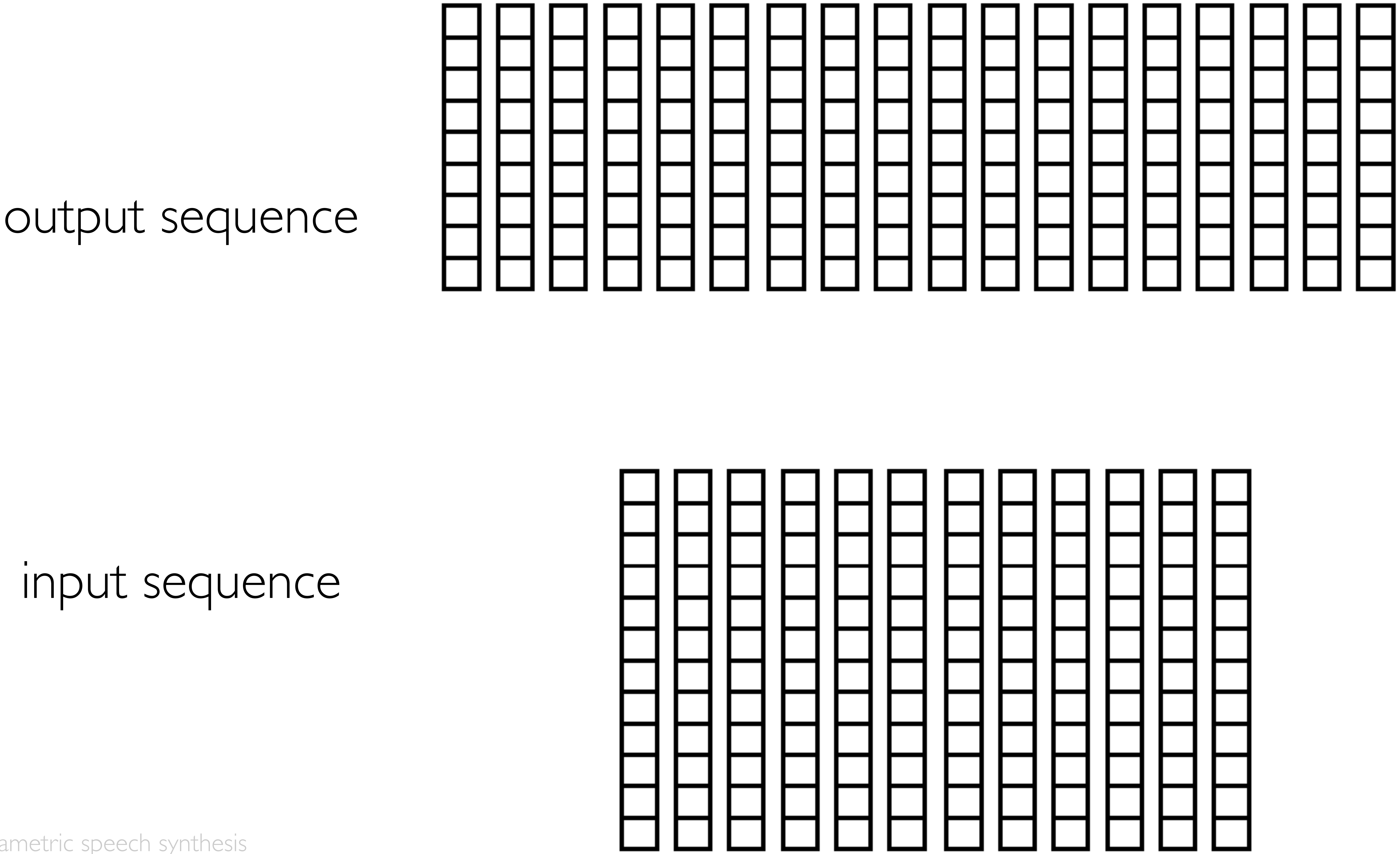
- "Information flow"

Orientation



- Statistical parametric synthesis
  - predict **speech parameters** from **linguistic specification**

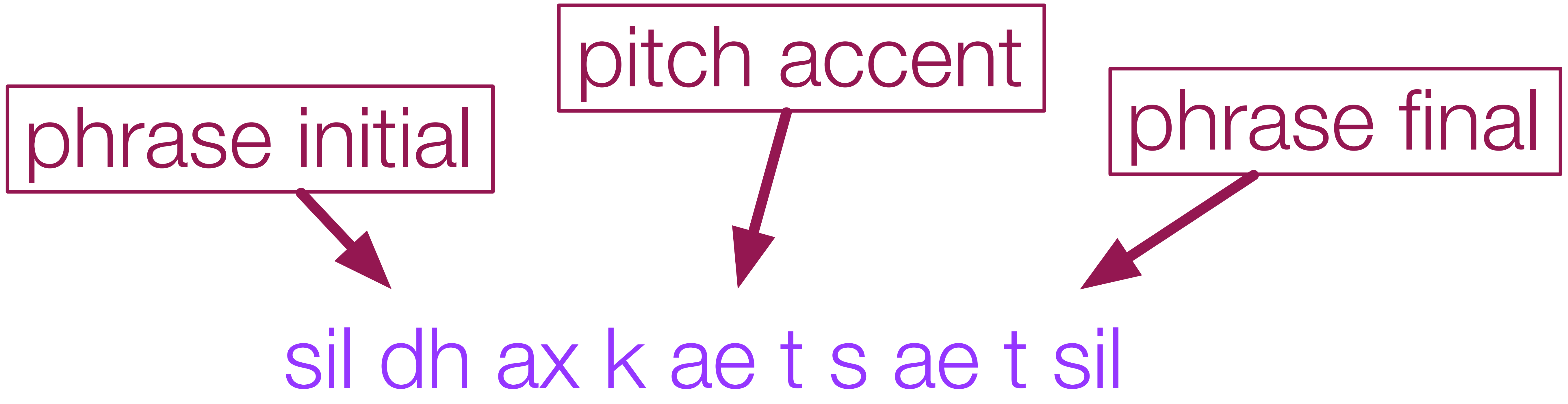
Solve text-to-speech as **sequence-to-sequence** regression using DNNs



output sequence

input sequence

Remind yourself that a decision tree effectively treats the input features as "one hot"



phrase initial

pitch accent

phrase final

sil dh ax k ae t s ae t sil

"the cat sat"

DET NN VB

((the cat) sat)

Exercise: represent this input text as a sequence of one-hot vectors

“Please call . . . .”

Visualising vector representations of linguistic information

- Assume the following
  - the phone set only contains 3 classes (instead of, say, 45 for English)

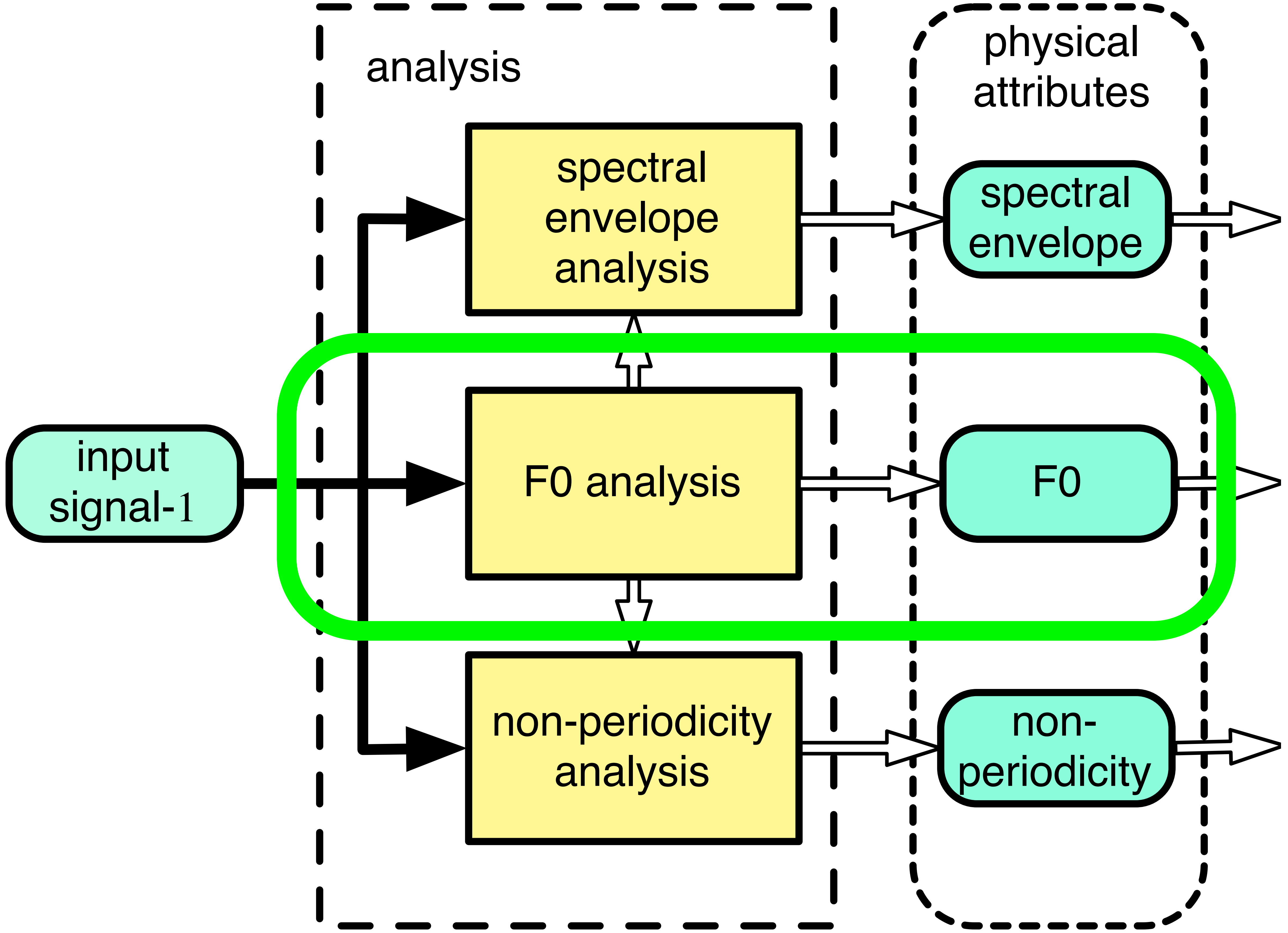
    θ , ð , ʃ

  - we only encode the current phone (ignore all context)

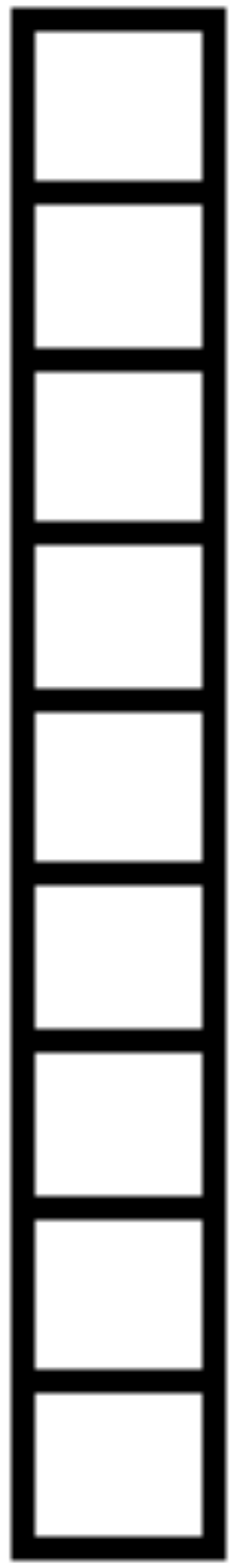
- Make a table of all the necessary one-hot codes (each code is a vector)
- Plot the codes (i.e., vectors) on a diagram
- Write down a sequence of vectors representing

