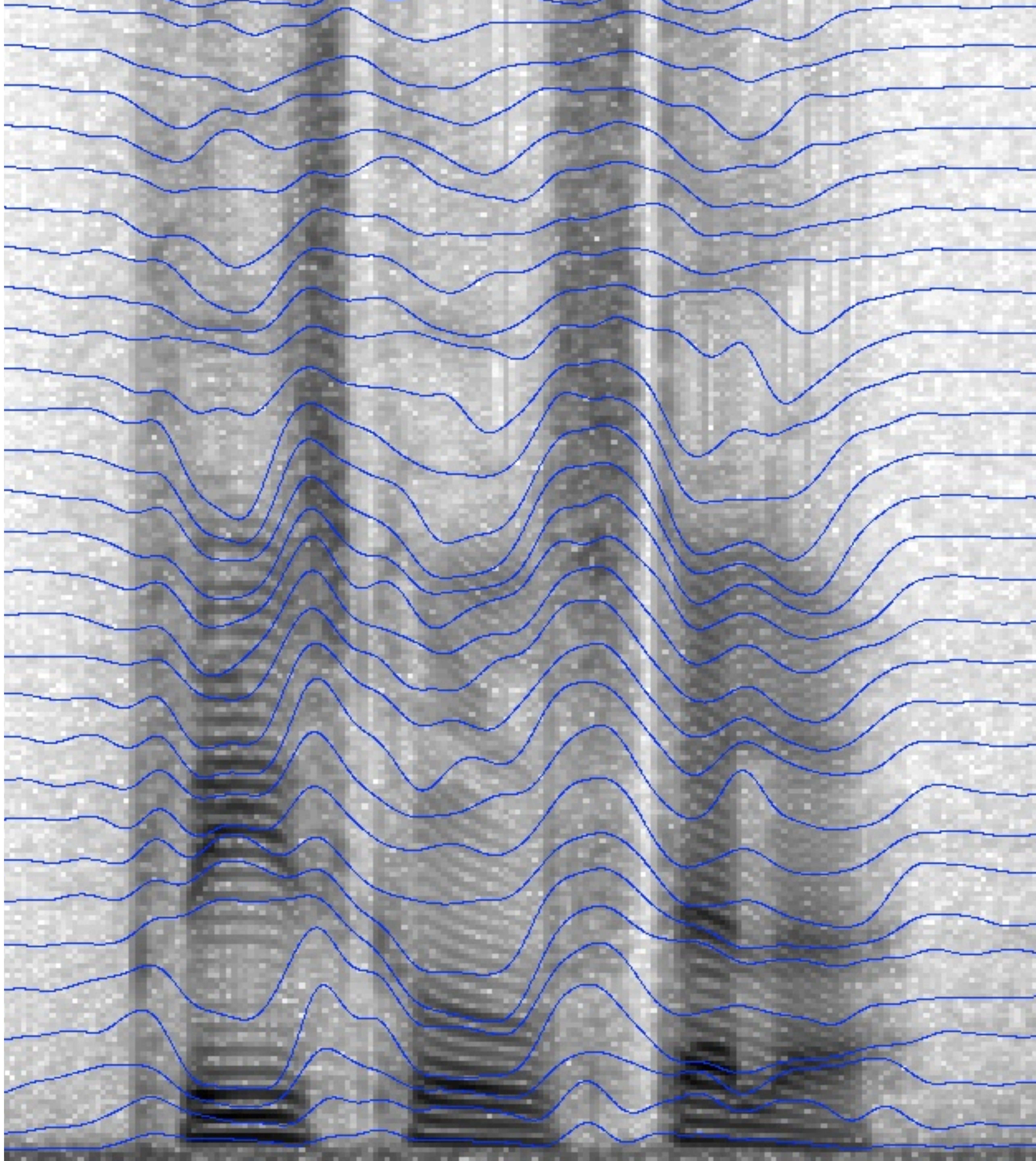
Speech Synthesis

Simon King
University of Edinburgh



Introduction to the course (2023-24 version)

- learning outcomes
- delivery
- timetable
- course outline
- introduction to the coursework

Learning outcomes

- Understand the **speech synthesis process**, and be familiar with the processing steps required to convert text to speech.
- Be familiar with the **different speech synthesis methods** currently used by speech synthesis systems and understand the advantages and disadvantages of each.
- Have a detailed understanding of the principles of **unit selection** speech synthesis, and the issues involved with choosing suitable candidate units to match a given target sequence.

Learning outcomes (continued)

- Understand the design issues associated with **recording data** suitable for building a unit selection voice.
- **Practical experience** of building a synthetic voice yourself.
- Be familiar with the different **speech coding** techniques that can be used for speech synthesis, and understand how these can be used to aid the joining of individual speech segments and how using different signal processing techniques to manipulate speech synthesis output affects the speech quality.
- Be in a position to discuss **current issues** in speech synthesis and see where speech synthesis research is heading in the future.

Delivery

- The website speech.zone contains almost everything you will need
  - **video material, slides for the videos, reading lists, forums, calendar, coursework instructions,, slides for classes**
    - you must have an **account** on this site, so that you can post on the forums - make sure you can log in, and email `Simon.King@ed.ac.uk` if you have any trouble

- You still need to use **Learn** for submitting your coursework
- We will also use **Learn** to send class announcements

Delivery

- Please give me **feedback** (email, forum posts, verbally, class reps, PPLS teaching offices, notes slipped under office doors,…) about **course structure** and **delivery mode**, **throughout** the course.

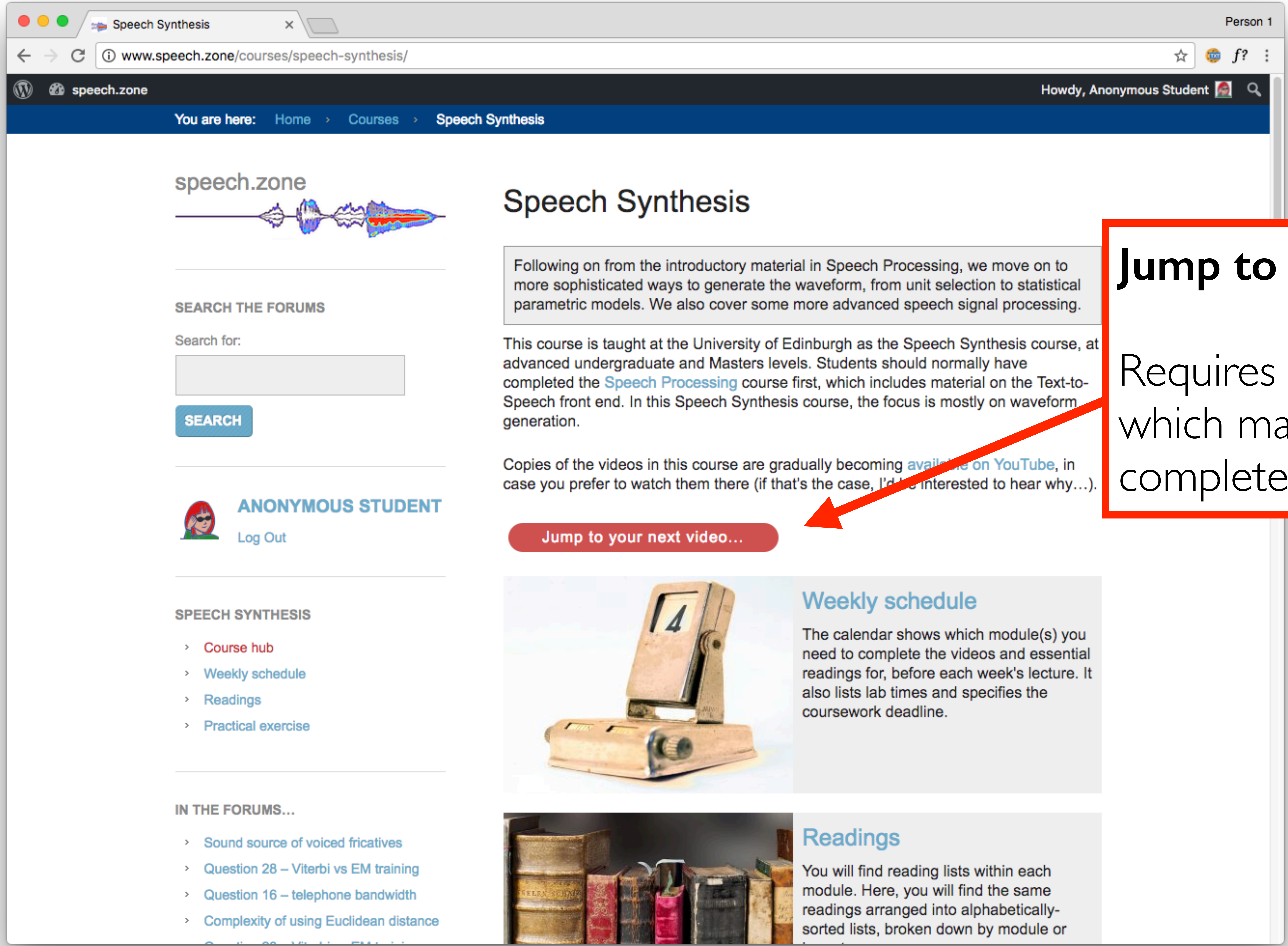
- I also want feedback on speech.zone
  - is it clearly organised?
  - is the website reliable and fast enough?
  - is it obvious what relates to this course, and what does not?
  - does everything work correctly on your device?

Delivery

- **Lectures** will cover the most popular current speech synthesis methods
  - unit selection
  - statistical parametric speech synthesis (SPSS) using HMMs or Neural Networks
  - the current state of the art: sequence-to-sequence models
- **Coursework** - a single major assignment
  - build and evaluate a unit selection speech synthesiser, using your own recordings
- **Readings** - lists provided on speech.zone
- *Background assumed*
  - most of you will have taken Speech Processing - if you have not taken this course, then please speak to the lecturer as soon as possible (if I give you permission to enrol, you'll need to catch up on Speech Processing content, including some videos and readings)

Delivery

- The material on speech.zone is divided into **modules**
  - the video content provides only the *bare bones*
    - this is especially true of the more advanced material towards the end of the course
  - **you** need to flesh out the details by taking **full advantage** of
    - readings
    - active participation in classes
    - labs (including discussion with other students, the tutors, and the lecturer)
    - forums (please attempt to answer each other's posts - I will correct any errors and provide definitive answers)



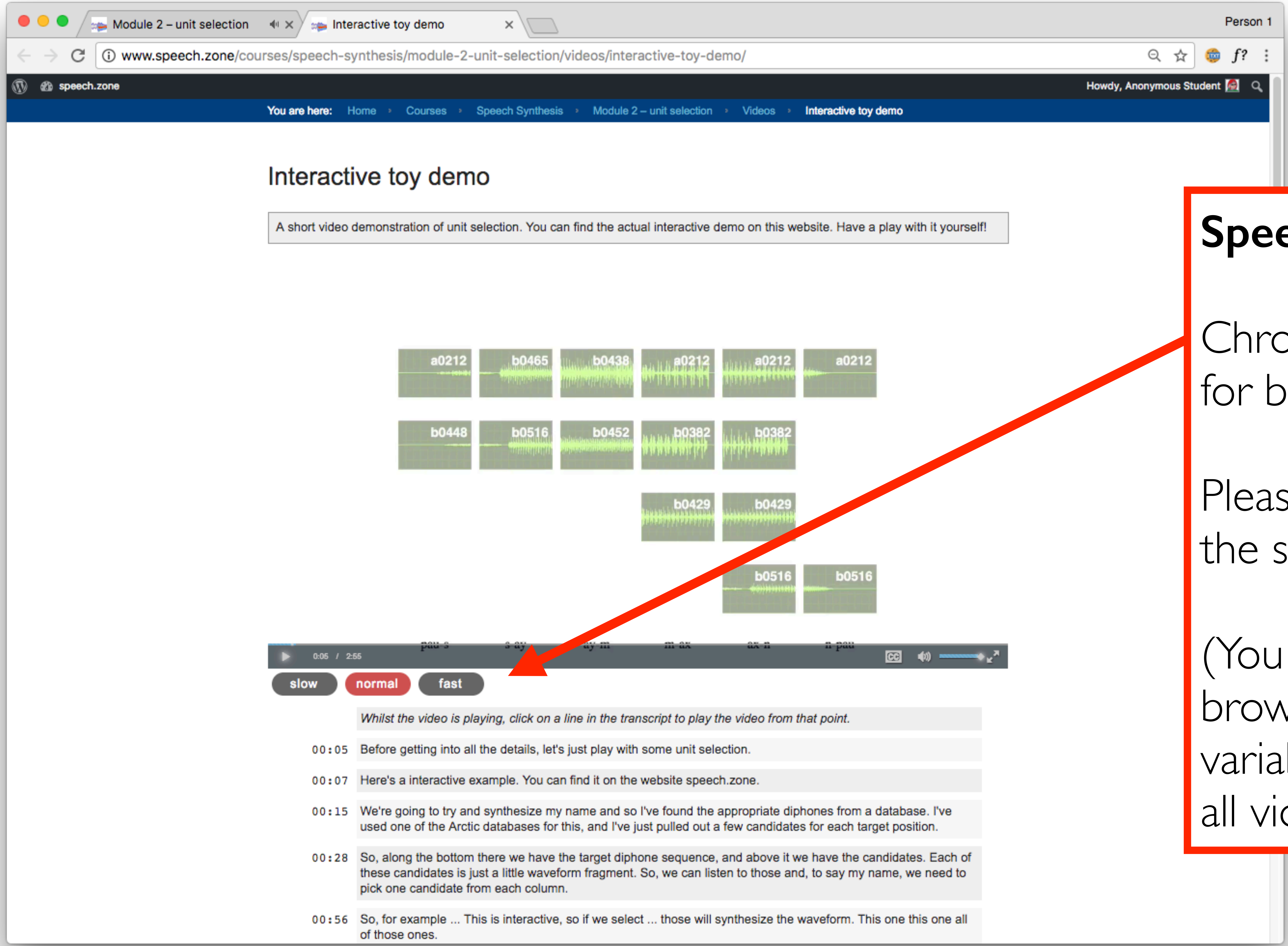
Jump to your next video:

Requires you to rate videos, which marks them as completed.

Subtitles
are being rolled out gradually across the modules.

Transcripts

Open the video in the pop-out window.
After starting playback, click on the transcript to jump to that point in the video.