  [θ ð θ ʃ ð]

What are the output features (i.e., speech parameters) ?



speech parameters



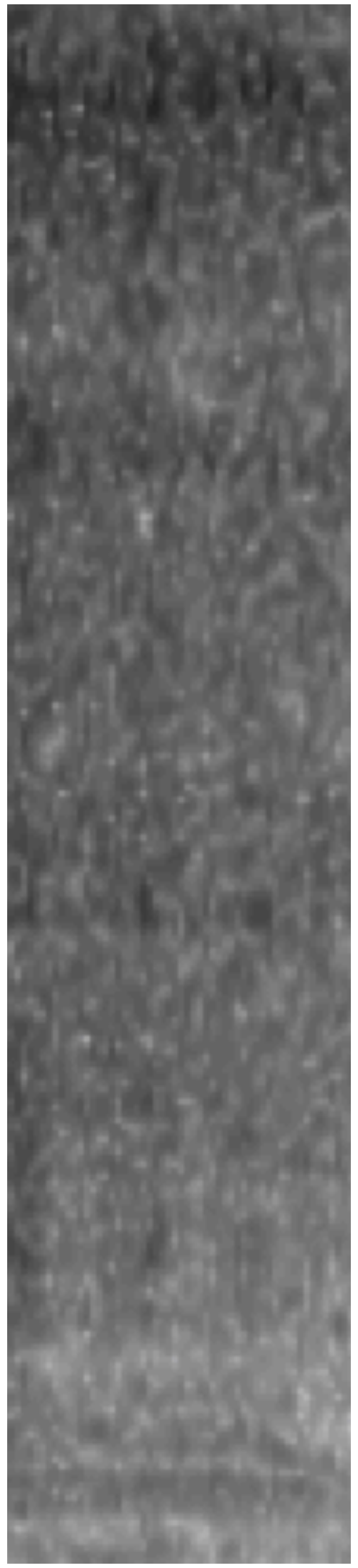
output feature vector

Output **representation**

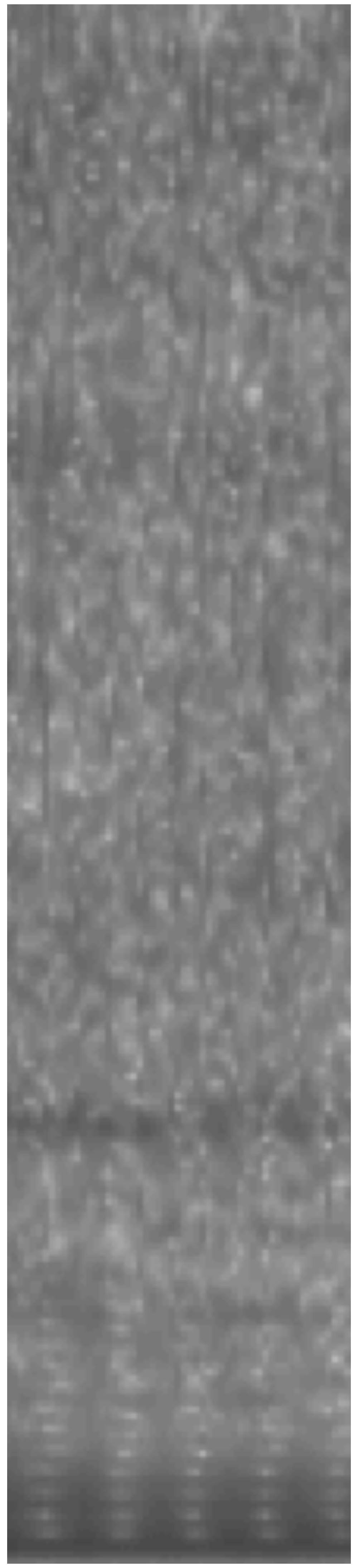
- It is no longer necessary to **decorrelate** the speech parameters
- So, is there any value in converting the spectral envelope to Mel Cepstral Coefficients ?

- Could a single feed-forward DNN predict **all of the following**
  - mel-cepstrum, F0, energy, duration ?

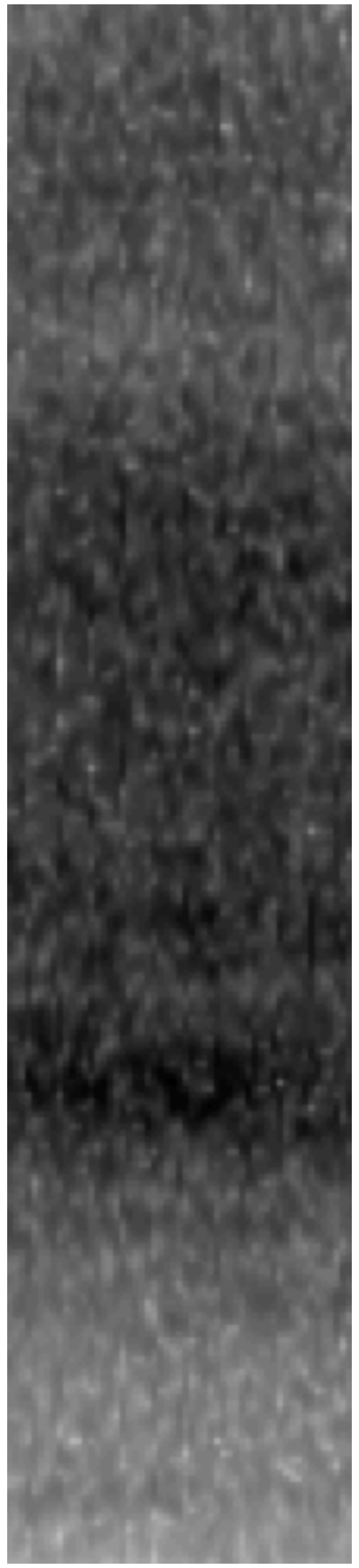
θ



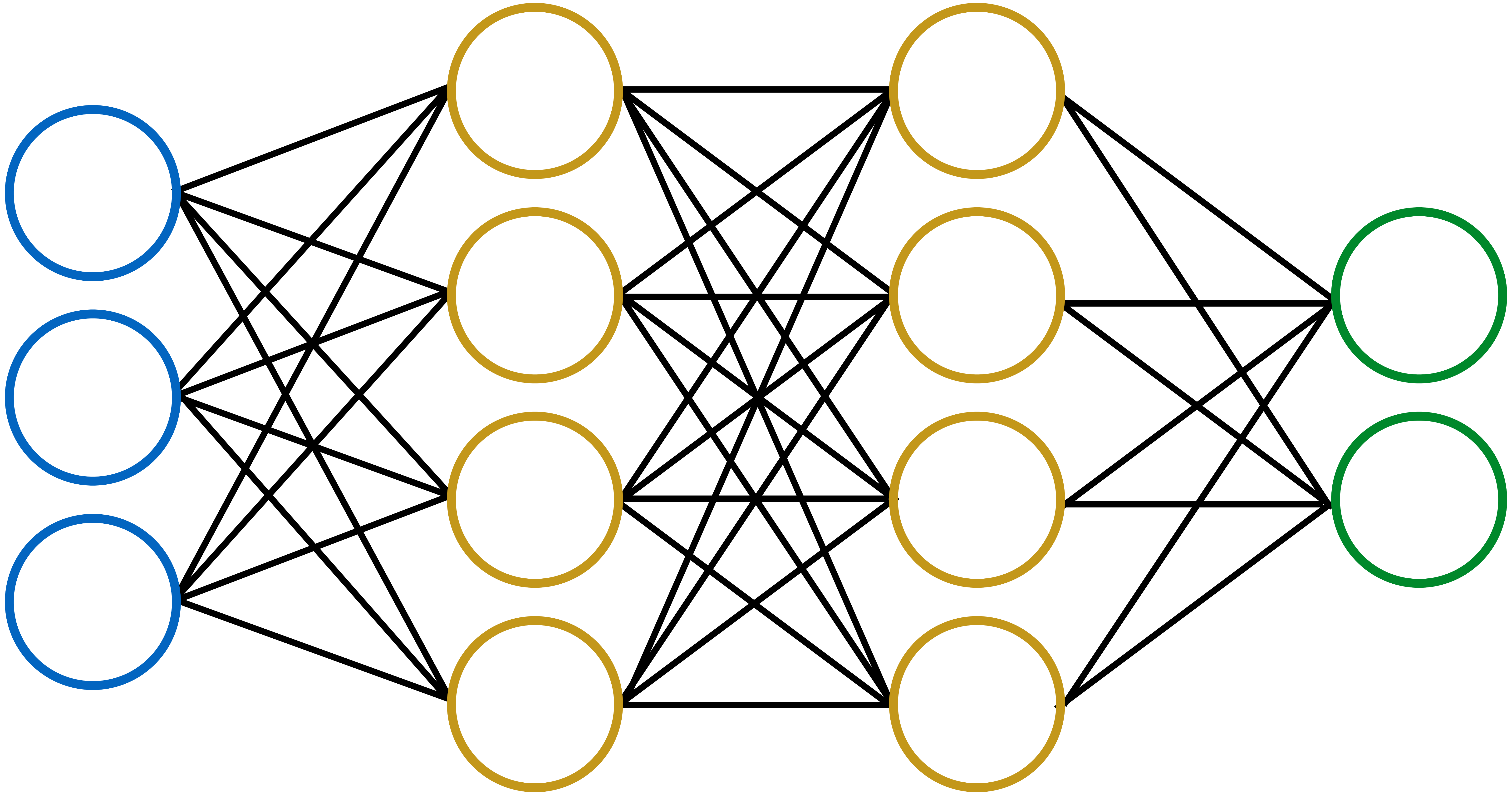
ð



ʃ

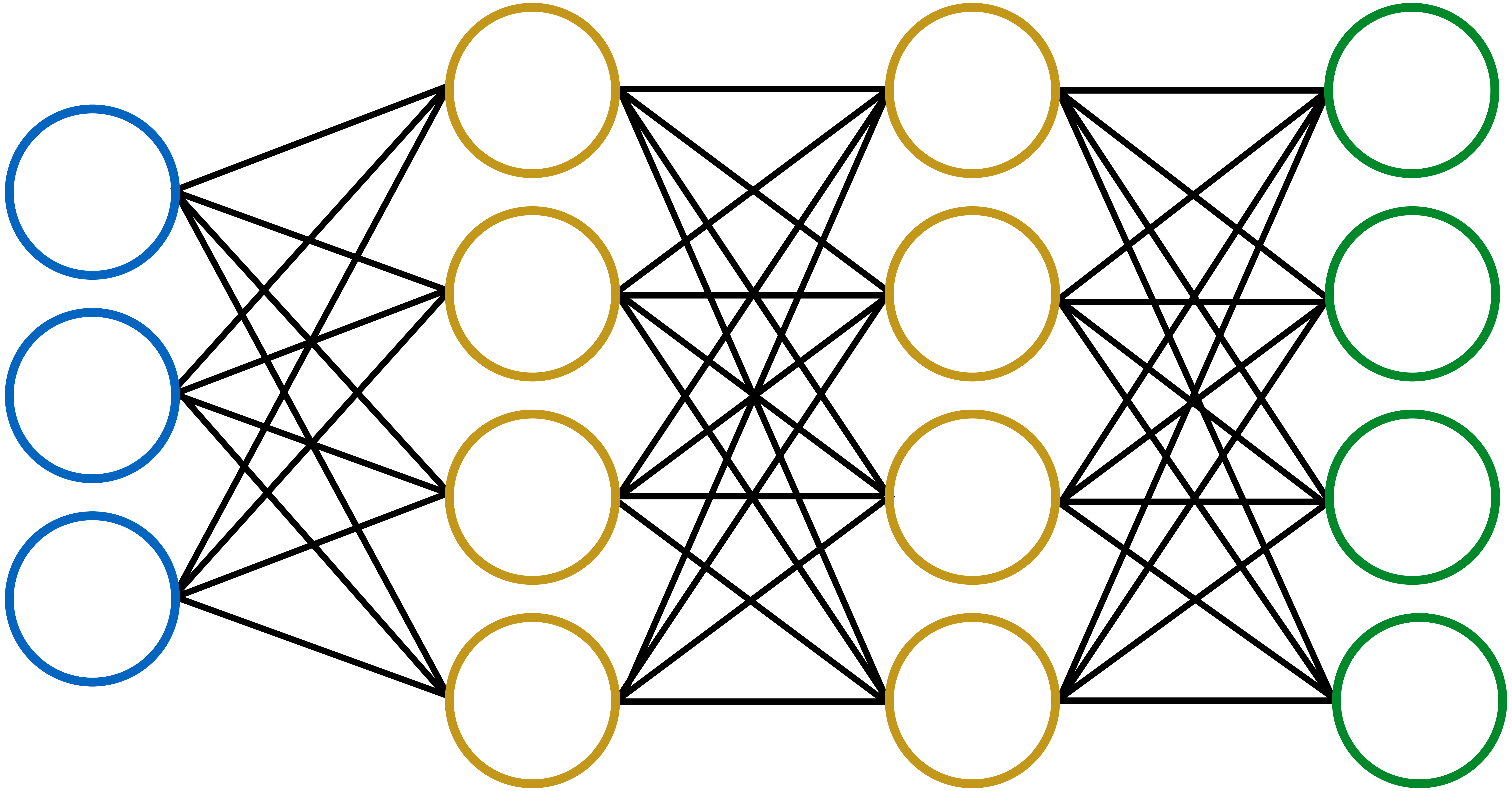


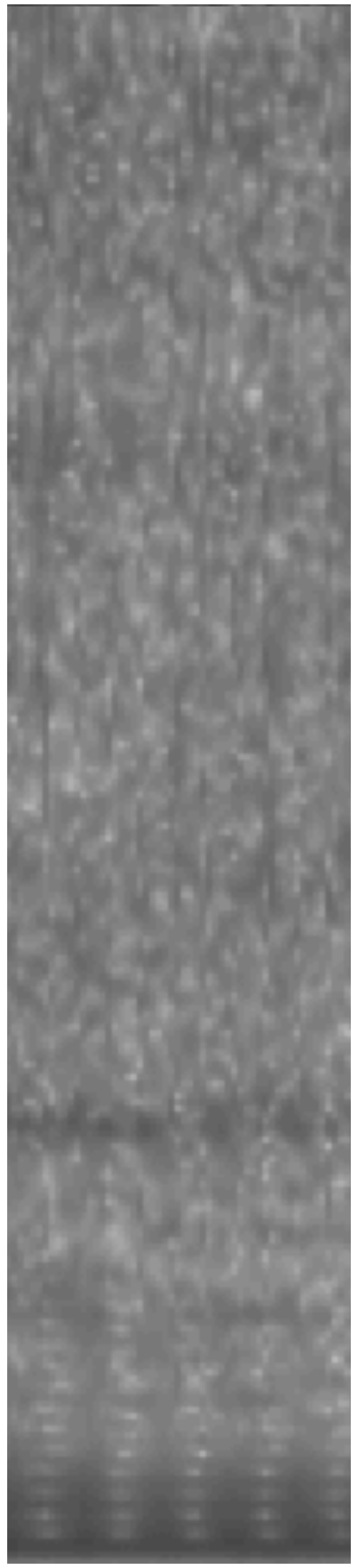
Hidden layer activations = intermediate **representations**



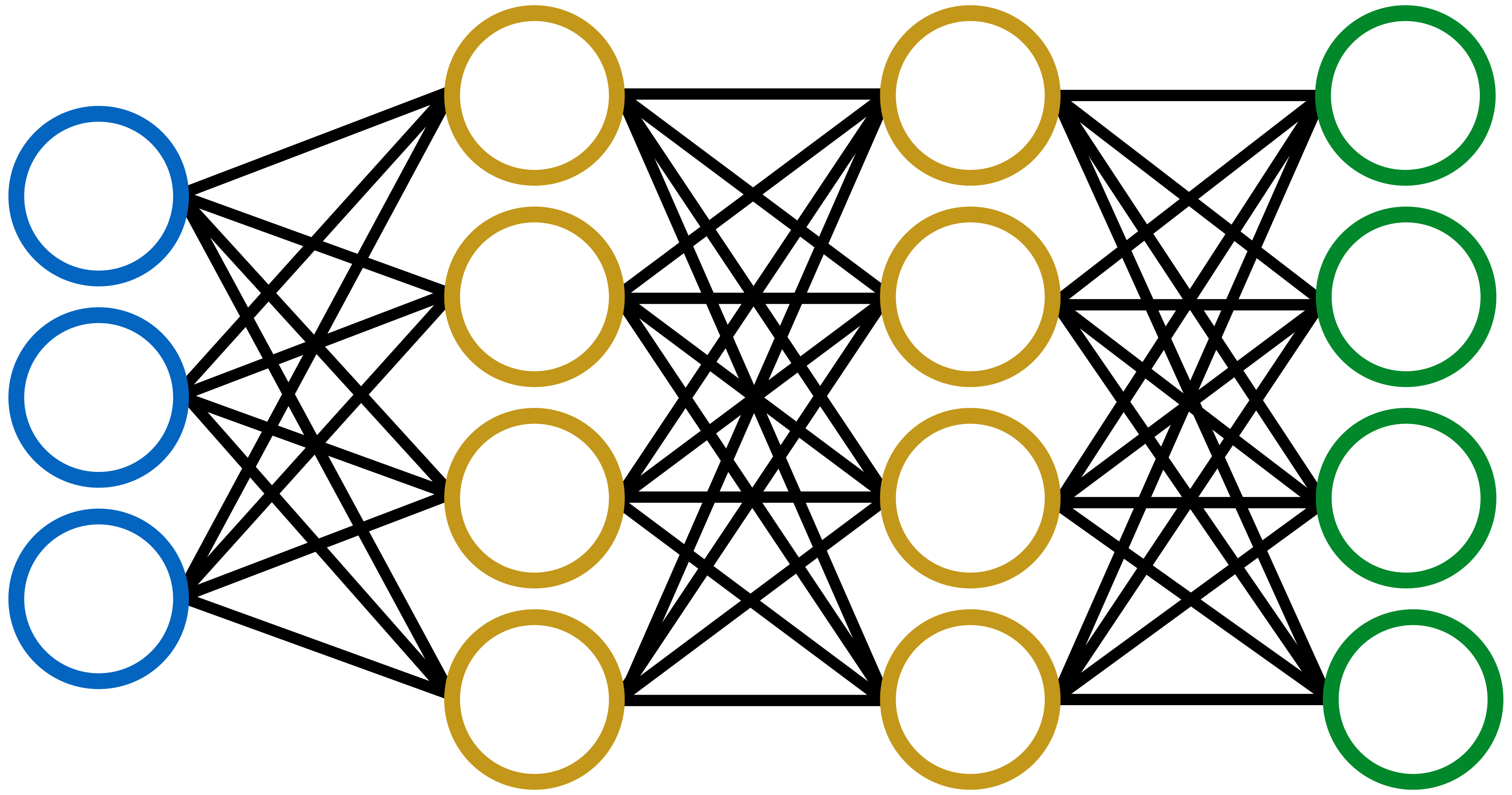
Doing regression

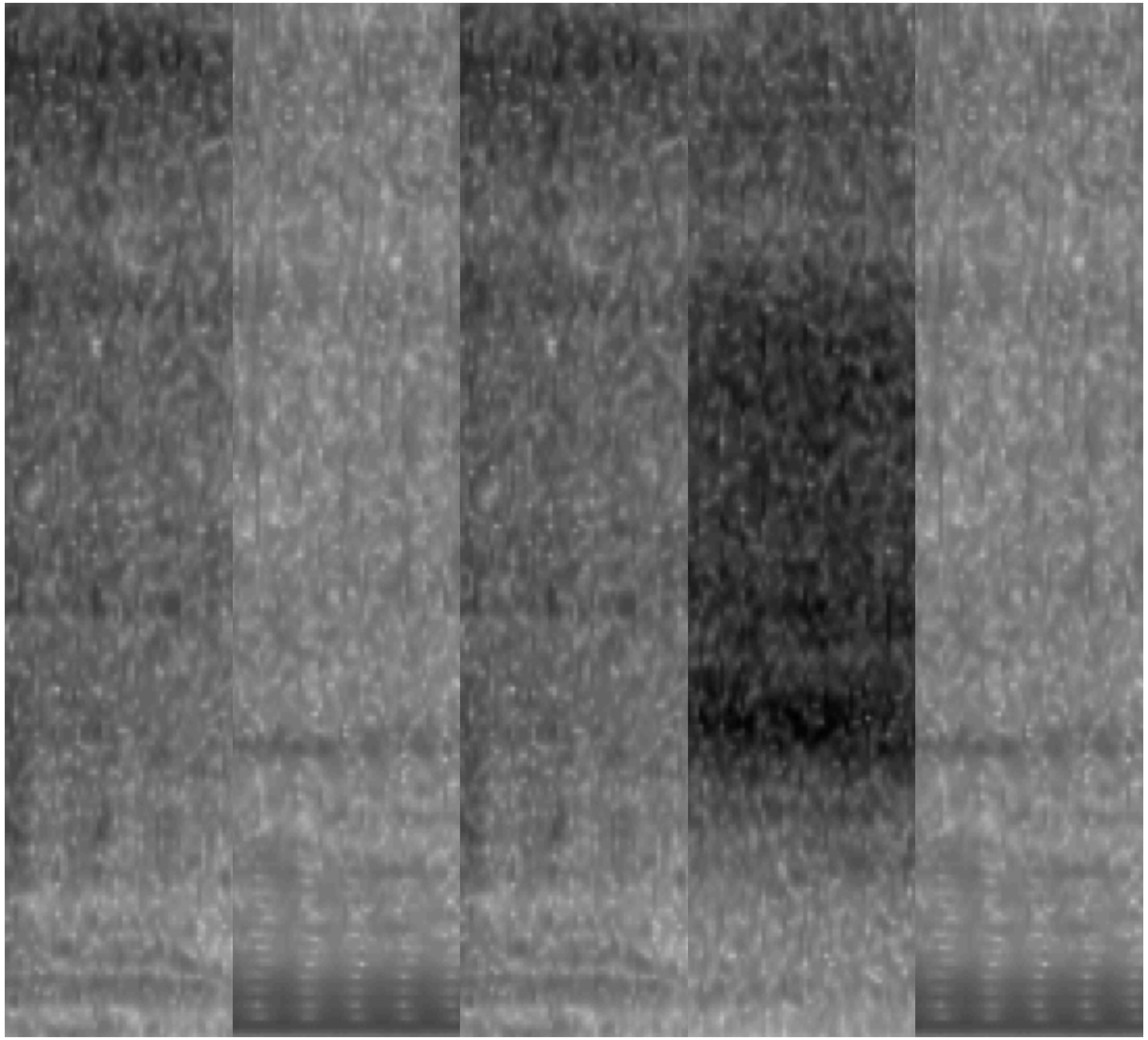
θ





Now do the sequence [θ ð θ ʃ ð]