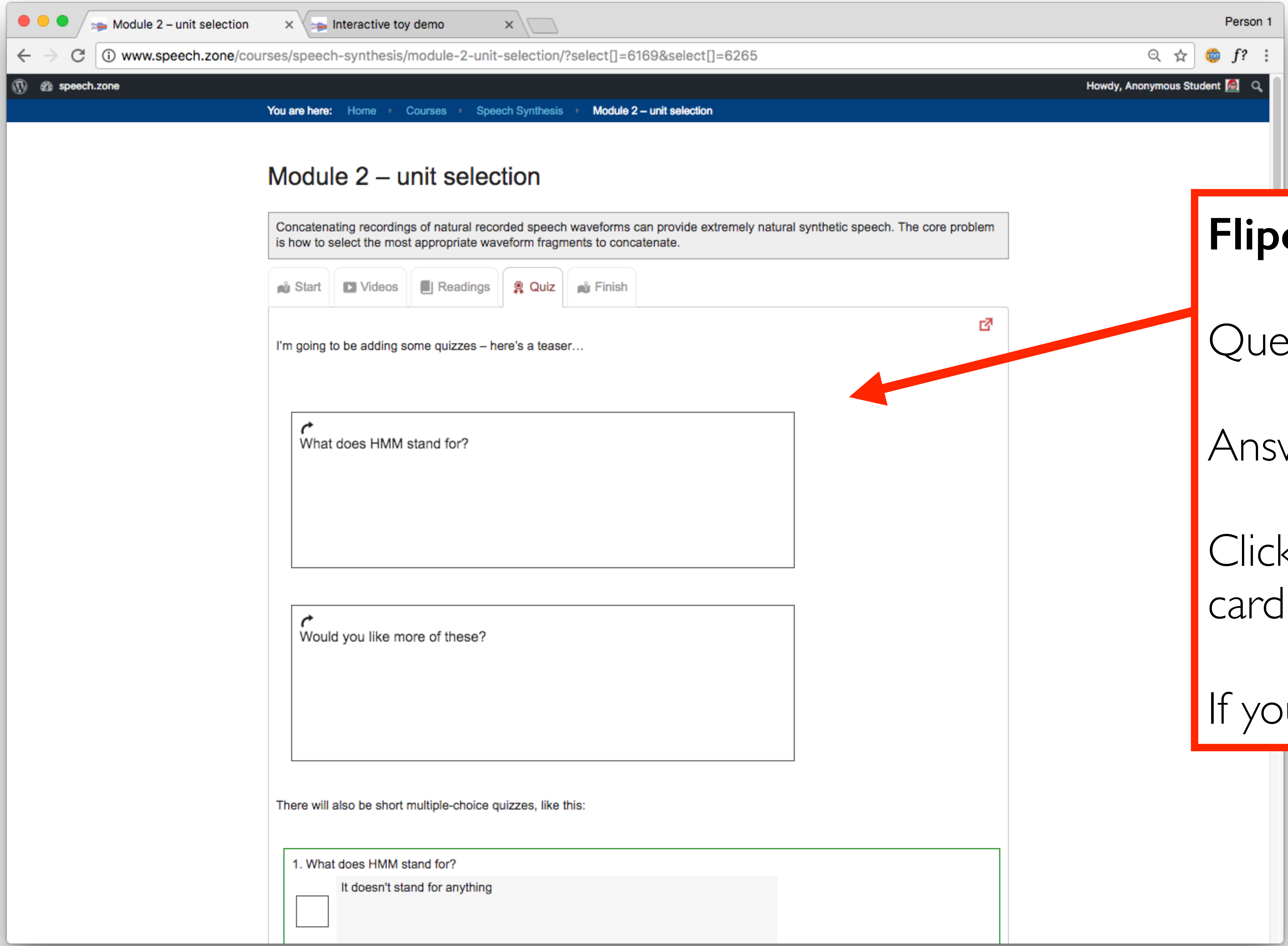


Speed controls

Chrome is recommended for best quality audio.

Please give feedback: are the speed settings right?

(You can also install a browser plugin to provide variable speed control for all videos on all websites.)



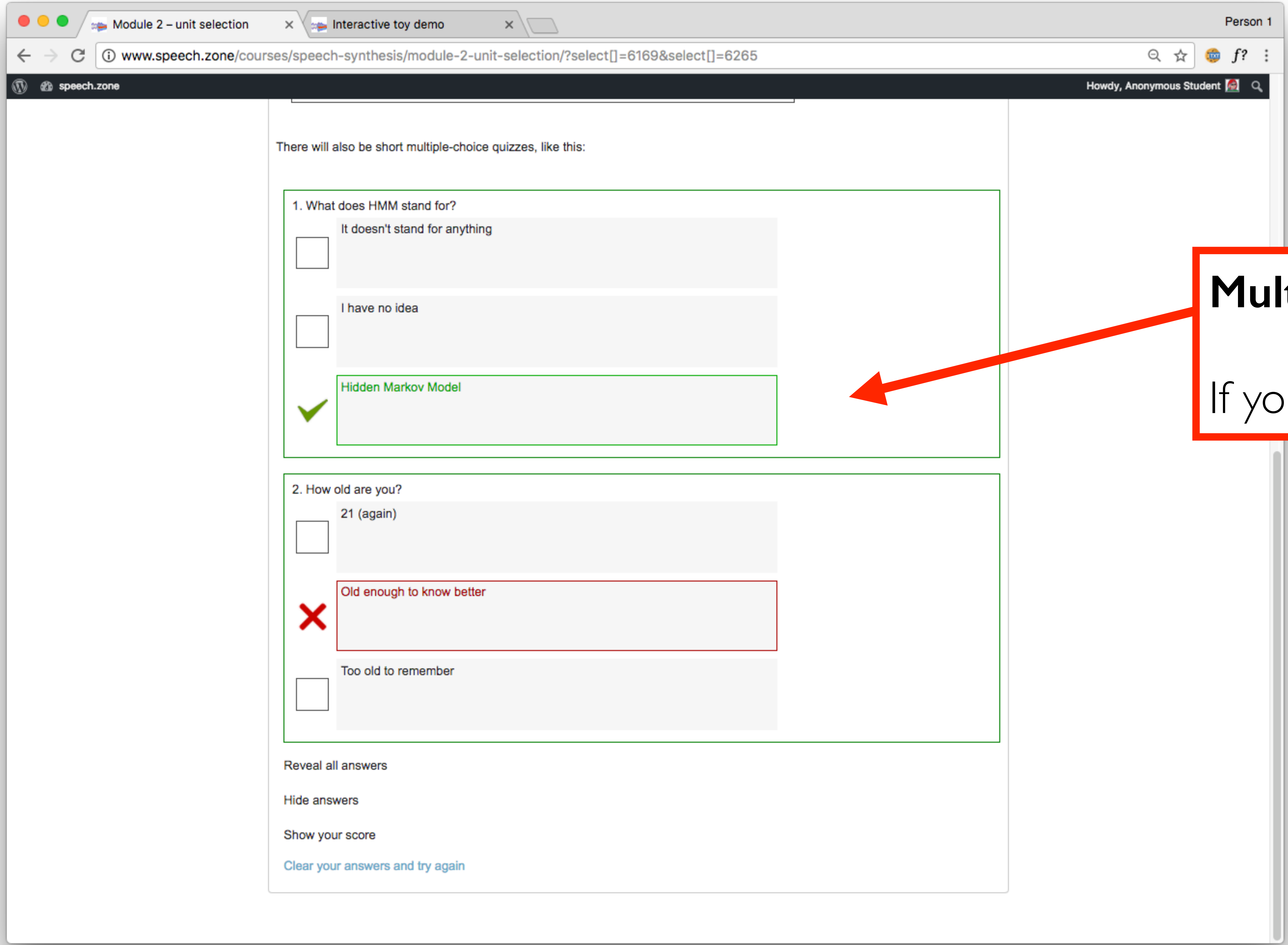
Flipcard quizzes

Question on one side.

Answer on the other.

Click anywhere on the card to flip it over.

If you like them, tell me!

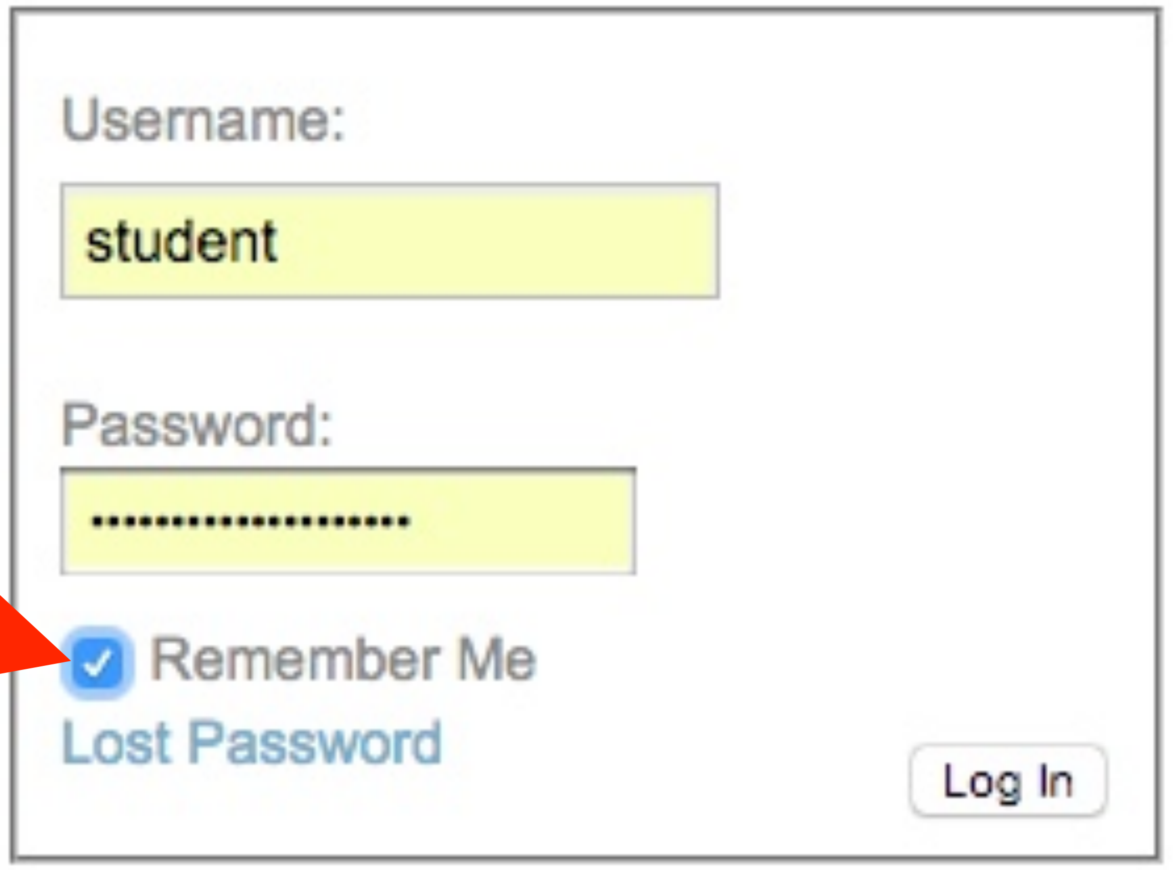


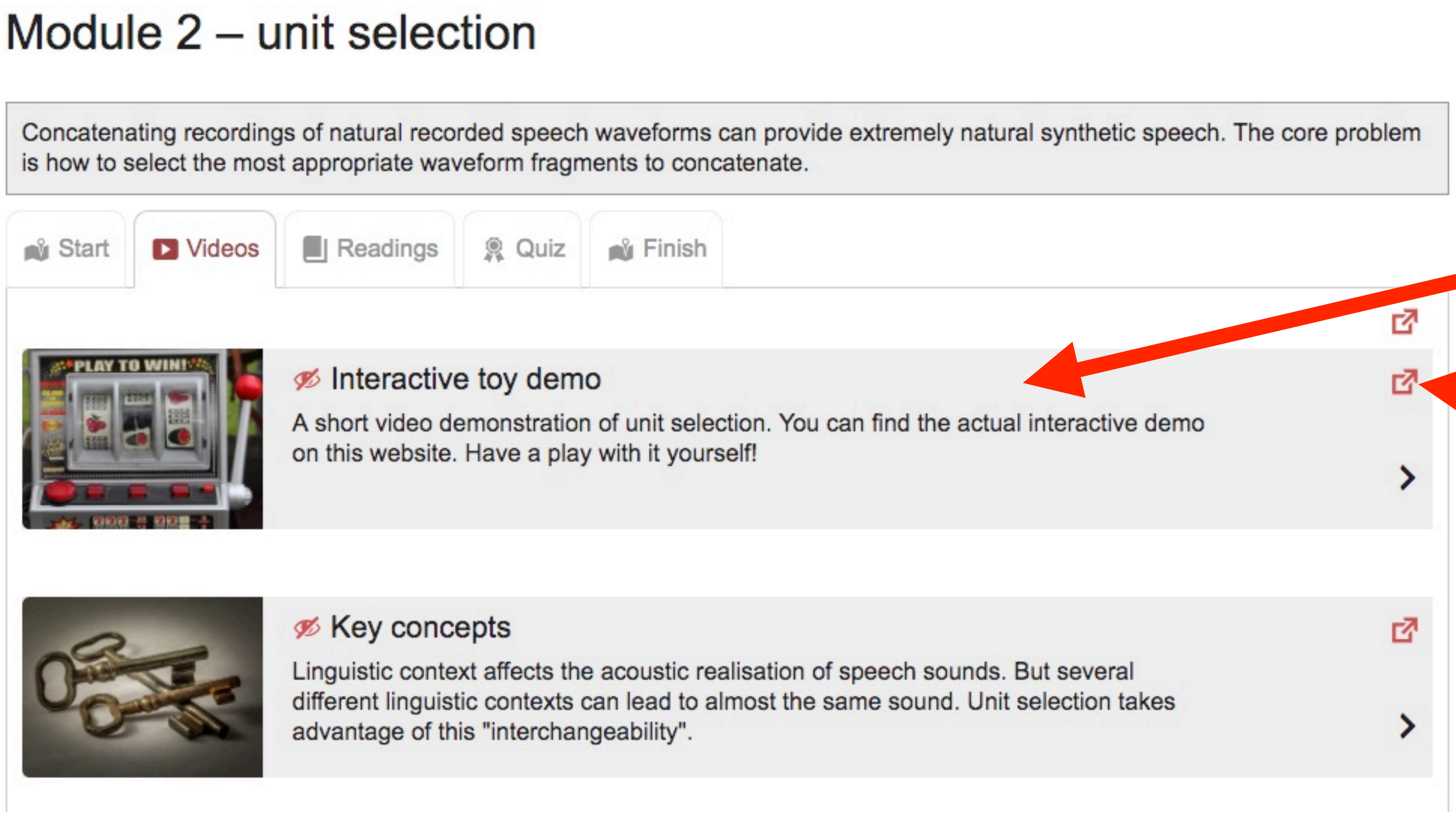
Multiple-choice quizzes

If you like them, tell me!

speech.zone tips

- check 'Remember me' to stay logged in for a year, in the current browser (otherwise, it's 2 days)





- click anywhere to open videos without leaving this page
- or, open in a new window to get extra features
  - transcripts
  - see other people's ratings

What **you** have to do (*)

- **Before each class**
  - complete the module specified in the course calendar, including
    - the videos + all <u>Essential</u> readings
  - post your questions on the forum

- **In each class**
  - actively participate in discussion of the course content
  - ask questions

(*) if you don't like this, then this course probably will not suit your learning style

Timetable

- **Class**
  - Tuesday 14:10 – 16:00
- **Lab** (you need to attend one session each week, but come to both if you fall behind)
  - group 1 : Wednesday 11:10 - 13:00 group 2 : Thursday 11:10 - 13:00
  - additional booked lab time (talking & discussion encouraged!) : Friday 10:00 - 11:50
- **Coursework**
  - deadline is in the course calendar on speech.zone
- **Class test** (UG only)
  - during the April/May exam period ; date to be announced later
  - examinable content = videos + Essential readings + class content for Module 8 onwards

Marking policy

- Same marking policy as **Speech Processing**
  - https://www.speech.zone/courses/speech-processing/marking-policy

- Please read the **Common Marking Scheme**

  60-69% = a good understanding of the video content and <u>Essential</u> readings

  70-79% = as above, plus most <u>Recommended</u> readings

  80%+ = as above, plus independent study, including further readings of your choice

Coursework: build your own unit selection voice

- Supervised lab sessions start this week
- The lab sessions will be led by Korin Richmond, with a tutor
- There is an introduction to the coursework within this lecture, after an outline of the course

Course outline

- **Introduction**
  - taster, brief history lesson, understanding the problem, list of current issues
- **Unit selection**
  - the method, and how to construct the speech database it relies upon
- **Signal processing**
  - vocoding, estimating F0 from speech signals
- **Statistical parametric speech synthesis**
  - the method, and its advantages over unit selection ; from HMMs to Deep Neural Networks
- **The latest developments (the "state of the art")**
  - from Deep Neural Networks to sequence-to-sequence models
  - open issues

Text-to-speech key challenges

- We can identify four main challenges for any builder of a TTS system.
  1. Semiotic classification of text
  2. Decoding natural-language text
  3. Creating natural, human-sounding speech
  4. Creating intelligible speech
- We can also identify two current and future main challenges
  1. Generating affective and augmentative prosody
  2. Speaking in a way that takes the listener's situation and needs into account

(Taylor 2009, Section 3.6, page 51)

Semiotic classification of text

- This is what we called "text normalisation" in Speech Processing

- Largely a solved problem (or at least solvable with current methods, given enough effort)
- Commercial systems do pretty good job of this
- Festival is reasonably good
  - improvements would be straightforward, but take a lot of effort

Decoding natural-language text

- In Speech Processing, we covered aspects of this, including:
  - homographs
    - disambiguate using POS tags
    - will fail for homographs with the same POS but different senses
  - shallow (“syntactic”) structure
    - phrase break prediction

- We can say that parts of this problem are solved
  - POS tagging, at least for well-resourced languages
- but that it’s not entirely clear how much ‘decoding’ is needed for speech synthesis
  - which prevents people solving the remaining problems

Creating natural, human-sounding speech

- Much to discuss here, from
  - **low-level** signal quality
    - concatenating waveforms vs. using models & classical vocoders vs. neural vocoders
  - **segmental** quality
    - pronunciation, stress, connected speech processes
  - **augmentative** prosody (text-related)
    - very much an open and important problem - even hard to define the scope!
  - **affective** prosody (not necessarily text-related)
    - some methods for generating 'affective' or 'emotional' speech, but few for predicting it (from what?)

Creating intelligible speech

- Closer to a solved problem than naturalness
  - interestingly, the most *natural-sounding* systems are **not** always the most *intelligible*
- Can achieve human levels of intelligibility
  - straightforward with good statistical parametric systems (example 1.6.1)
- Unit selection systems
  - generally less intelligible than natural speech (example 1.6.2)
  - but this is in lab conditions with semantically-unpredictable sentences

- In real applications, with 'normal' sentences, intelligibility is often at ceiling levels anyway, so differences between systems cannot be measured, and may not matter

Understanding the problem

- **Input is text**
  - what *properties of text* do we need to know about?
- **Output is speech**
  - what *properties of speech* do we need to know about?
- **How hard is the conversion from text to speech?**
  - Do we need to understand the text?
  - If so, *how* would we do that?
  - If not, what *do* we need to extract from the text?

What properties of text do we need to know about?

"*it is not necessary to go all the way and uncover the meaning from the written signal;*

we have to perform just the job of text decoding, not also that of text understanding

……

by and large, the identity and order of the words to be spoken is all we require to synthesise speech;

no higher-order analysis or understanding is necessary."

(Taylor 2009, Section 3.1.2, page 29)

but Taylor adds two caveats:

- word sense disambiguation (e.g., "polish", "lead", "bass")
- **prosody (a *huge* caveat !!)**

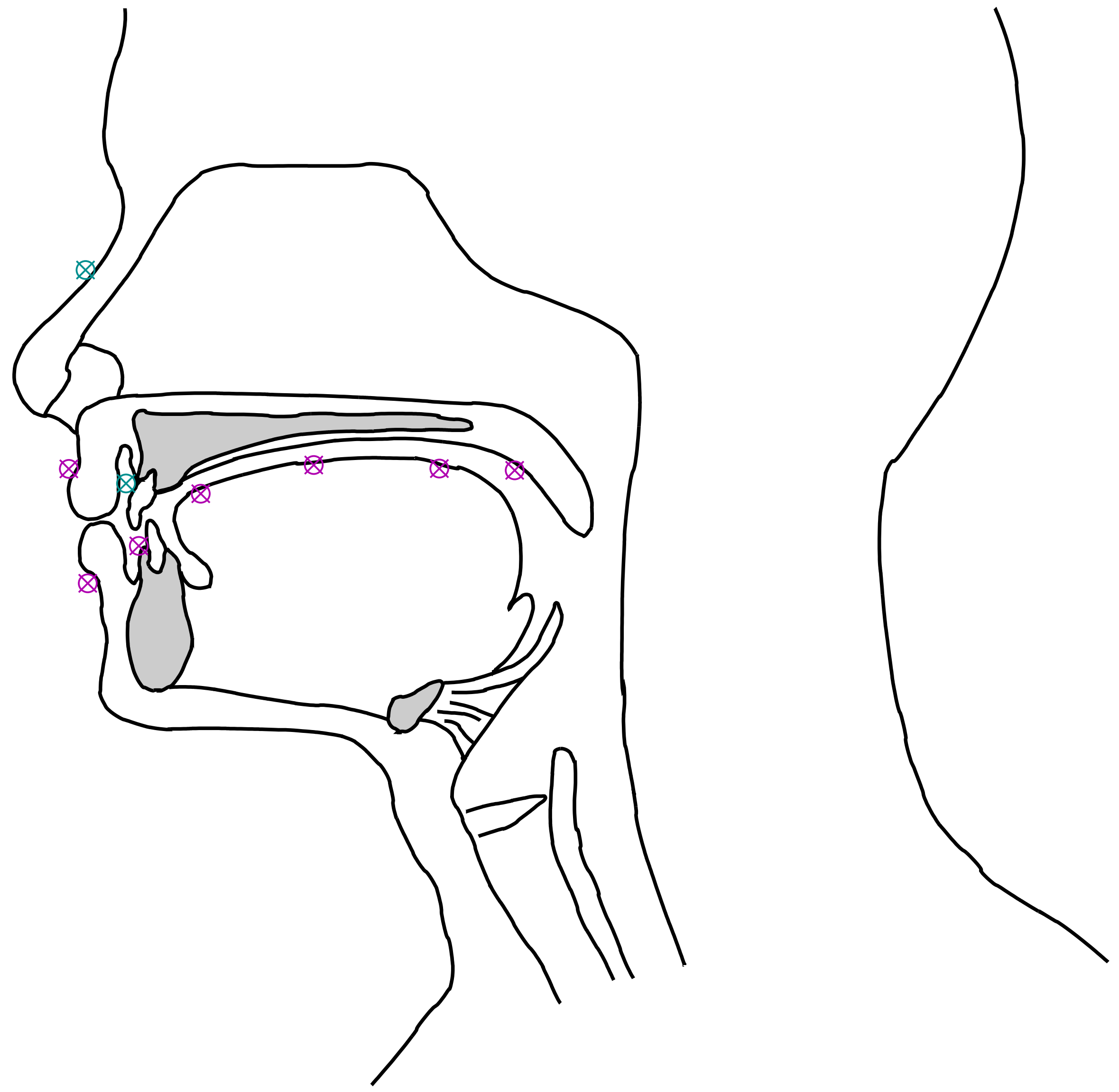
What properties of speech do we need to know about?

- To start us thinking about the issues involved in creating synthetic speech, let's think first about what speech is "made of", because
  - in speech synthesis, we need to say **new** things (i.e., utterances not in our recorded database)
  - in speech recognition, we need to **generalise** from the examples in the training data to the speech we have to recognise

- It is convenient to think about speech as a **linear** sequence of units
  - enables a concatenative approach to speech synthesis
  - in speech recognition, allows us to string together models of small units (e.g. phonemes) to make models of larger units (e.g. words)

Speech production

- Observed signal is result of several interacting processes
- The **context** in which a speech sound is produced affects that sound
  - articulatory constraints: where the articulators are coming from / going to
  - phonological effects
  - prosodic environment



Units of speech

- The speech signal we observe (the waveform) is the product of interacting processes operating at different time scales
  - at any moment in time, the signal is affected not just by the current phoneme, but many other aspects of the context in which it occurs
  - the context is complex - it's not just the preceding/following sounds

- How can we reconcile this conflict, when we want to simultaneously:
  - model speech as a simple string of units
  - take into account all the long-range effects of context, before, during and after the current moment in time

Context is the key

- Context-dependent units offer a solution
  - engineer the system in terms of a simple linear string of units
  - then account for context by having a different version of each unit for every different context
- But, how do we know what all the different contexts are?
- If we enumerate all possible contexts, they will be practically infinite
  - there are an infinite number of different sentences in a language
  - context potentially spans the whole sentence (or further)
- However, what is important is the **effect** that the context has on the current speech sound - so next we can think about reducing the number of *effectively different* contexts