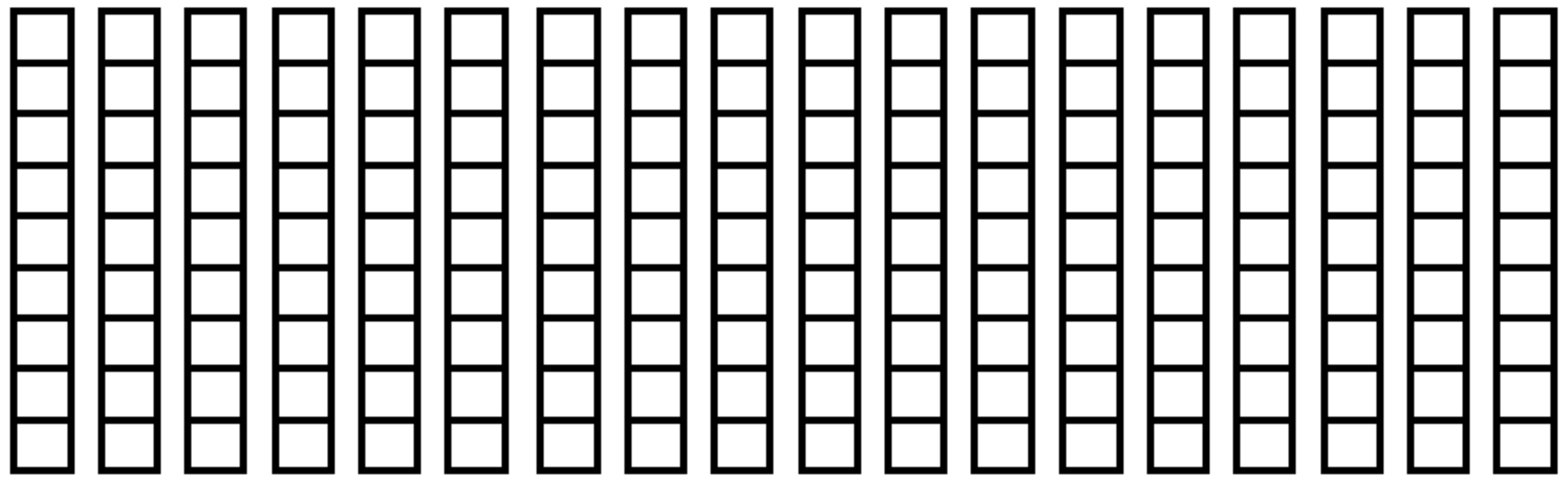
Network architectures

- Different types of unit
    - “standard” memoryless: sigmoid, tanh, or similar
    - Long Short-Term Memory

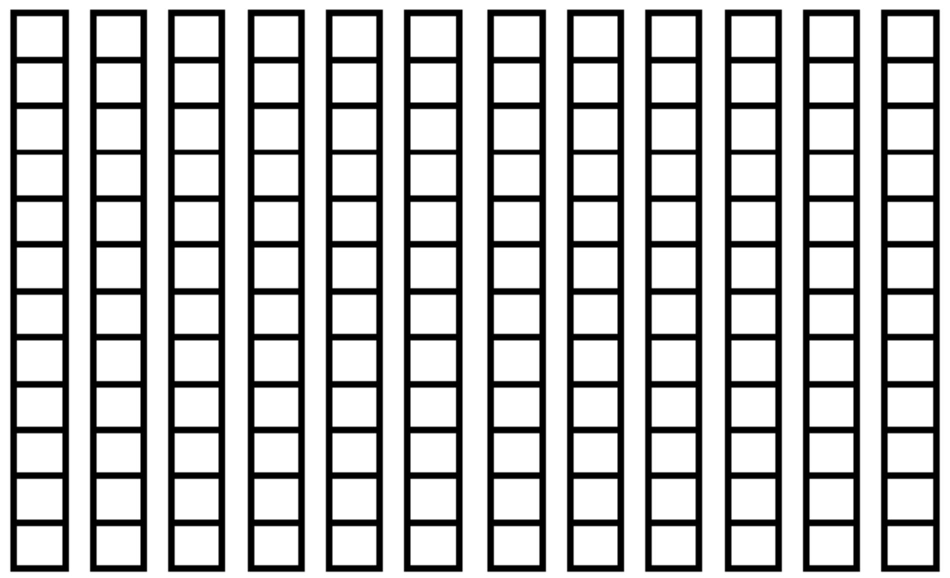
- Different architectures
    - feed-forward
    - recurrent (as used in language modelling for ASR)
    - special patterns of connections between layers

Finally, draw a diagram of **sequence-to-sequence** regression using a DNN

output sequence

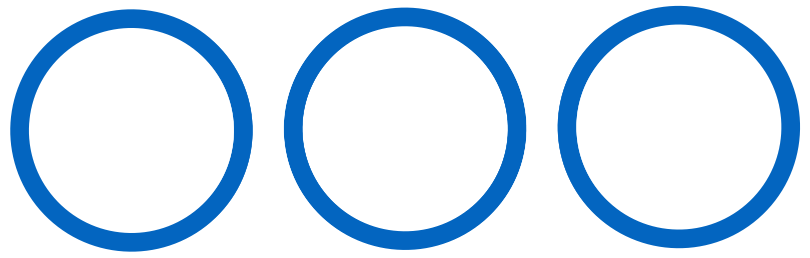


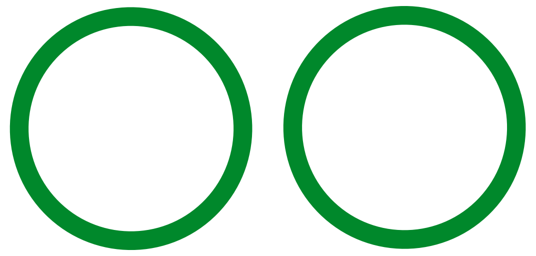
input sequence



Training a neural network: pairs of input/output vectors

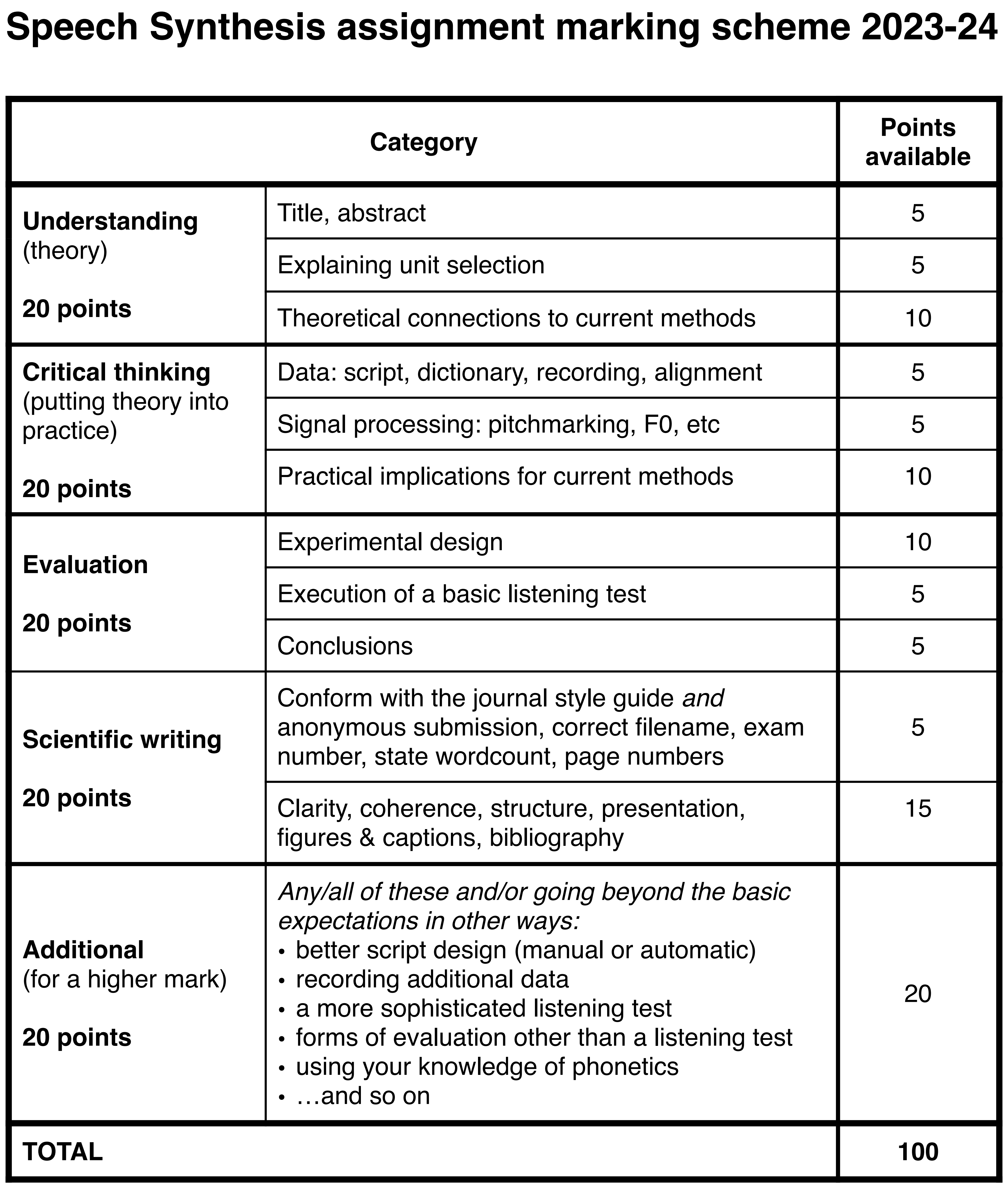
```
[0 0 1 0 0 1 0 1 1 0 … 0.2  0.0]      [0.12 2.33 2.01 0.32 6.33 … ]
[0 0 1 0 0 1 0 1 1 0 … 0.2  0.1]      [0.43 2.11 1.99 0.39 4.83 … ]
…
[0 0 1 0 0 1 0 1 1 0 … 0.2  1.0]      [1.11 2.01 1.87 0.36 2.14 … ]
[0 0 1 0 0 1 0 1 1 0 … 0.4  0.0]      [1.52 1.82 1.89 0.34 1.04 … ]
[0 0 1 0 0 1 0 1 1 0 … 0.4  0.5]      [1.79 1.74 2.21 0.33 0.65 … ]
[0 0 1 0 0 1 0 1 1 0 … 0.4  1.0]      [1.65 1.58 2.68 0.31 0.73 … ]
…
[0 0 1 0 0 1 0 1 1 0 … 1.0  1.0]      [1.55 1.03 3.44 0.30 1.07 … ]
[0 0 0 1 1 1 0 1 0 0 … 0.2  0.0]      [1.92 0.99 3.89 0.29 1.45 … ]
[0 0 0 1 1 1 0 1 0 0 … 0.2  0.2]      [2.38 1.13 4.02 0.28 1.98 … ]
[0 0 0 1 1 1 0 1 0 0 … 0.2  0.4]      [2.65 1.98 3.94 0.29 2.16 … ]
…
```





Optional: bonus slides that derive back-propagation (non-examinable)