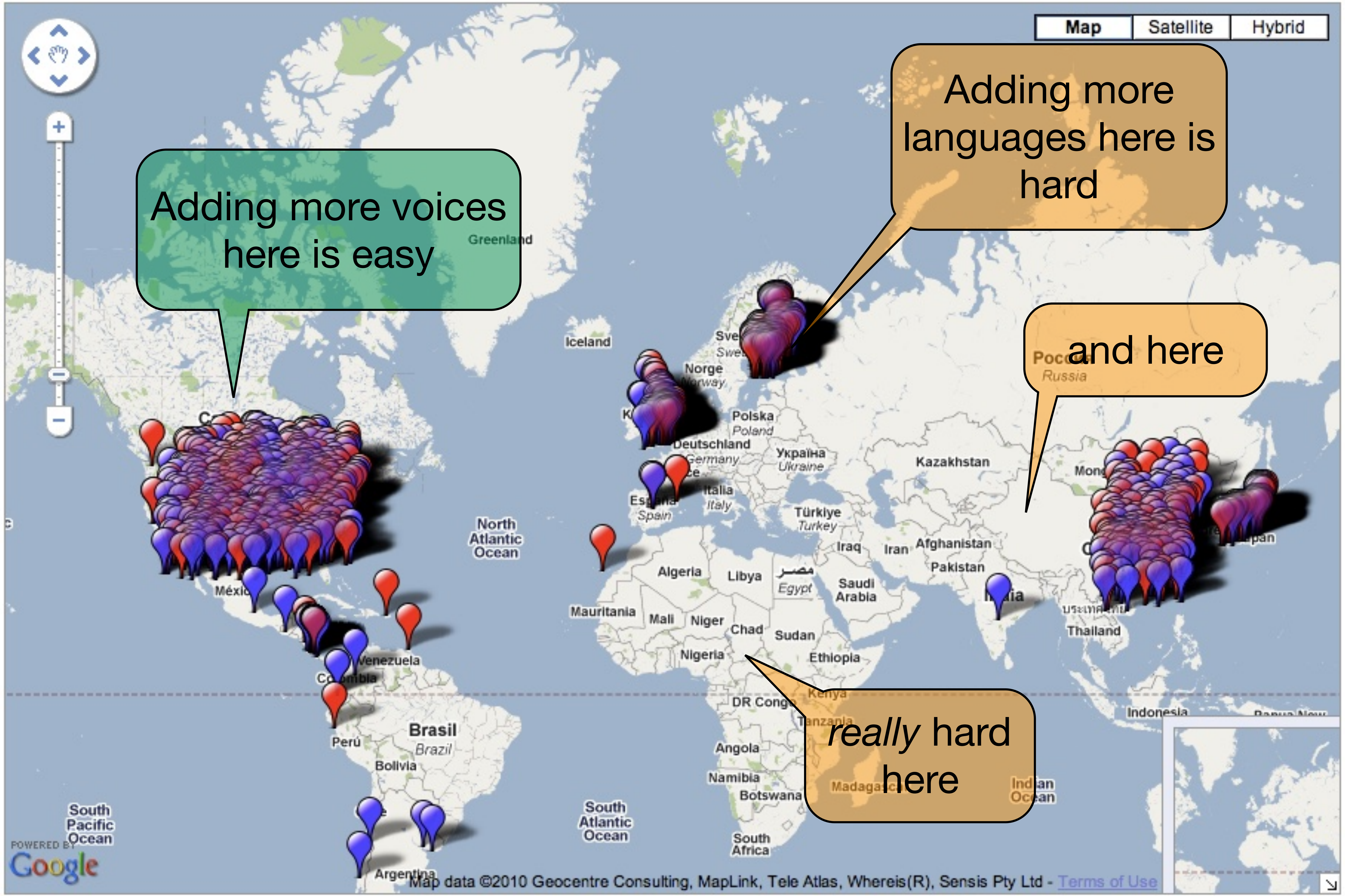
Current issues

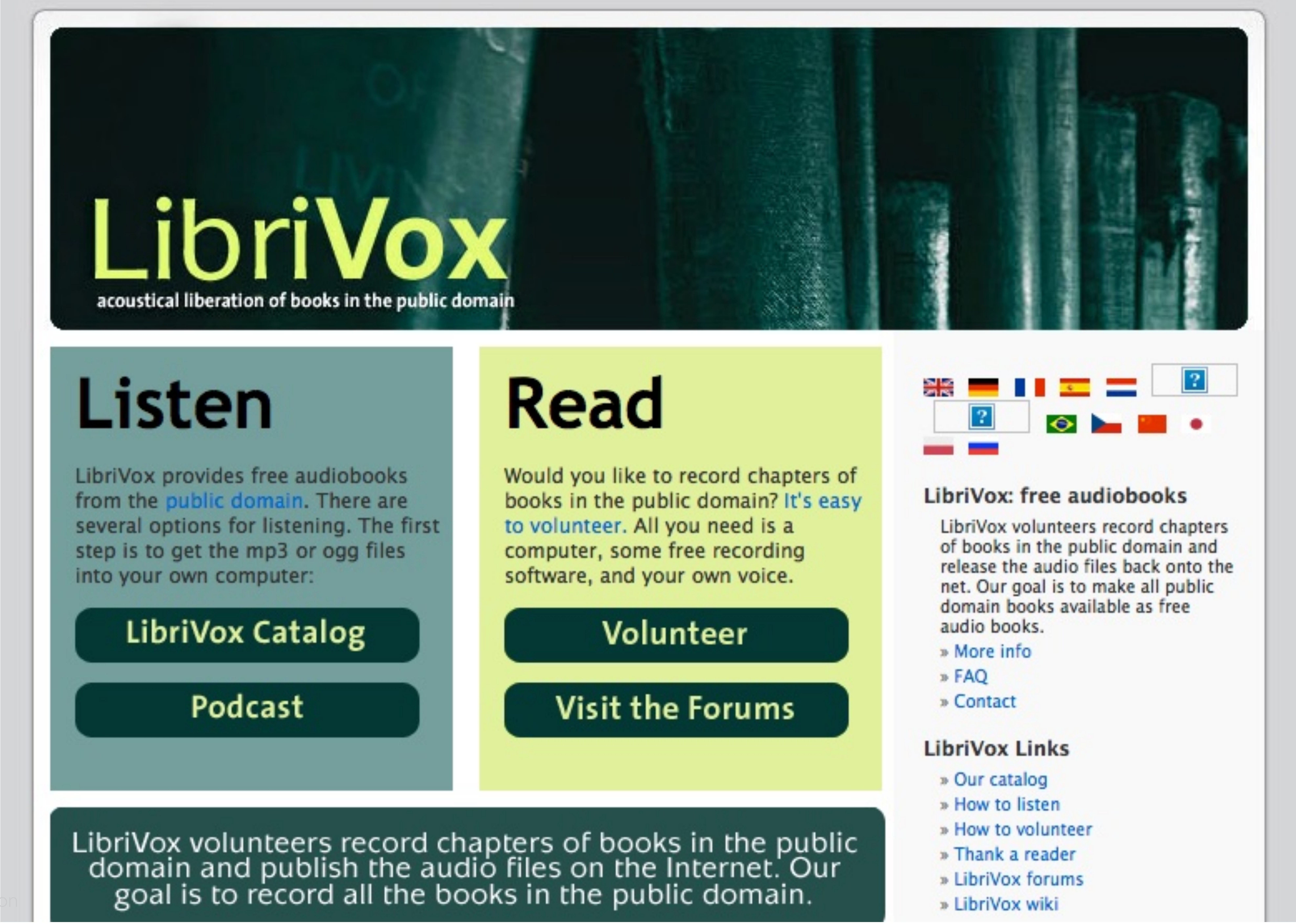
- **Deployed commercial systems**
  - heavily reliant on high-quality speech, professionally recorded in a studio
  - multiple languages (but typically fewer than 50)
  - speaking styles from a fixed set (e.g., newscaster, narrator)
  - adaptation and control using markup, speech exemplars, Human-in-the-Loop
- **Recent and emerging techniques**
  - extensive use of 'found data' (necessitated by models that require a lot of data)
  - more powerful forms of control over speaking style
  - rapid expansion to more languages

Assistive communication devices

THE UNIVERSITY OF EDINBURGH

Voices easy. Languages harder!




LibriVox
acoustical liberation of books in the public domain
Listen
LibriVox provides free audiobooks from the public domain. There are several options for listening. The first step is to get the mp3 or ogg files into your own computer:
LibriVox Catalog
Podcast
Read
Would you like to record chapters of books in the public domain? It's easy to volunteer. All you need is a computer, some free recording software, and your own voice.
Volunteer
Visit the Forums
LibriVox: free audiobooks
LibriVox volunteers record chapters of books in the public domain and release the audio files back onto the net. Our goal is to make all public domain books available as free audio books.
» More info
» FAQ
» Contact
LibriVox Links
» Our catalog
» How to listen
» How to volunteer
» Thank a reader
» LibriVox forums
» LibriVox wiki
LibriVox volunteers record chapters of books in the public domain and publish the audio files on the Internet. Our goal is to record all the books in the public domain.

Synthetic speech created from audiobooks

1 paragraph example

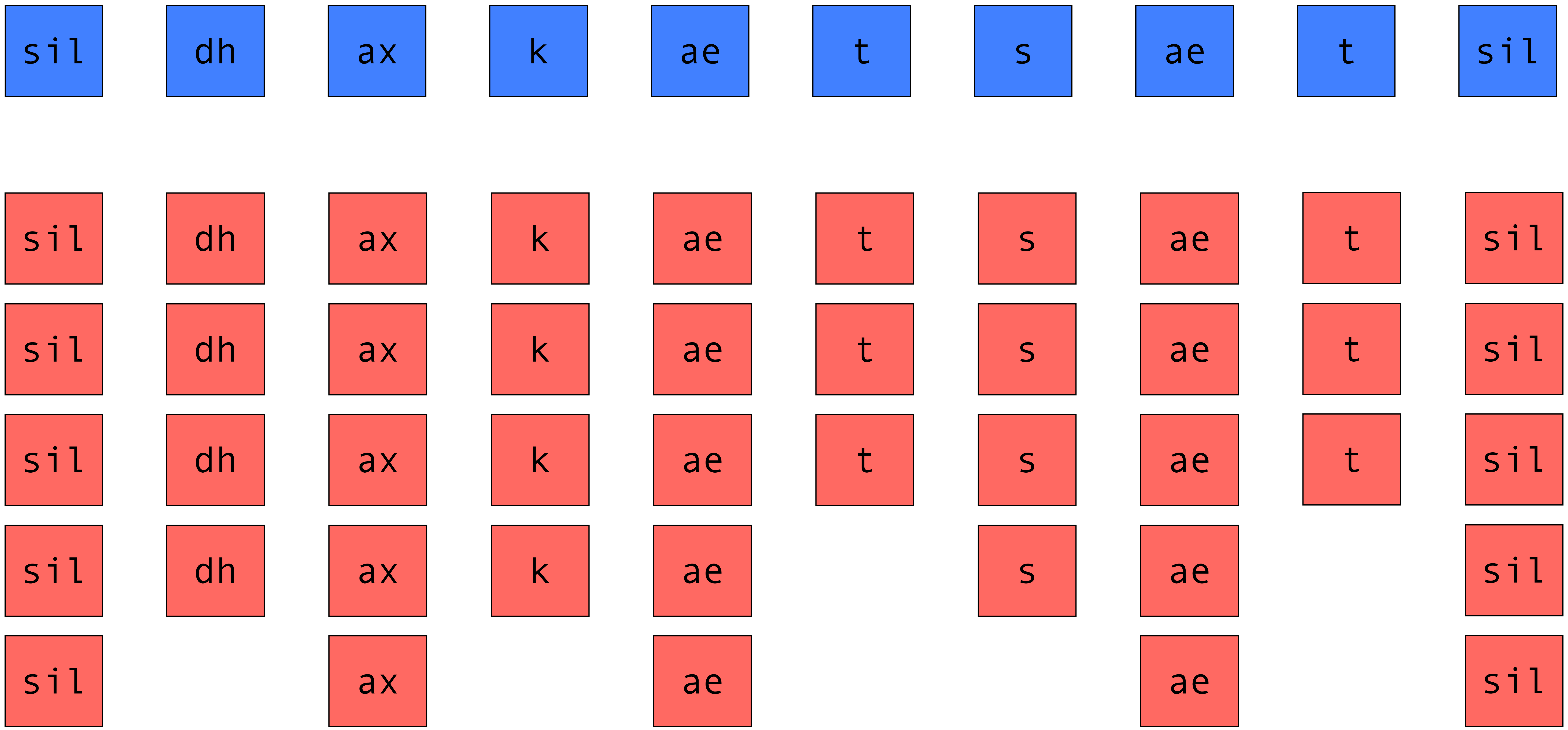
Audio credits:Speech and Hearing Research Center, Peking University

Current issues

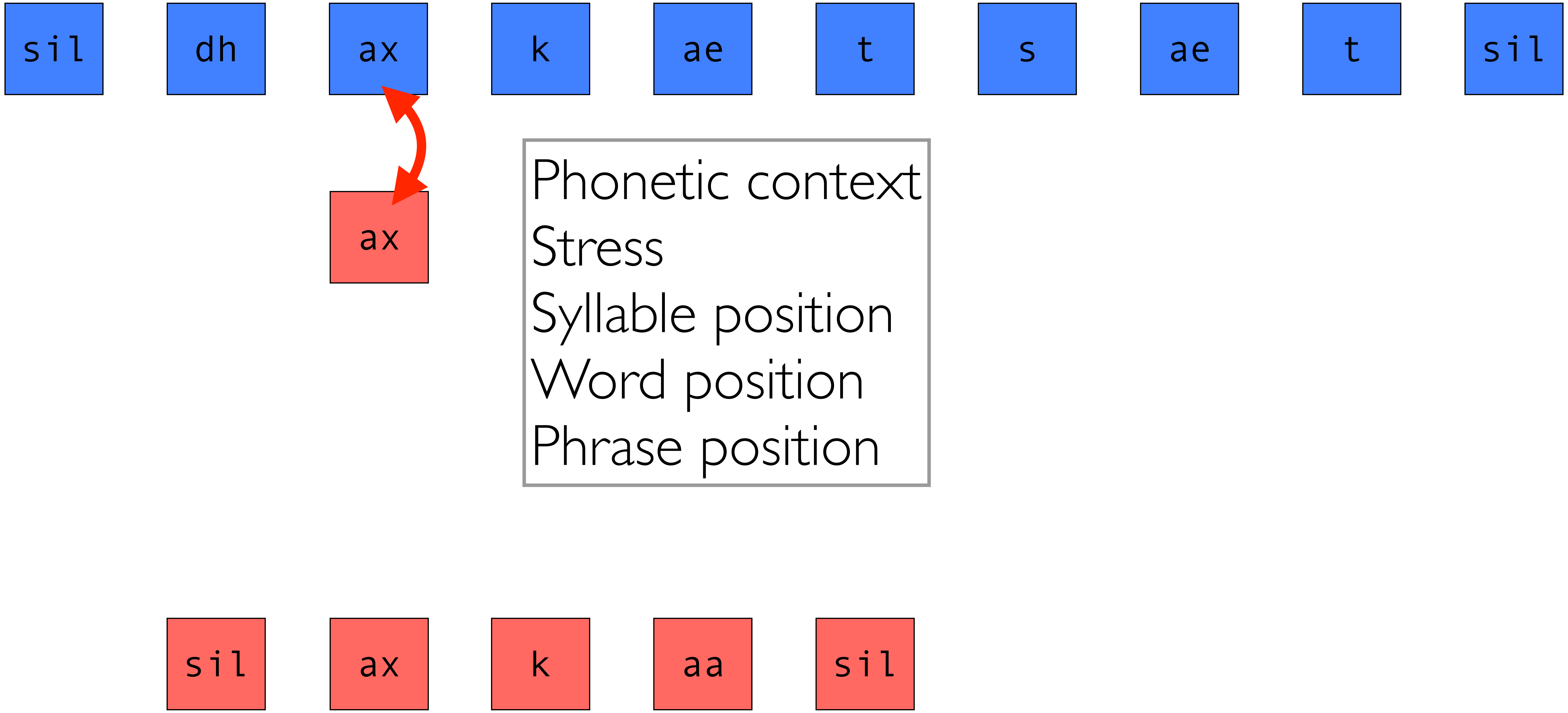
- **What is being actively researched**
  - neural **deep learning** approaches - mostly sequence-to-sequence models
  - better and faster neural **vocoders**
  - semi-, self-, and **un-supervised learning**, to reduce reliance on expensive labelled data
  - **prosody**, including its relationship to the meaning of the text (what Taylor calls "Generating affective and augmentative prosody")
  - listener and situation-**appropriate synthesis** (what Taylor calls "Speaking in a way that takes the listener's situation and needs into account.")
    - for impaired listeners
    - for speech-to-speech translation, including dubbing

A tour of the remaining modules

Module 2 - unit selection



Module 3 - unit selection target cost functions



Phonetic context
Stress
Syllable position
Word position
Phrase position

Module 4 - the database

So I came here.	sil_s s_ow ow_ay ay_k k_ey ey_m m_hh hh_ih ih_r r_sil
Now we have finally heard her.	sil_n n_aw aw_w w_iy iy_hh hh_ae ae_v v_f f_ay ay_n n_ax ax_l l_iy ~~iy_hh~~ hh_er er_d d_hh ~~hh_er~~ er_sil
Those chefs know who they are.	sil_dh dh_ow ow_z z_sh sh_eh eh_f f_s s_n n_ow ow_hh hh_uw uw_dh dh_ey ey_aa aa_r r_sil
…etc	

aa_aa	aa_f		ay_ey		ey_f		hh_f		zh_f	zh_p
aa_ae	aa_g		ay_f		ey_g		hh_g		zh_g	zh_r
aa_ah	aa_hh		ay_g		ey_hh		hh_hh		zh_hh	zh_s
aa_ao	aa_ih		ay_hh		ey_ih		hh_ih		zh_ih	zh_sh
aa_aw	aa_iy		ay_ih		ey_iy		hh_iy		zh_iy	zh_t
aa_ay	aa_jh		ay_iy		ey_jh		hh_jh		zh_jh	zh_th
aa_b	aa_k	•••	ay_jh	•••	ey_k	•••	hh_k	•••	zh_k	zh_uh
aa_ch	aa_l		ay_k		ey_l		hh_l		zh_l	zh_uw
aa_d	aa_m		ay_l		ey_m		hh_m		zh_m	zh_v
aa_dh	aa_n		ay_m		ey_n		hh_n		zh_n	zh_w
aa_eh	aa_ng		ay_n		ey_ng		hh_ng		zh_ng	zh_y
aa_er	aa_ow		ay_ng		ey_ow		hh_ow		zh_ow	zh_z
aa_ey	aa_oy		ay_ow		ey_oy		hh_oy		zh_oy	zh_zh

Module 5 - evaluation

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]**.

4 : Mostly Natural

Submit

Module 6 - speech signal analysis & modelling

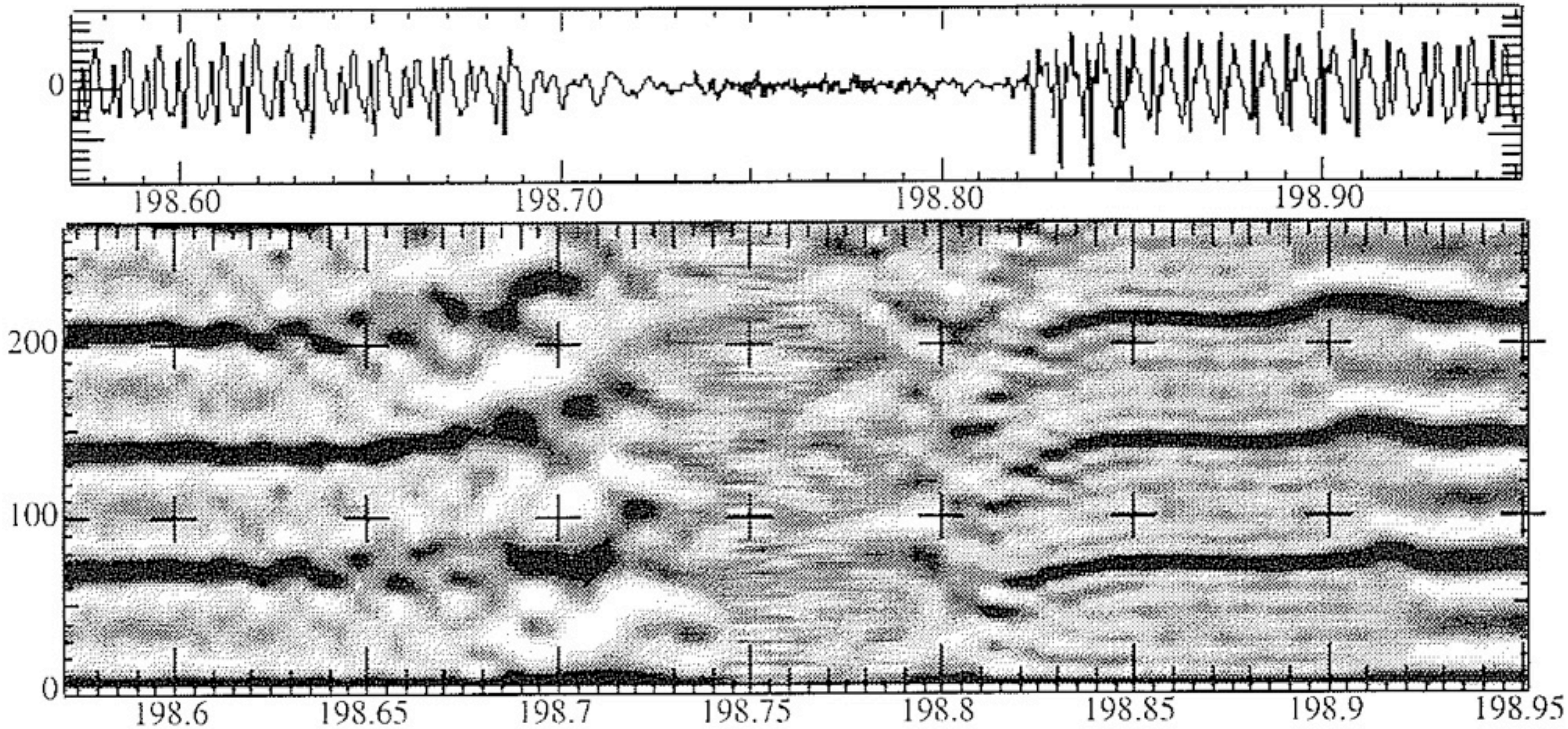
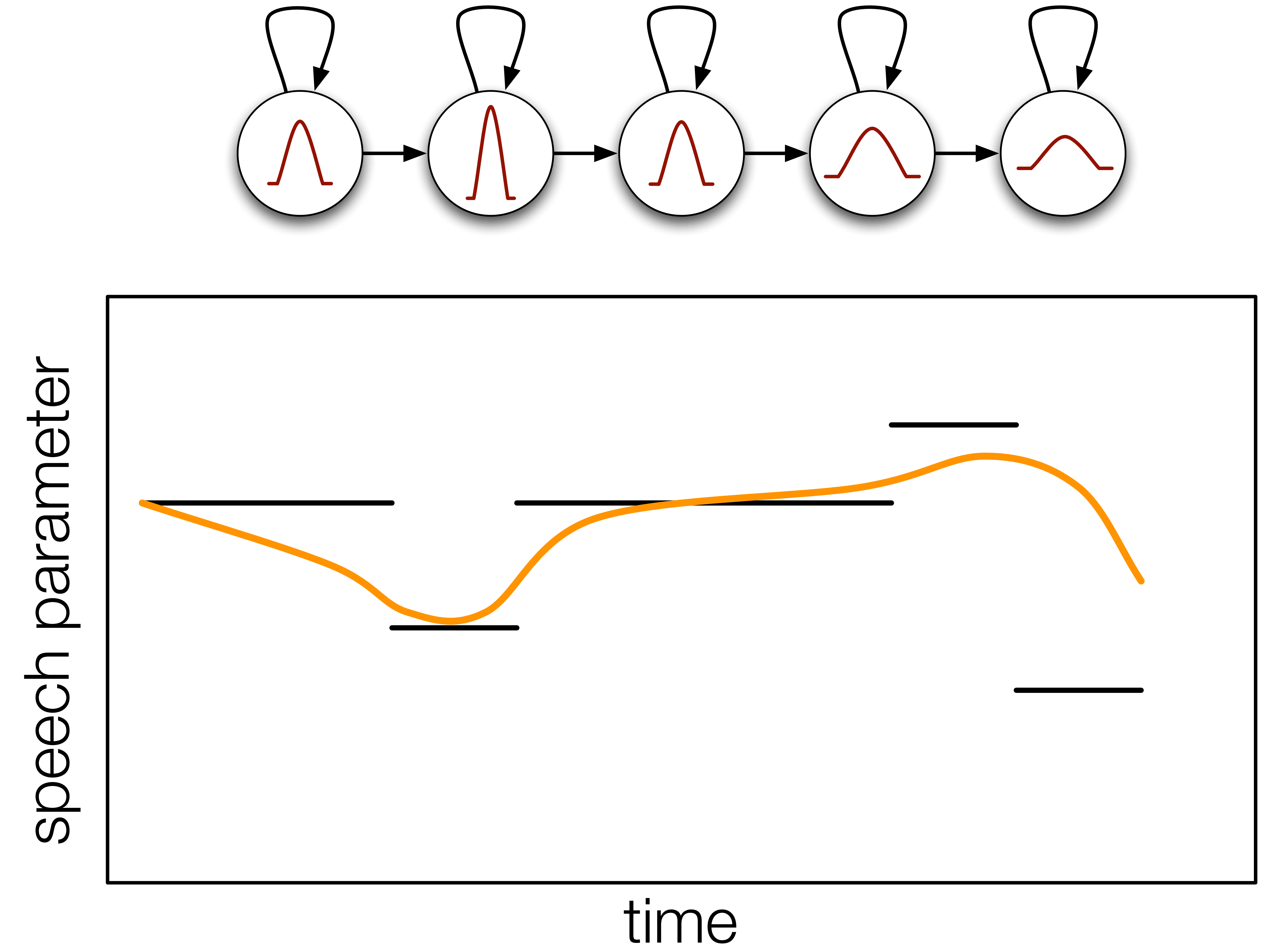
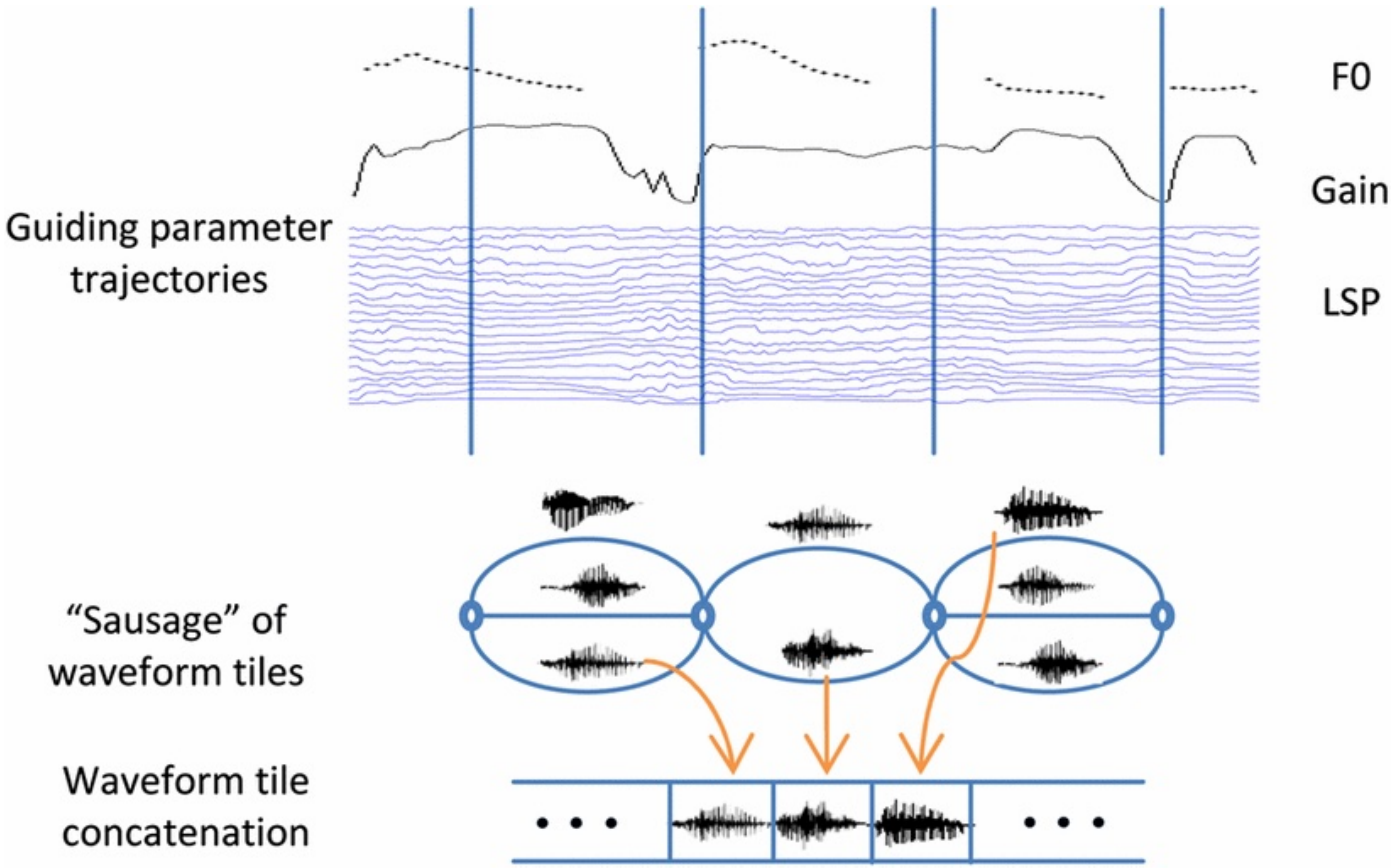


Figure 2 from David Talkin "A Robust Algorithm for Pitch Tracking (RAPT)" in Speech Coding and Synthesis, W. B. Kleijn and K. K. Palatal (eds), pages 497-518 Elsevier Science B.V., 1995

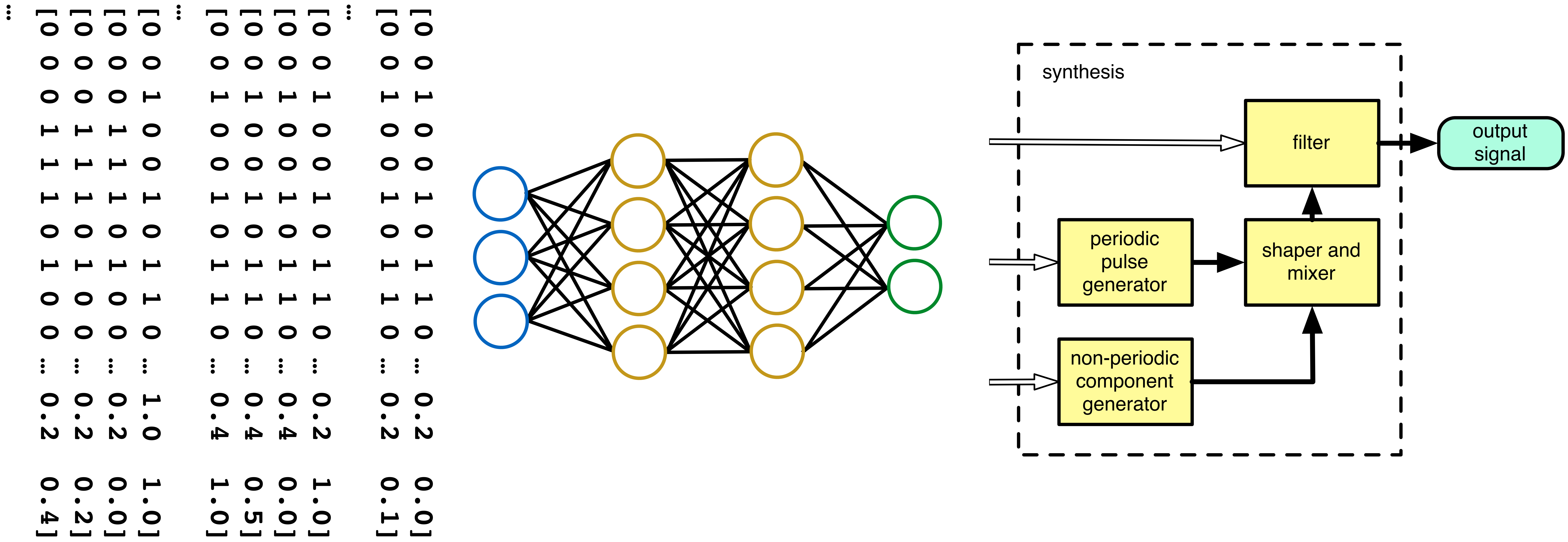
Module 7 - Statistical Parametric Speech Synthesis (SPSS)



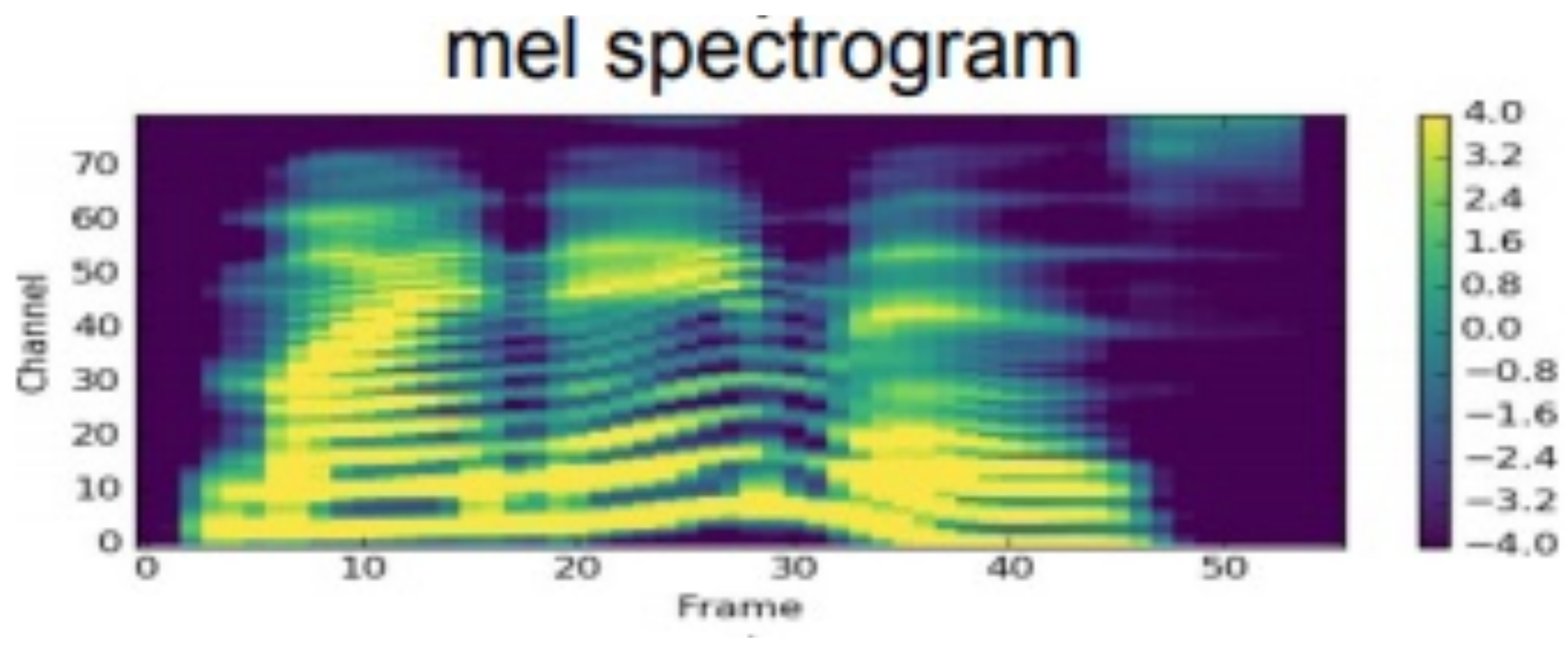
Module 7 bonus material (non-examinable) - hybrid speech synthesis