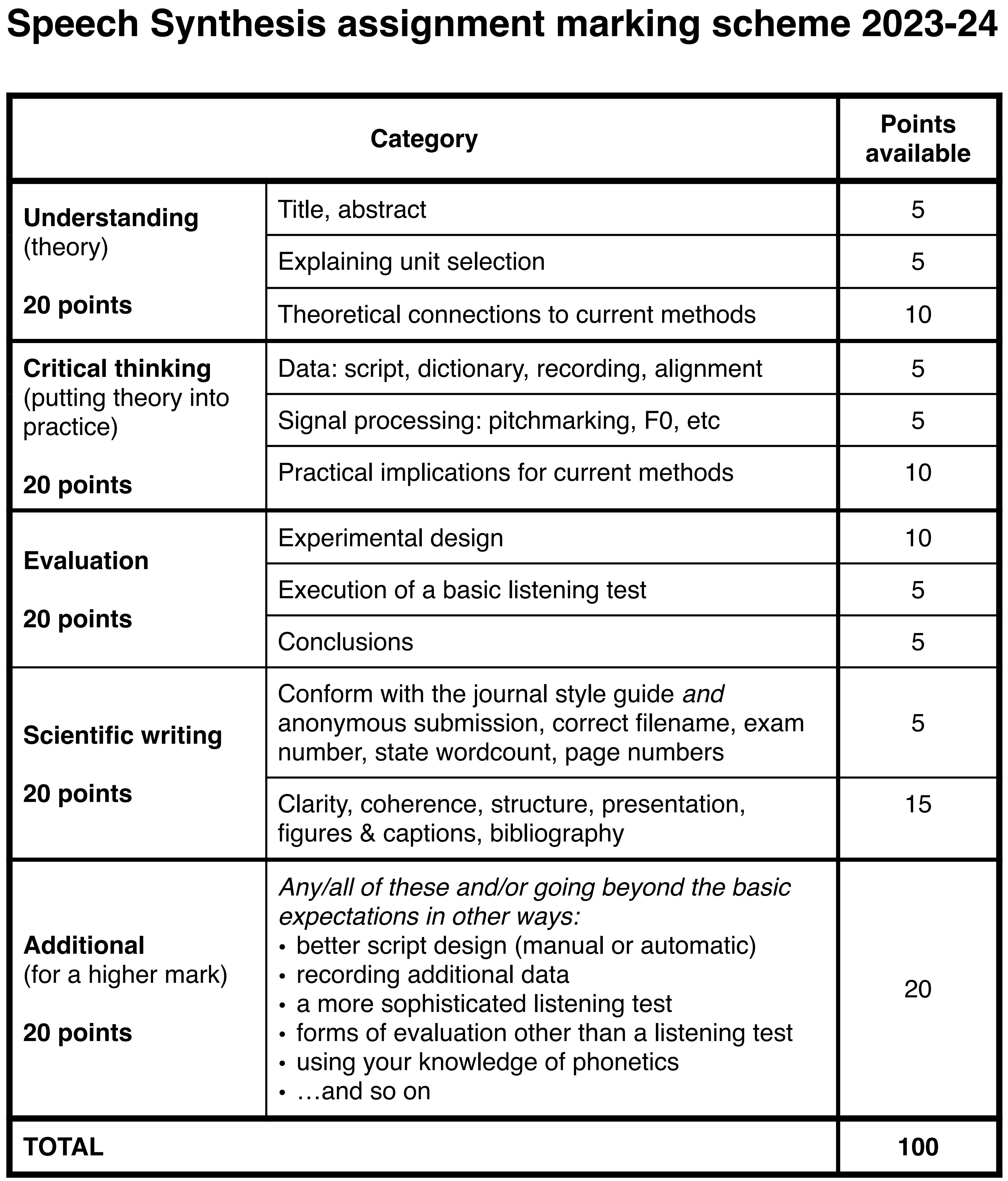
Tips on interpreting the marking sheet



Speech Synthesis assignment marking scheme 2023-24

Category		**Points available**
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

The marking sheet is *not* a table of contents for your paper

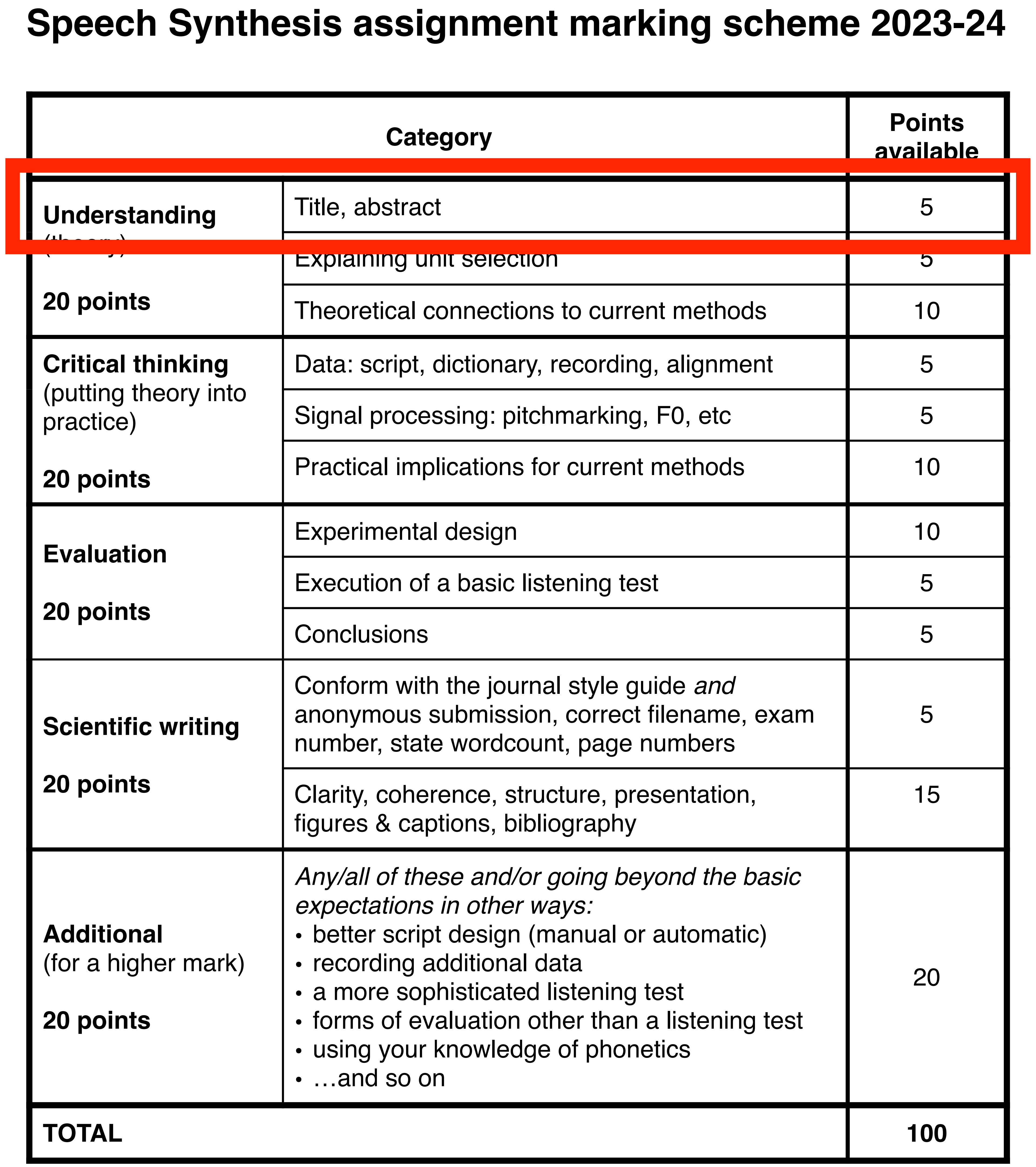


Speech Synthesis assignment marking scheme 2023-24

Category		**Points available**
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

A well-structured, polished report showing good effort, with interesting and justified investigations and claims supported by evidence, will get a good grade.

Tips on interpreting the marking sheet



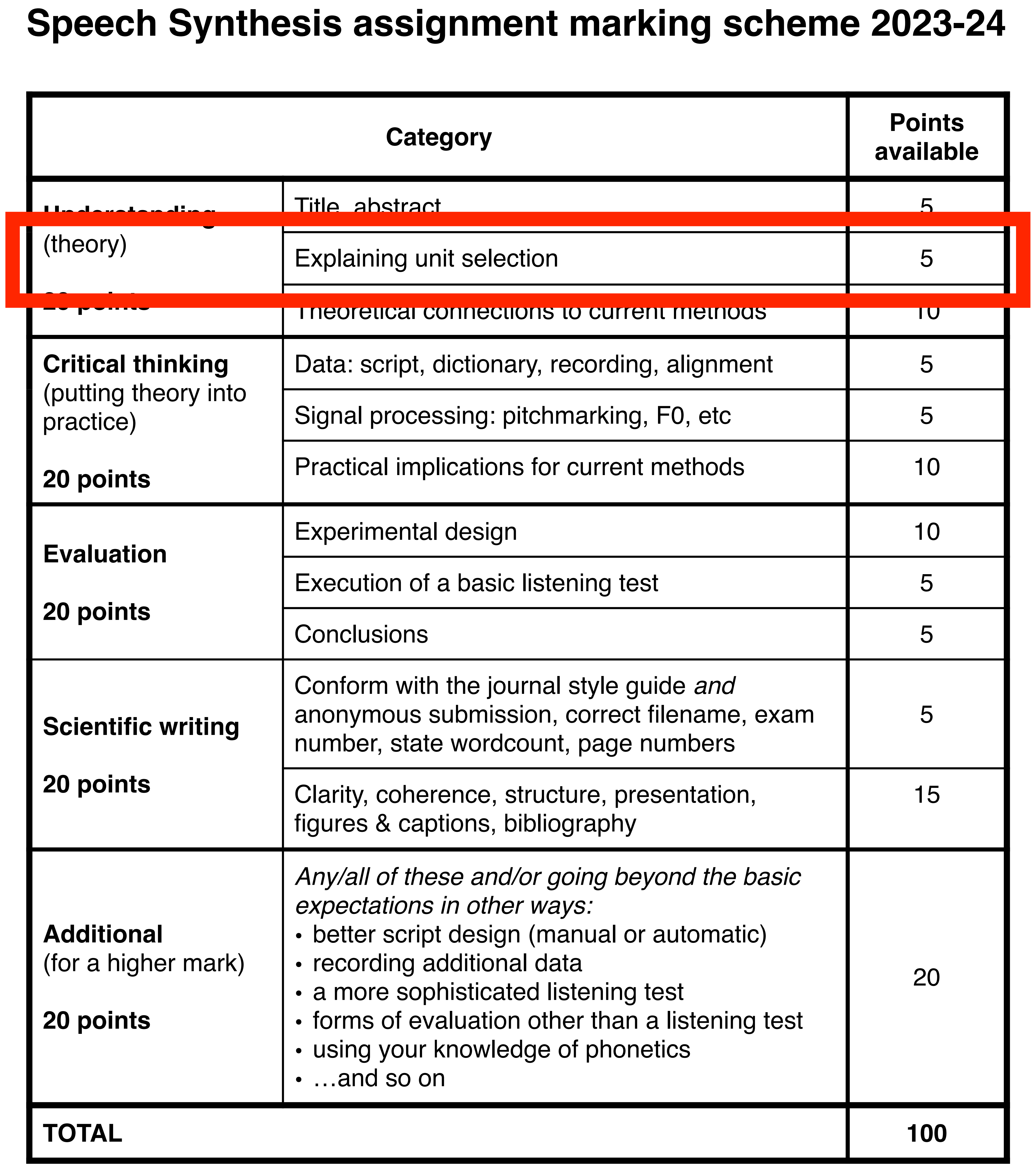
Speech Synthesis assignment marking scheme 2023-24

Category		**Points available**
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- **Informative** title
- **Structured** abstract
- A brief introduction to **this** paper
  - "scene setting"
  - relevant background (within reason)
  - clear motivation for the work
  - (paper outline/what to expect)
  - (not results or conclusions)

https://www.annaclemens.com/blog/how-to-write-the-perfect-abstract

Tips on interpreting the marking sheet

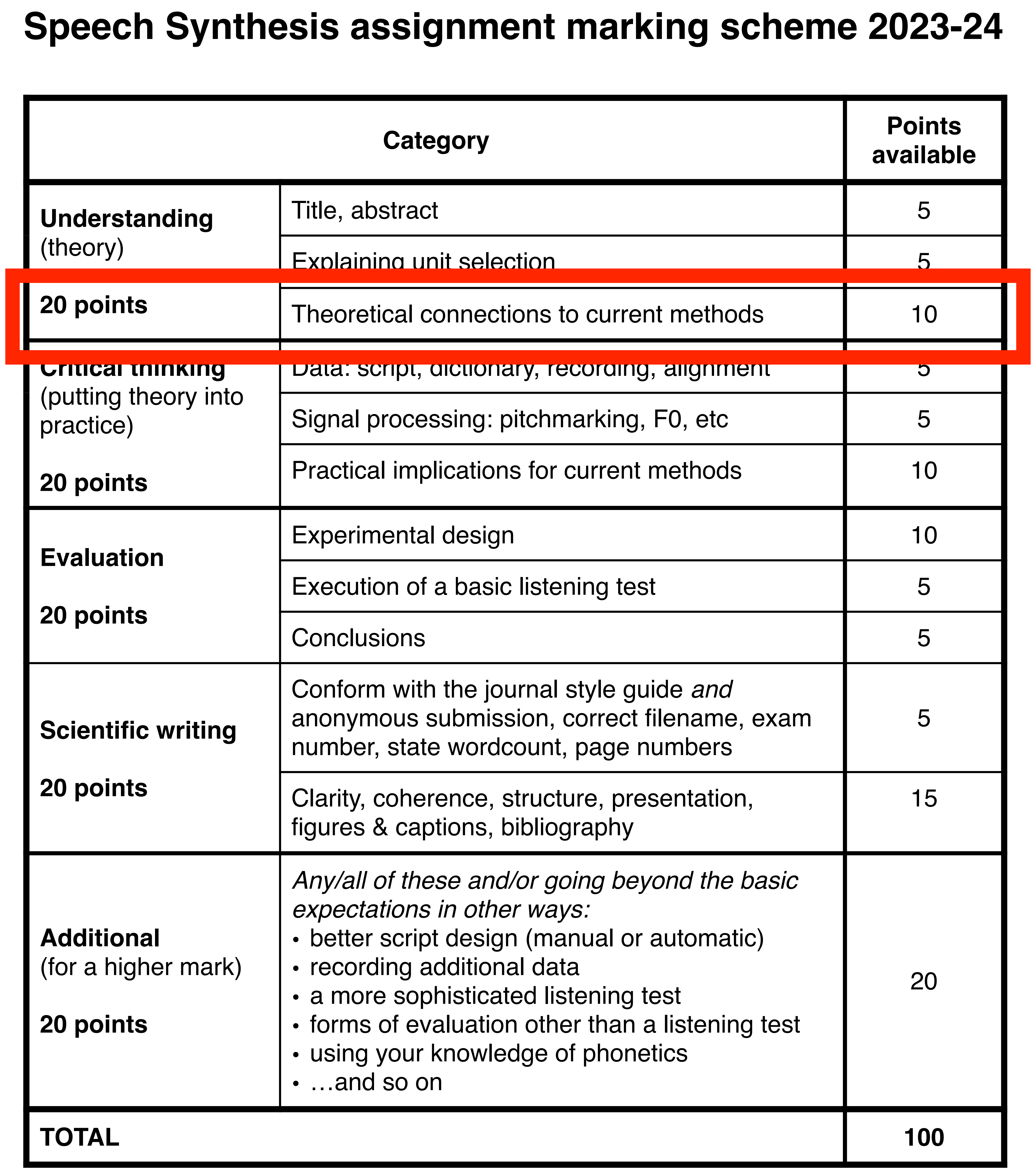


Speech Synthesis assignment marking scheme 2023-24

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- **Brief** explanation only
- Keep it relevant to **this** paper
- Demonstrate your **understanding**

Tips on interpreting the marking sheet

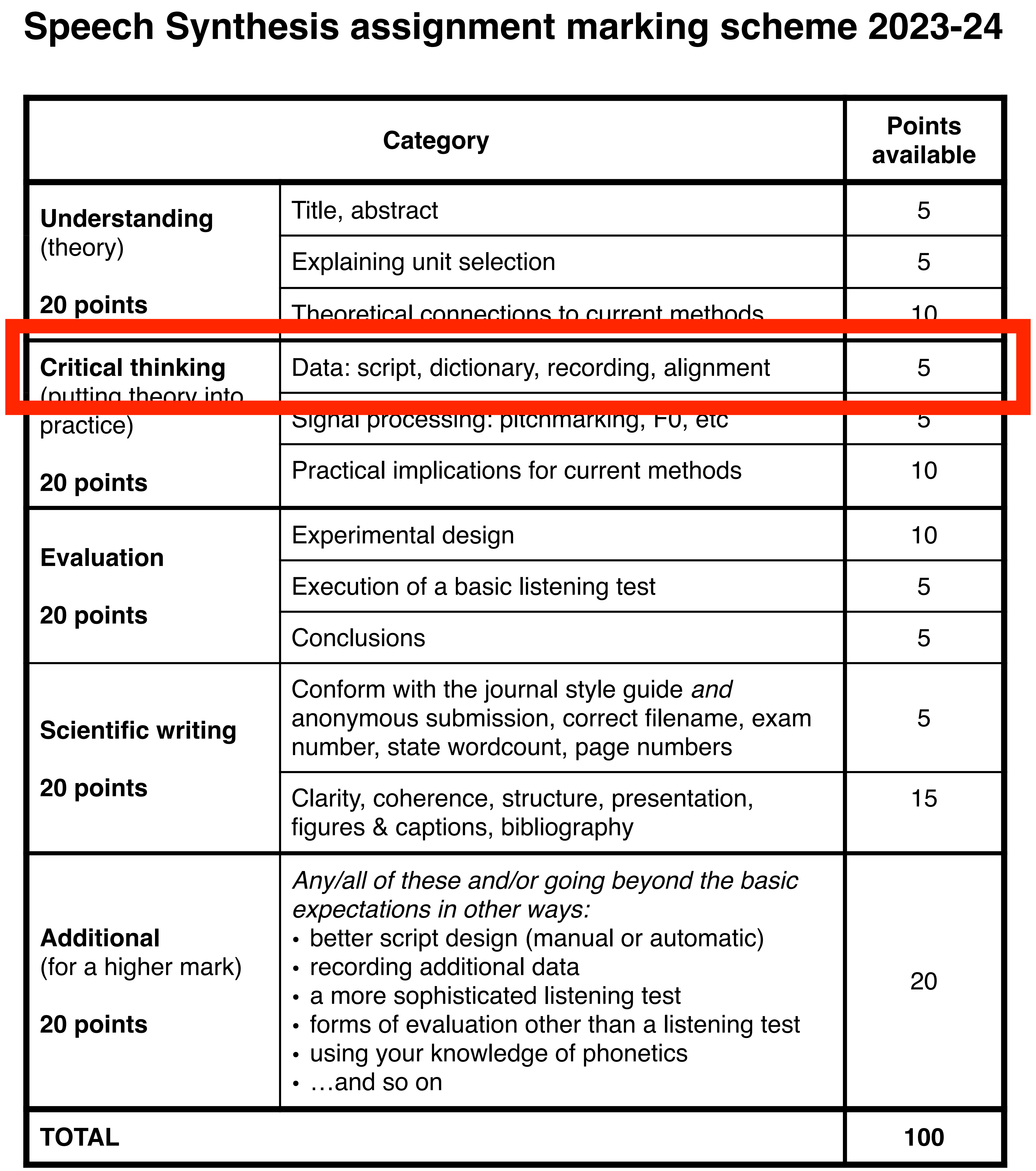


Speech Synthesis assignment marking scheme 2023-24

Category		**Points available**
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- **Incorporate** these throughout the paper
- Example 1:
  - Unit selection performs implicit **regression** from linguistic features to acoustic properties
  - How do various current methods do that?
- Example 2:
  - In unit selection, several choices are available for **waveform representation**
  - Are these the same or different in current methods?
- etc.

Tips on interpreting the marking sheet

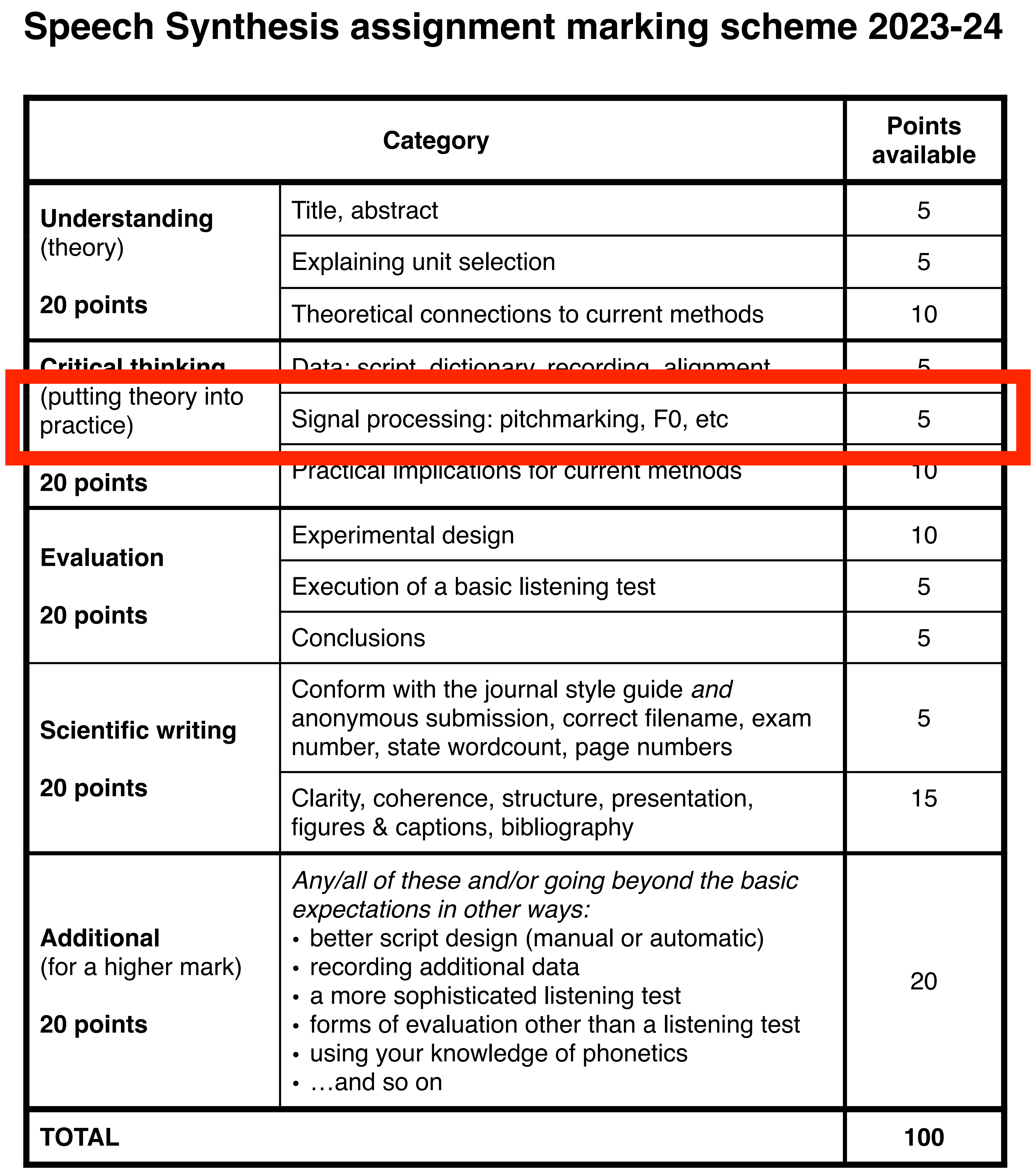


Speech Synthesis assignment marking scheme 2023-24

Category		**Points available**
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- Self-explanatory
- Look at the mark available and keep the basics really tight and to the point
- Optional extra work, experiments, etc, will attract marks in other categories