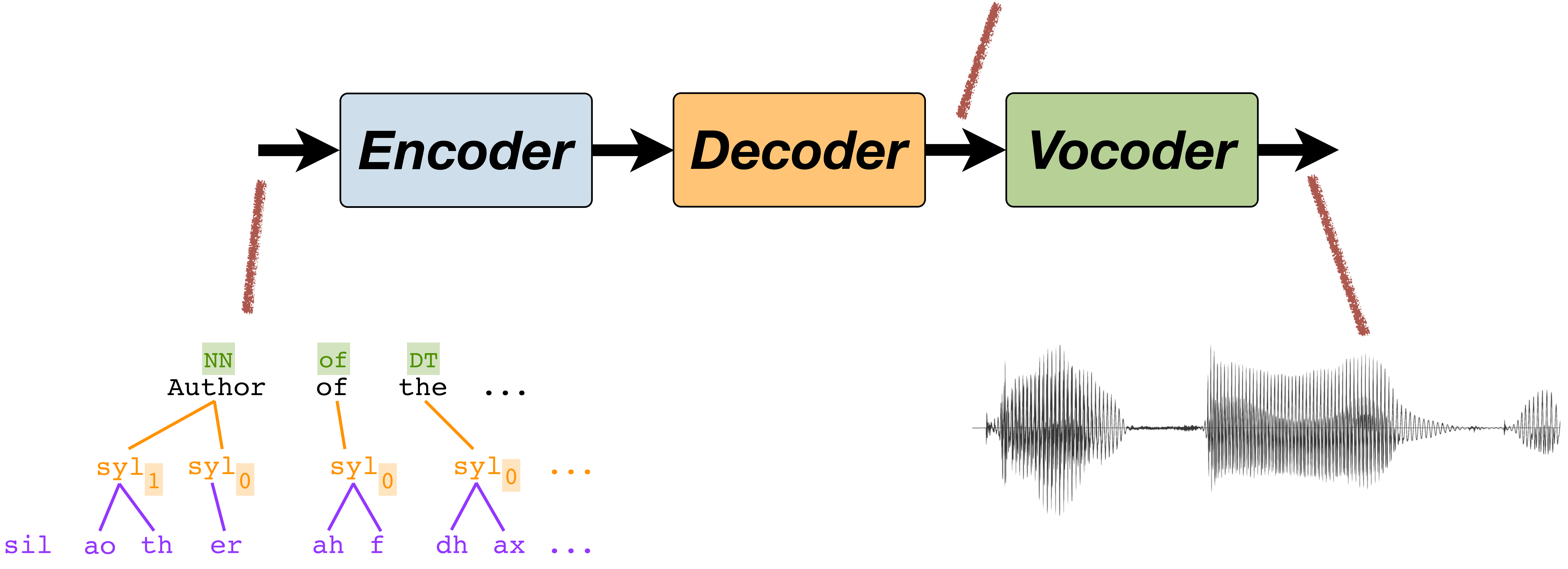


Module 8 - Deep Neural Networks (DNNs)



Module 9 - sequence-to-sequence models



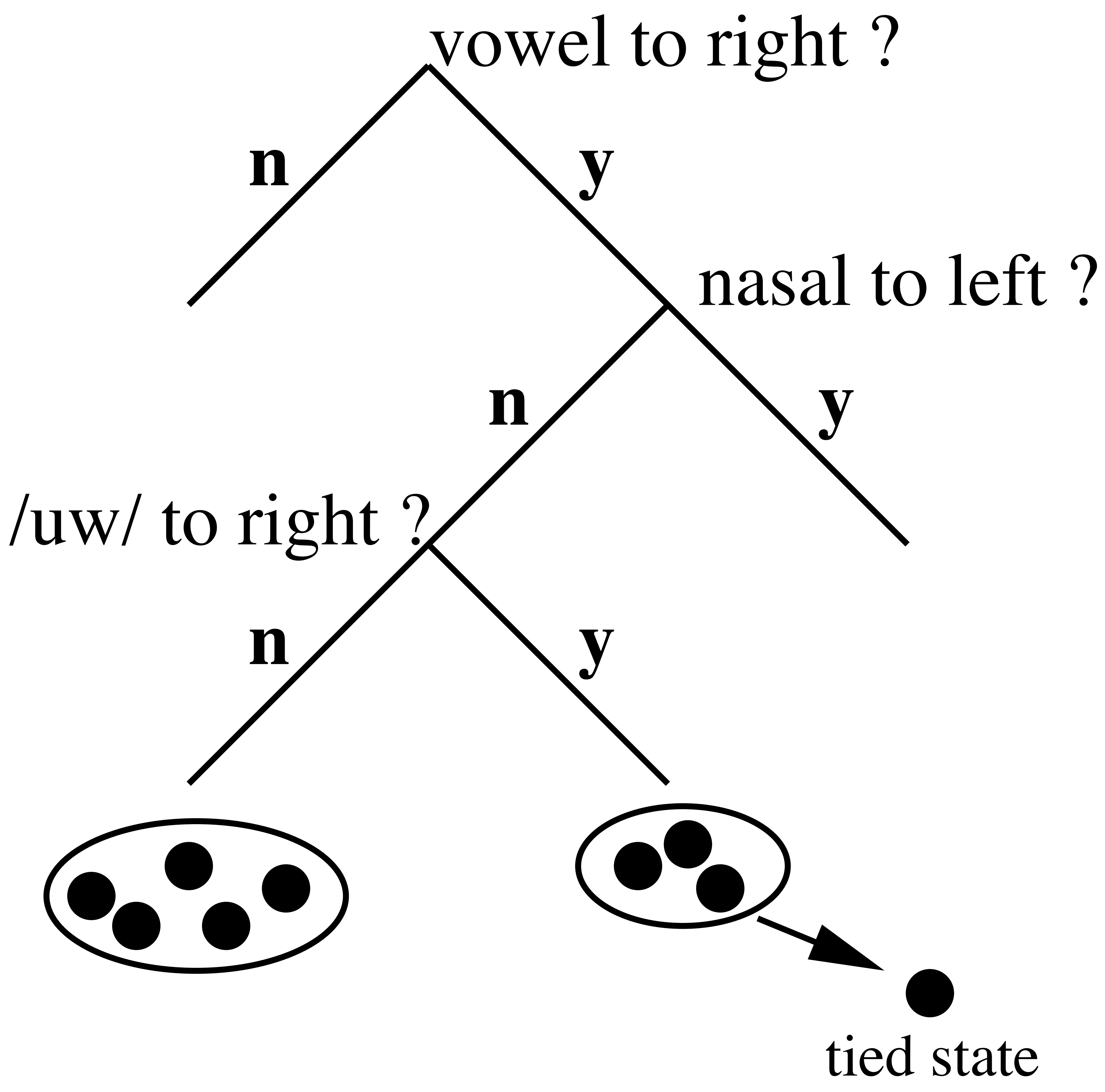


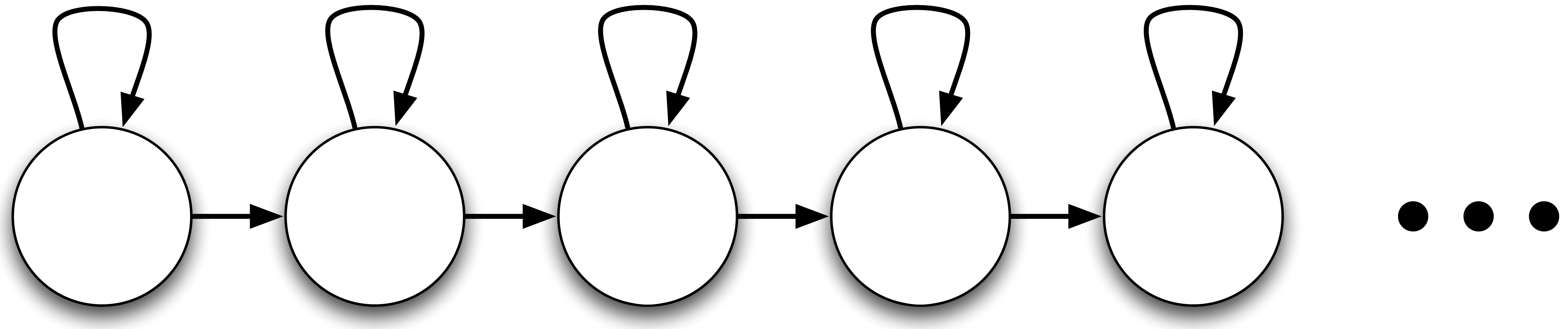
The state of the art (1 or 2 further lectures after Module 9)

Context is everything : unit selection

sil	dh	ax	k	ae	t	s	ae	t	sil
sil	dh	ax	k	ae	t	s	ae	t	sil
sil	dh	ax	k	ae	t	s	ae	t	sil
sil	dh	ax	k	ae	t	s	ae	t	sil
sil	dh	ax	k	ae		s	ae		sil
sil		ax		ae			ae		sil

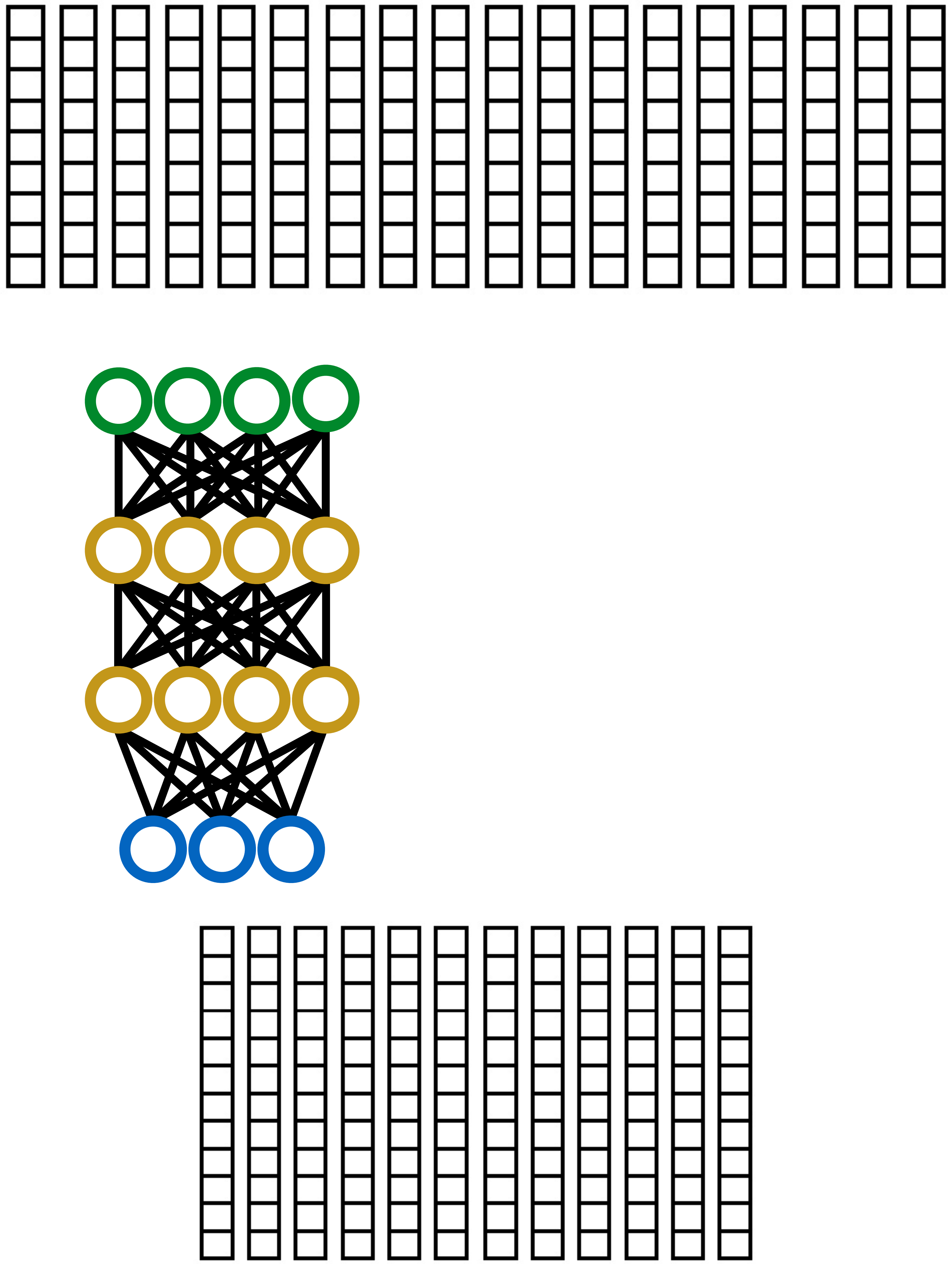
Context is everything : HMMs (+ regression trees)