Tips on interpreting the marking sheet

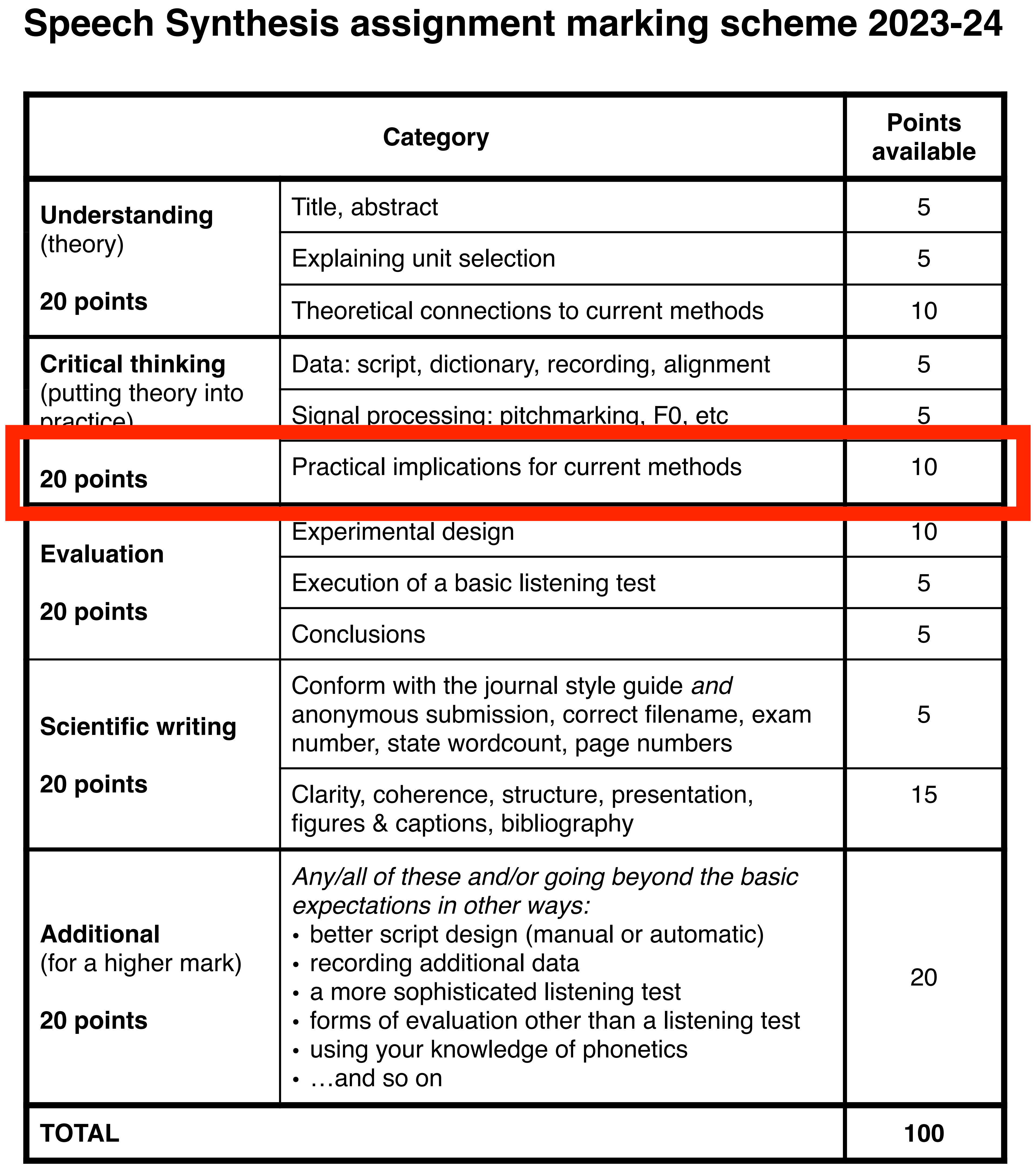


Speech Synthesis assignment marking scheme 2023-24

Category		**Points available**
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- Often overlooked, but easy marks available!
- Just show that you understand the various forms of **signal processing** that are happening
  - in voice building
  - during synthesis

Tips on interpreting the marking sheet

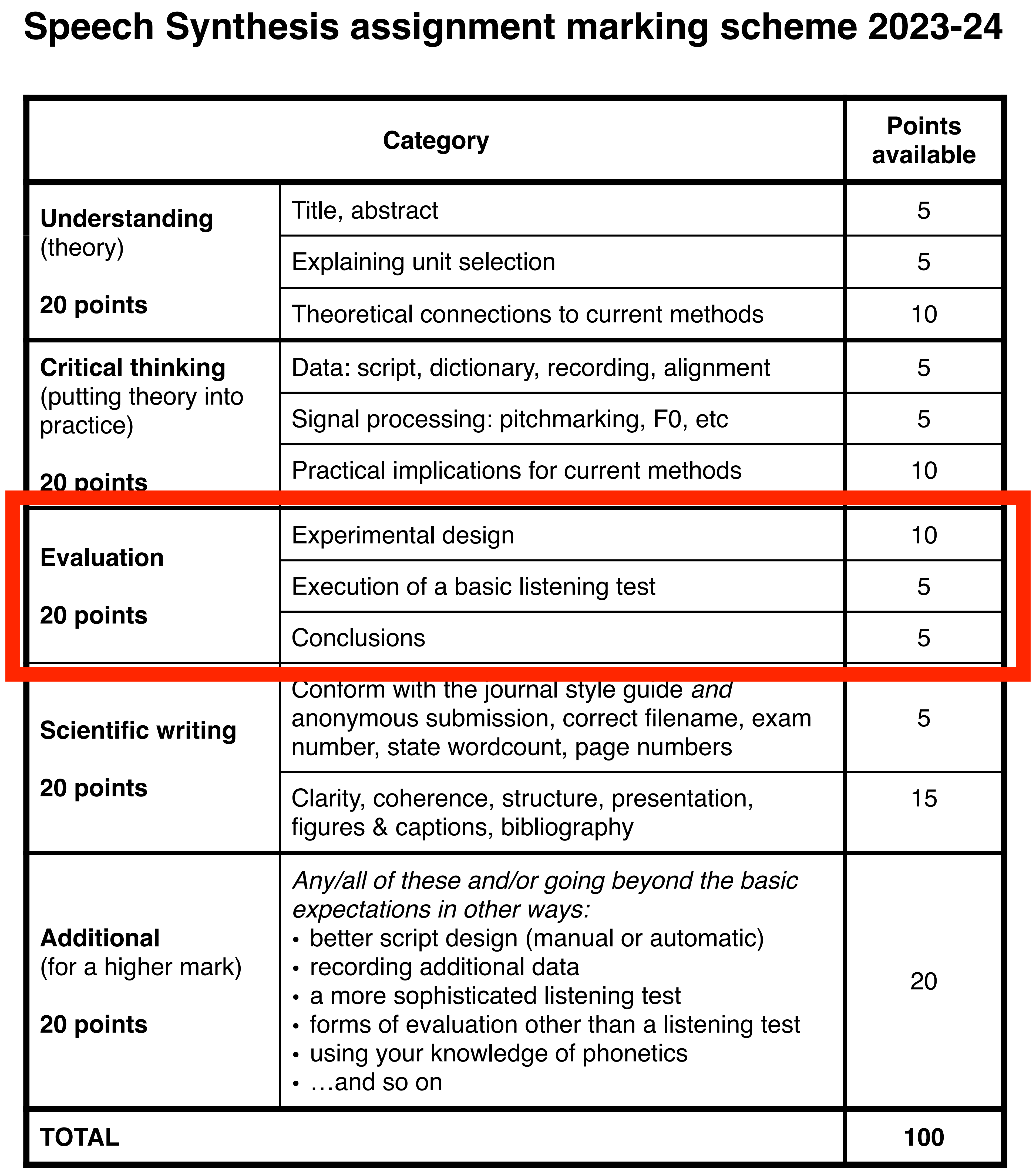


Speech Synthesis assignment marking scheme 2023-24

Category		**Points available**
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- Link everything to **current methods**. Do not do experiments with current methods, but use the literature to back up your claims.
- Example:
  - You will have discovered how sensitive (or not) unit selection is to many **design choices**, such as database contents, pitchmark accuracy, ...
  - Would current methods be more or less sensitive to each choice?

Tips on interpreting the marking sheet

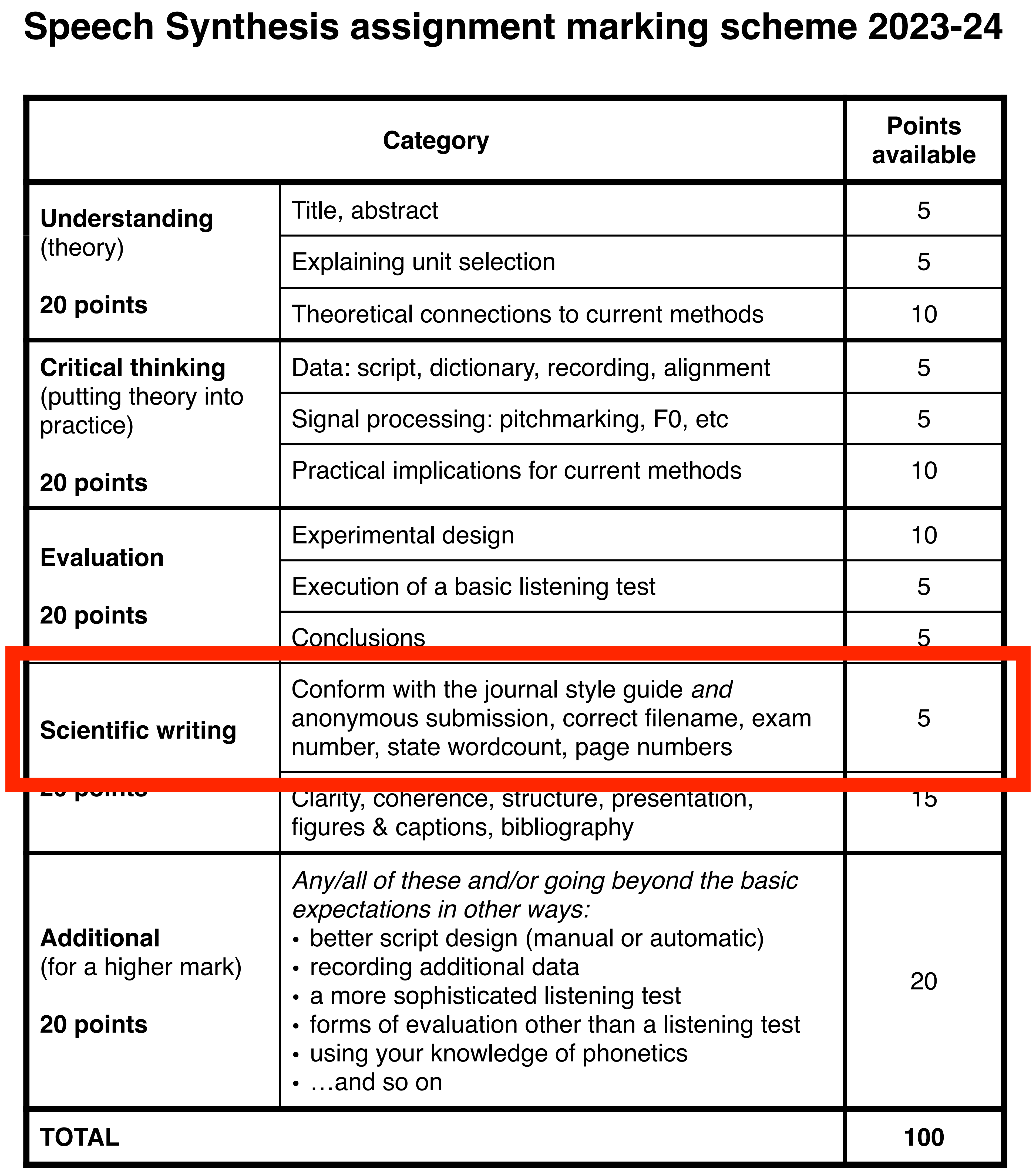


Speech Synthesis assignment marking scheme 2023-24

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- This is where you get marks for your experimental work and basic listening test
- Further marks available under Additional for going further

Tips on interpreting the marking sheet



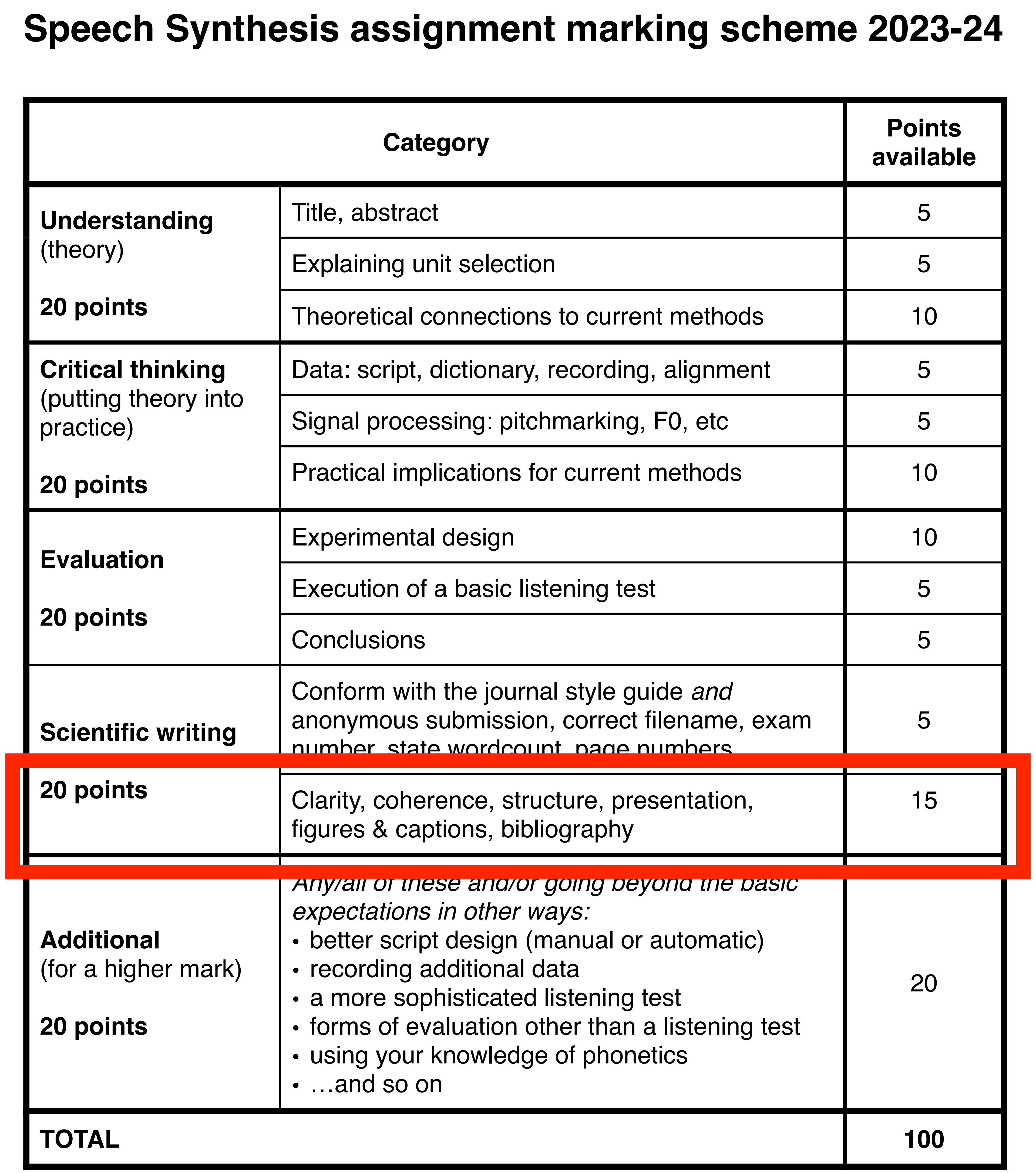
Speech Synthesis assignment marking scheme 2023-24

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- The easiest 5 marks you'll ever get!
- Don't miss out!

- Note: badly formatted work, missing exam number, lack of page numbers, etc - all create extra work for the marker and course organiser.

Tips on interpreting the marking sheet



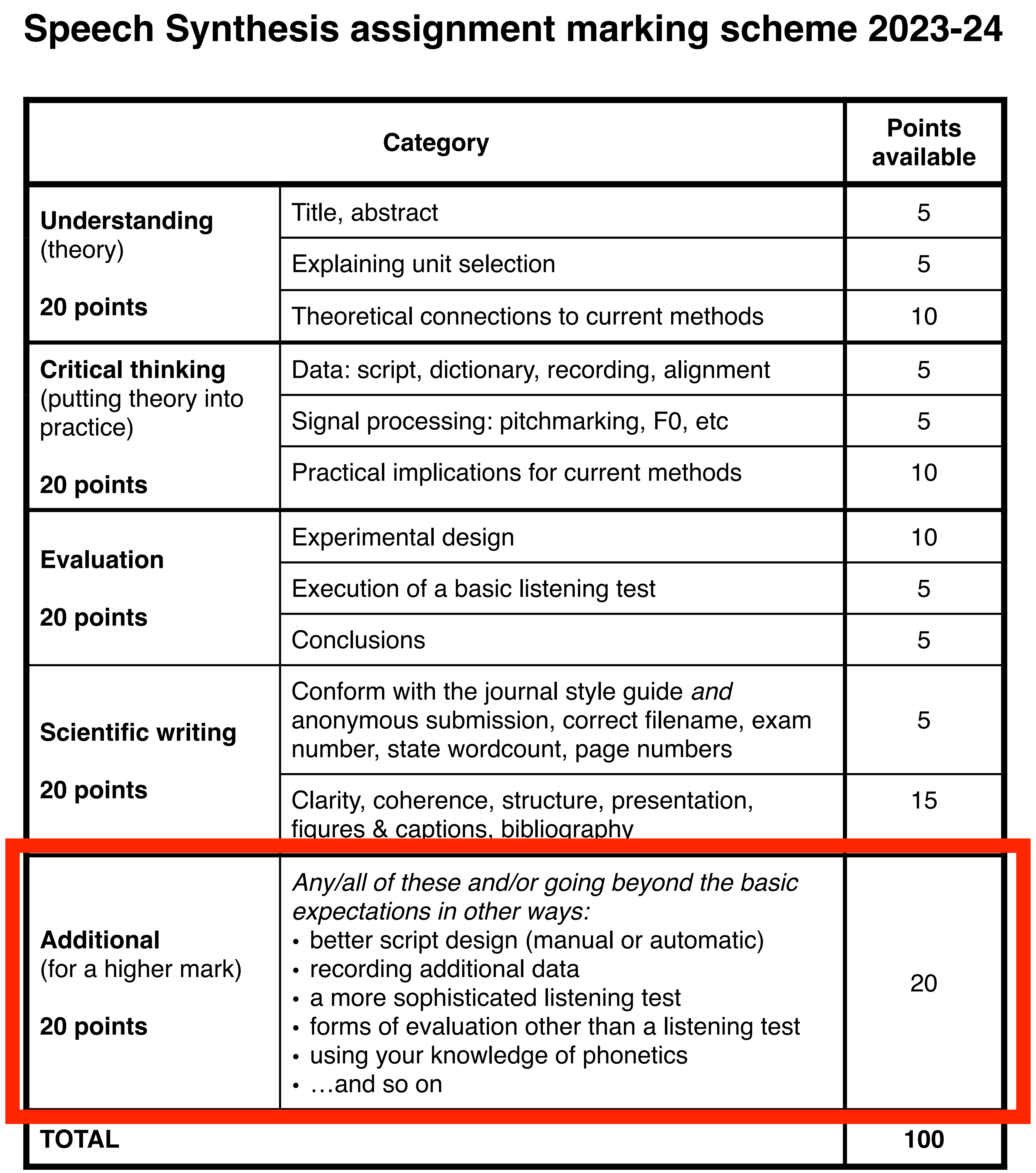
Speech Synthesis assignment marking scheme 2023-24

Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- Use the **feedback** from Speech Processing (*)
- **Scientific writing** should be clear, simple, and unambiguous
- Plan your paper's **structure** carefully
- Have your **reader** in mind at all times
- Good **presentation** makes a paper more enjoyable to read
- A happy marker is a generous marker

(*) If you didn't take Speech Processing, contact Simon for additional 1-on-1 help with your writing

Tips on interpreting the marking sheet



Speech Synthesis assignment marking scheme 2023-24

Category		**Points available**
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

- You are *not* expected to do **all** of these!
- But be tactical:
  - do aim for some marks in **multiple categories**
  - do not try to get all 20 points for going too deep in only one category (e.g., script design)
  - the list on the marking sheet is not exhaustive: creativity will be rewarded

Final tips

- Focus on **demonstrating your understanding**, not on how Festival and the scripts work
- **Figures** can say a lot with only a few words
- Present your experimental results in an **attractive** way
- A **bibliography** and in-text **citations** *must* be provided
    - Go beyond the Essential readings if you are aiming for a high mark
    - Cite **peer-reviewed** work whenever possible
    - Never cite a **preprint** (e.g., arXiv) when a peer-reviewed version is available
- The actual quality of your synthetic voice will **not** influence your mark