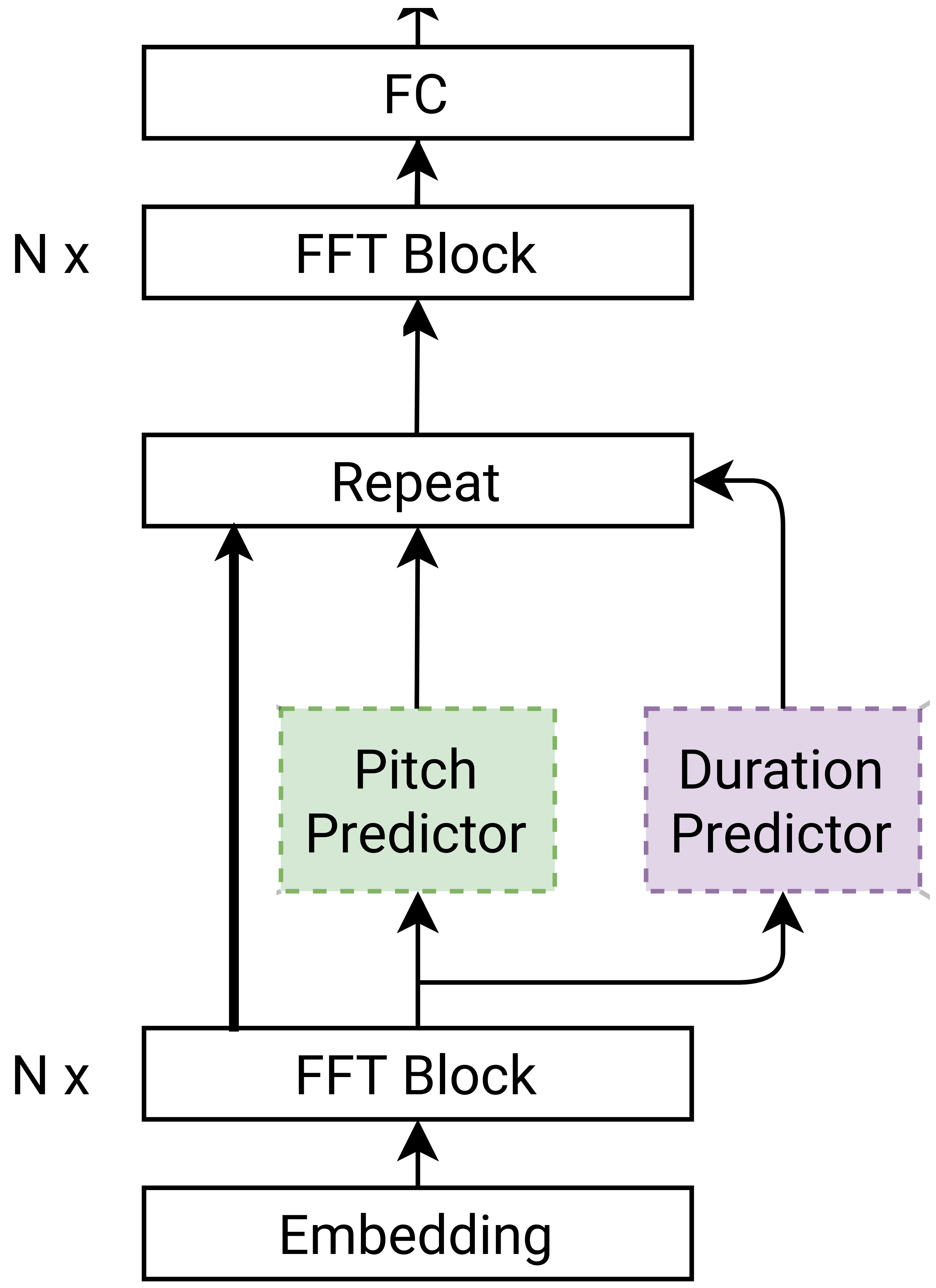
Context is everything : first attempts using Deep Neural Networks

output sequence

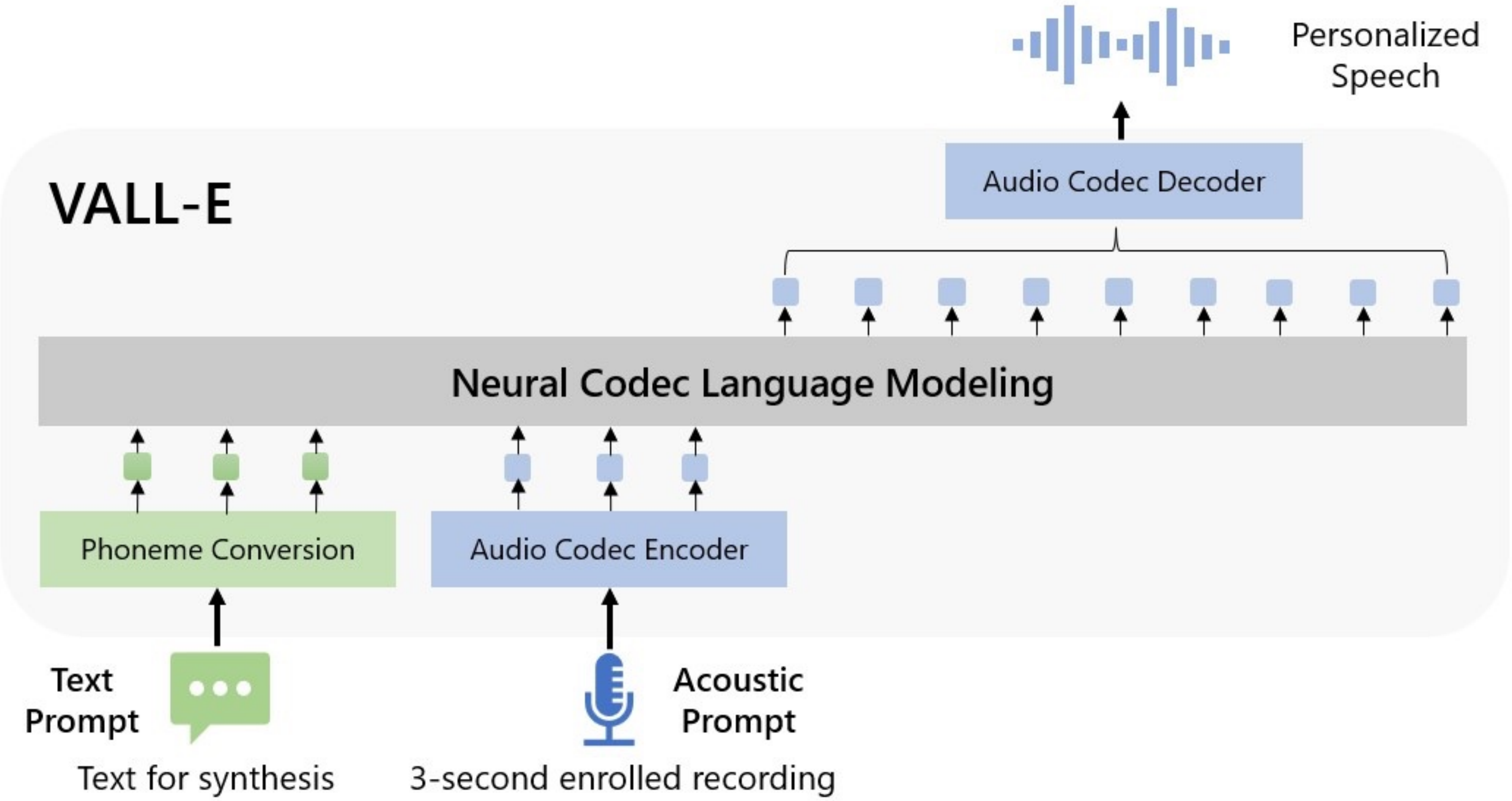


input sequence

Context is everything : encoder-decoder (e.g., FastPitch)



Context is everything : (large) language model approaches



https://www.microsoft.com/en-us/research/project/vall-e-x

Representation of written form, spoken form, and "everything in-between"

phrase initial

pitch accent

phrase final

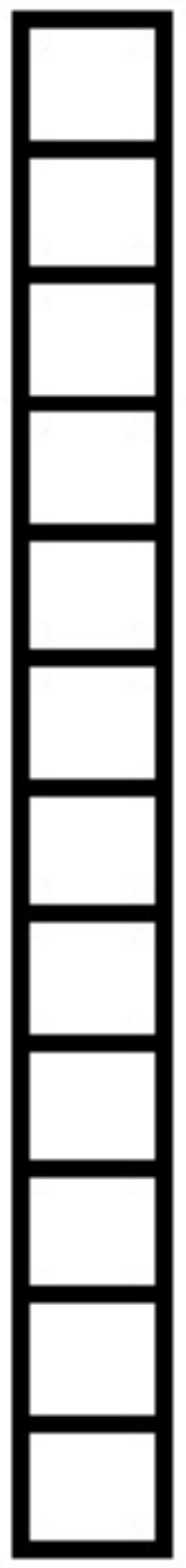
sil dh ax k ae t s ae t sil

"the cat sat"

DET NN VB

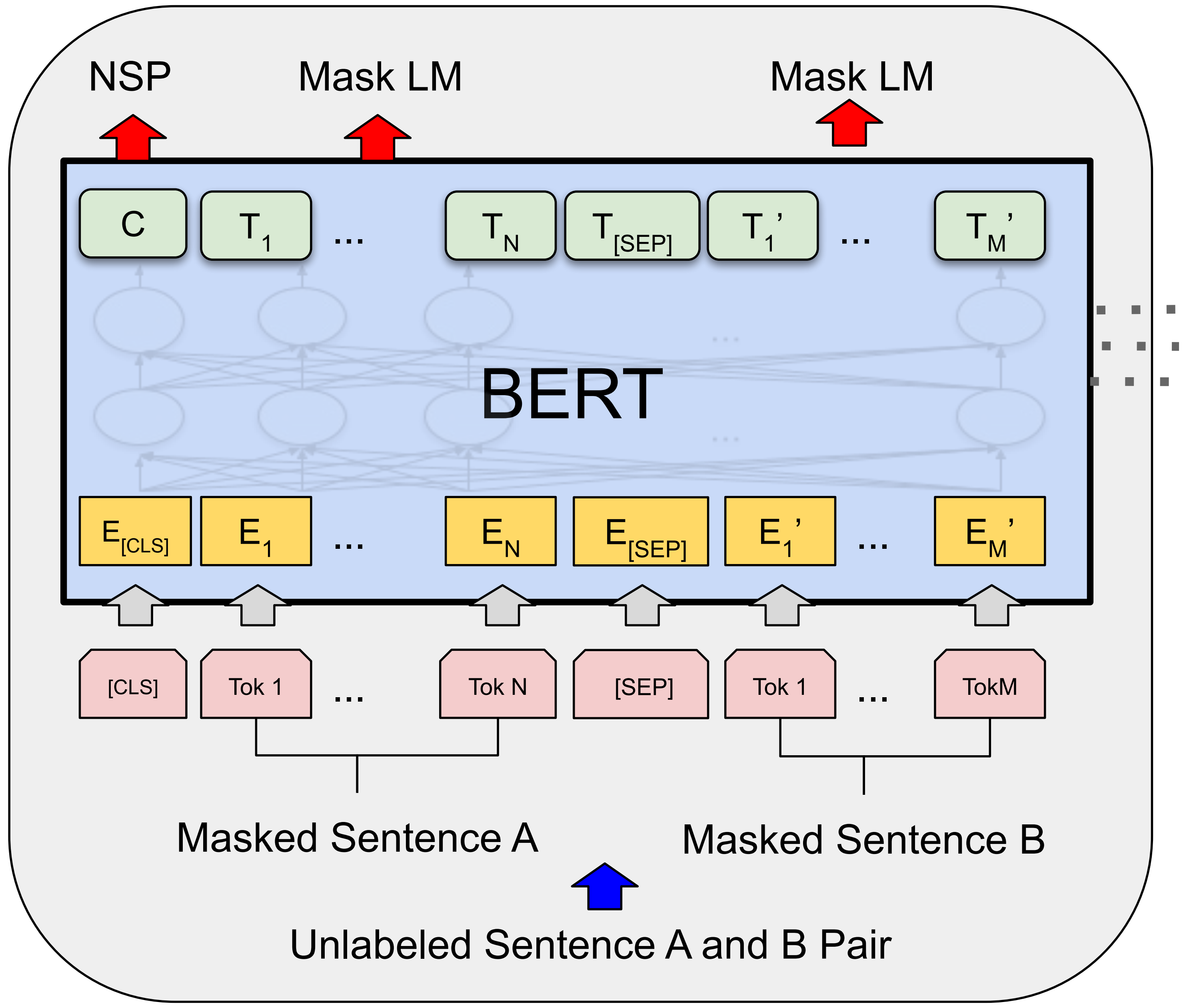
((the cat) sat)

sil^dh-**ax**+k=ae, "phrase initial", "unstressed syllable", ...

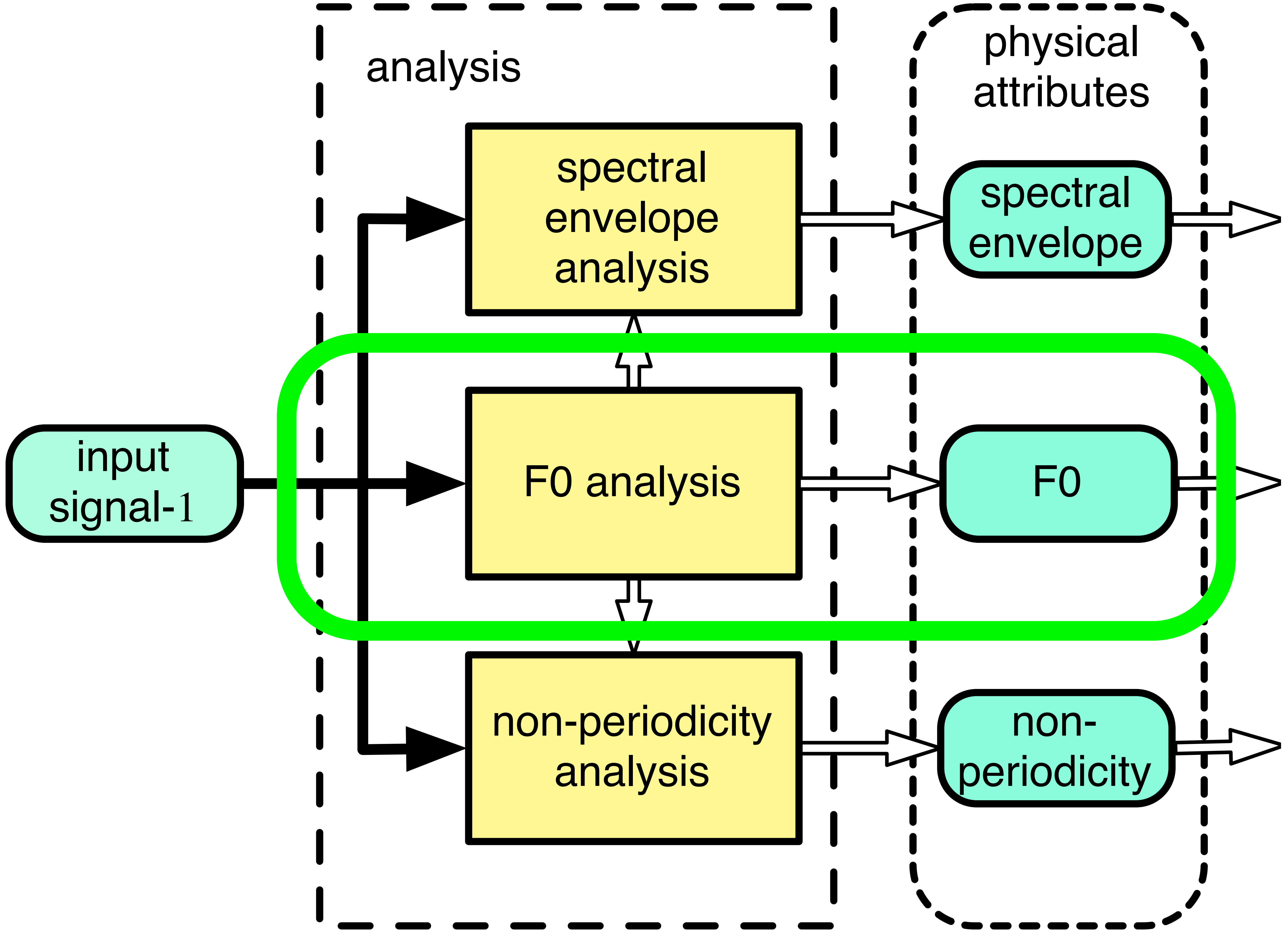


input feature vector

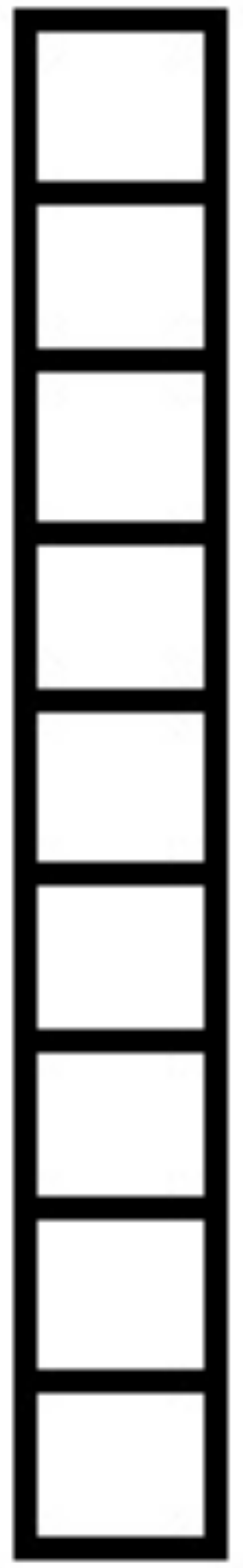
Representation of written form, spoken form, and "everything in-between"



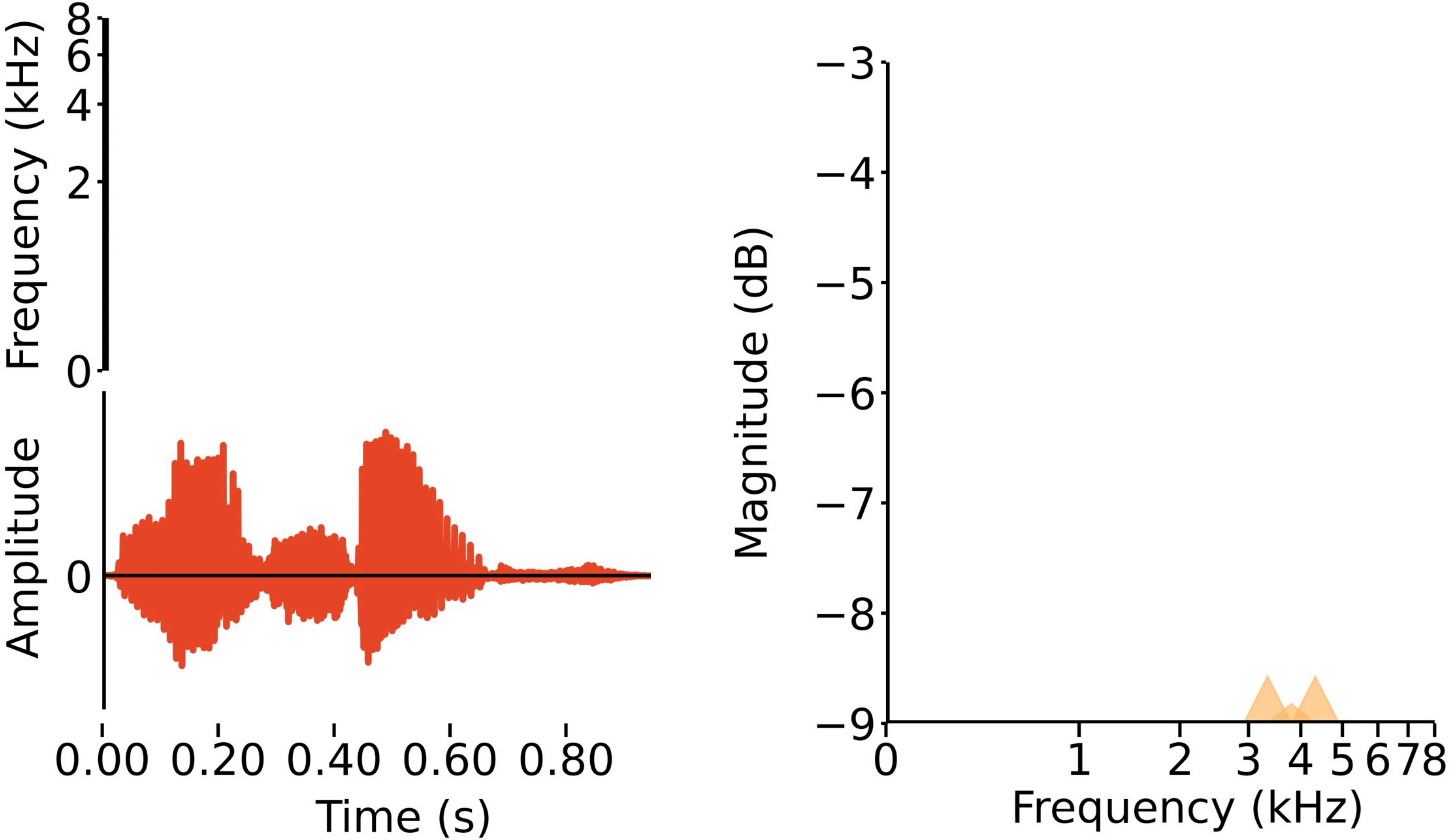
Representation of written form, spoken form, and "everything in-between"



speech parameters

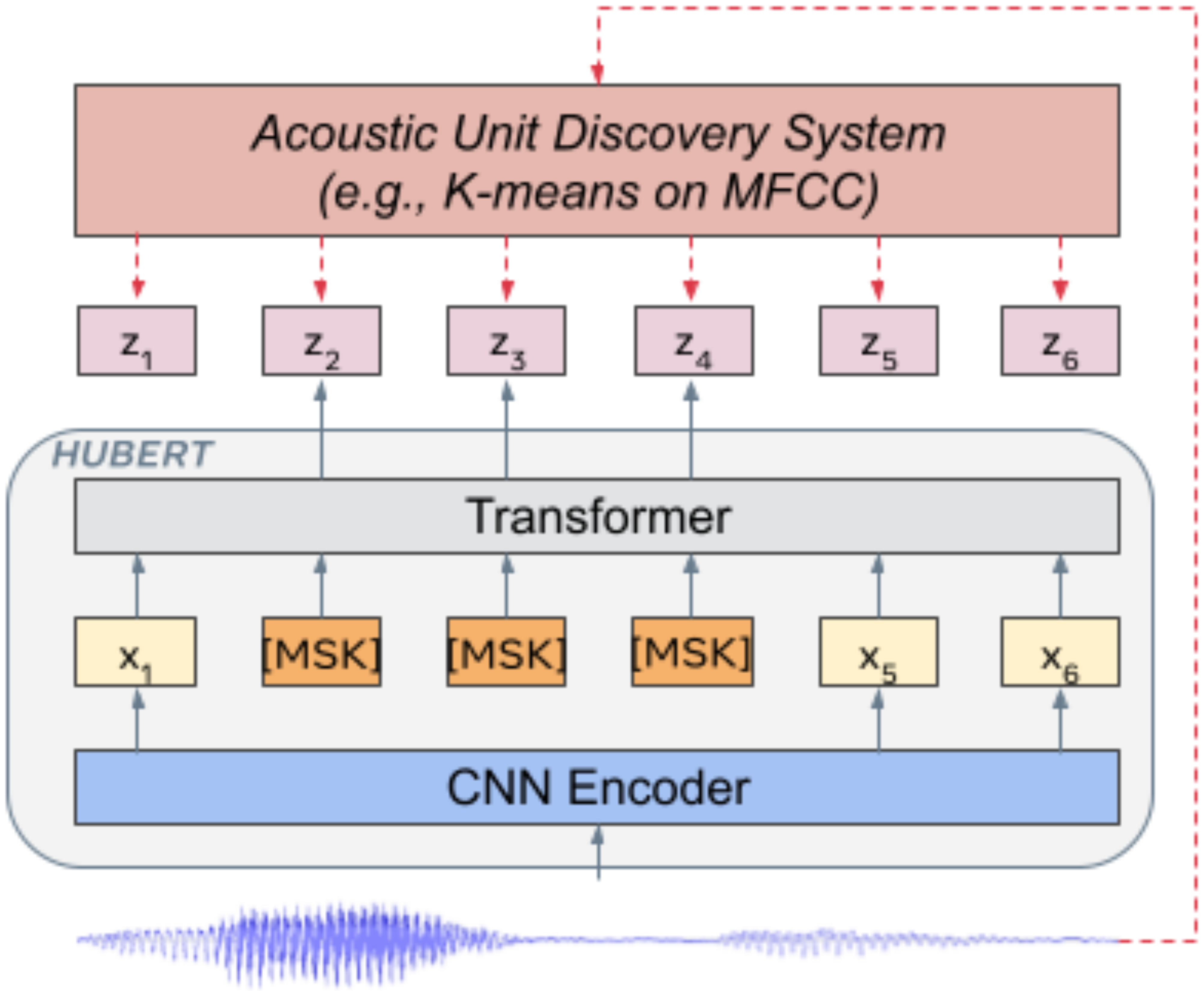


output feature vector



Frequency (kHz)
8
6
4
2
0
Amplitude
0
0.00
0.20
0.40
0.60
0.80
Time (s)
Magnitude (dB)
−3
−4
−5
−6
−7
−8
−9
0
1
2
3
4
5
6
7
8
Frequency (kHz)

Representation of written form, spoken form, and “everything in-between”



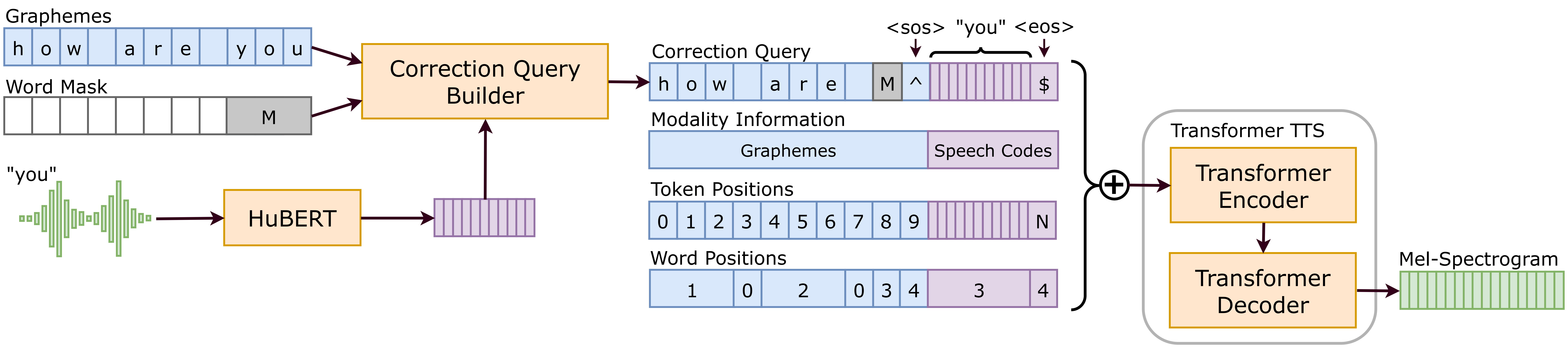
Hsu et al "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451-3460, 2021, doi: 10.1109/TASLP.2021.3122291.

Representation of written form, spoken form, and "everything in-between"

Interspeech 2022
18-22 September 2022, Incheon, Korea

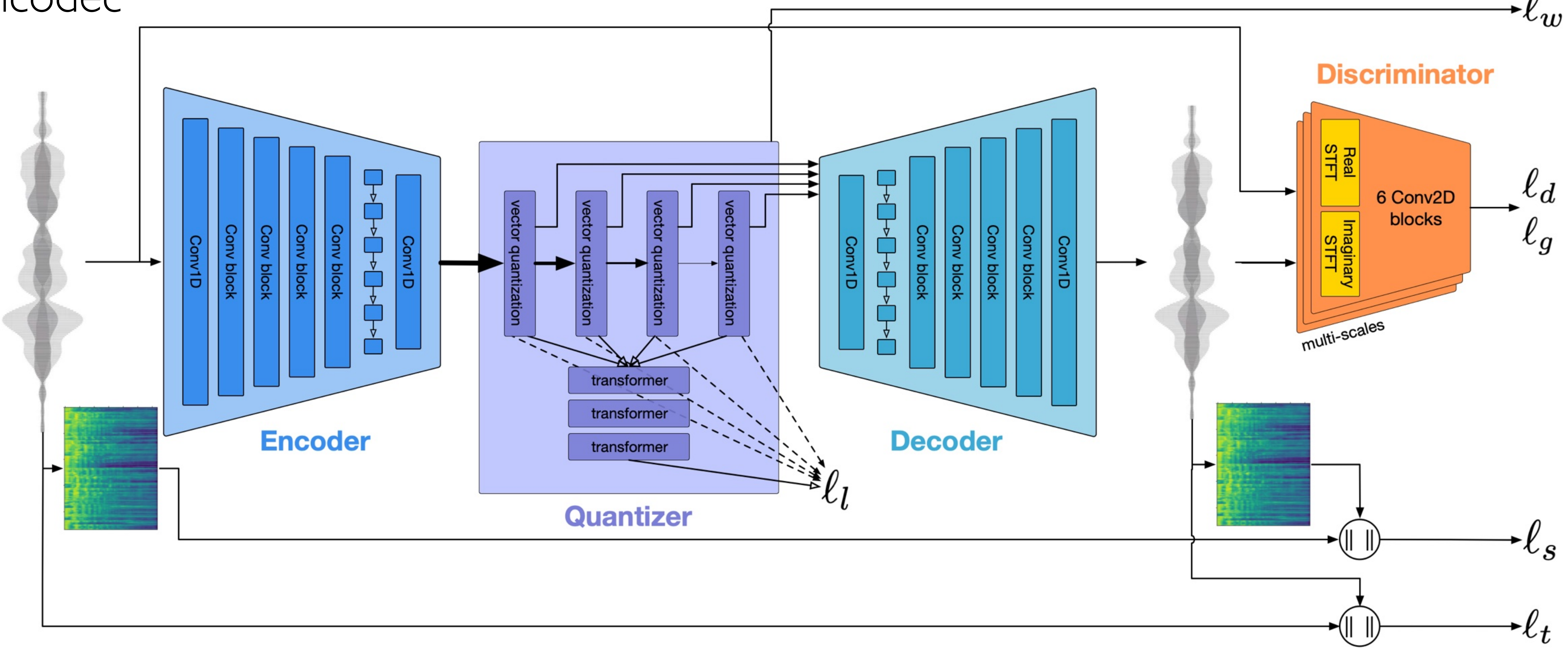
Speech Audio Corrector: using speech from non-target speakers for one-off correction of mispronunciations in grapheme-input text-to-speech

Jason Fong[1], *Daniel Lyth*[1], *Gustav Eje Henter*[2], *Hao Tang*[1], *Simon King*[1]



Representation of written form, spoken form, and "everything in-between"

Encodec



https://github.com/facebookresearch/encodec

Introduction to the coursework - see speech.zone