What we will cover in this class

- Brief recap of video content and Q&A
- Discussion points

Orientation

- Unit selection
  - selection of waveform units based on
    - target cost
    - join cost

Let's just consider the **IFF** type of target cost, which is based only on the **linguistic specification**

- Speech signal modelling
  - generalised source+filter model

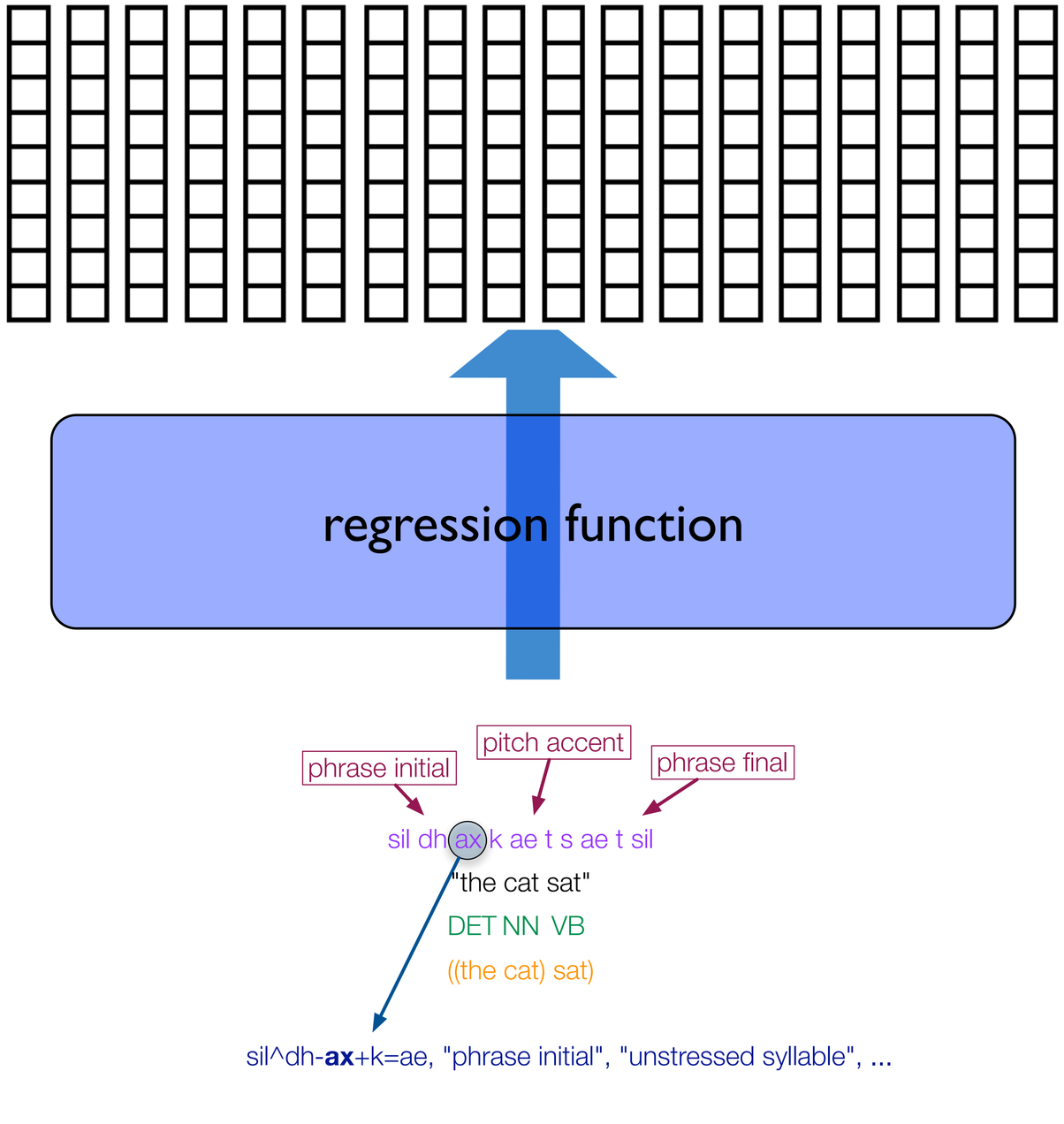
There are several ways to do this, but we need to be able to

- **separate** excitation & spectral envelope
- **reconstruct** the waveform

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

A **regression** task!

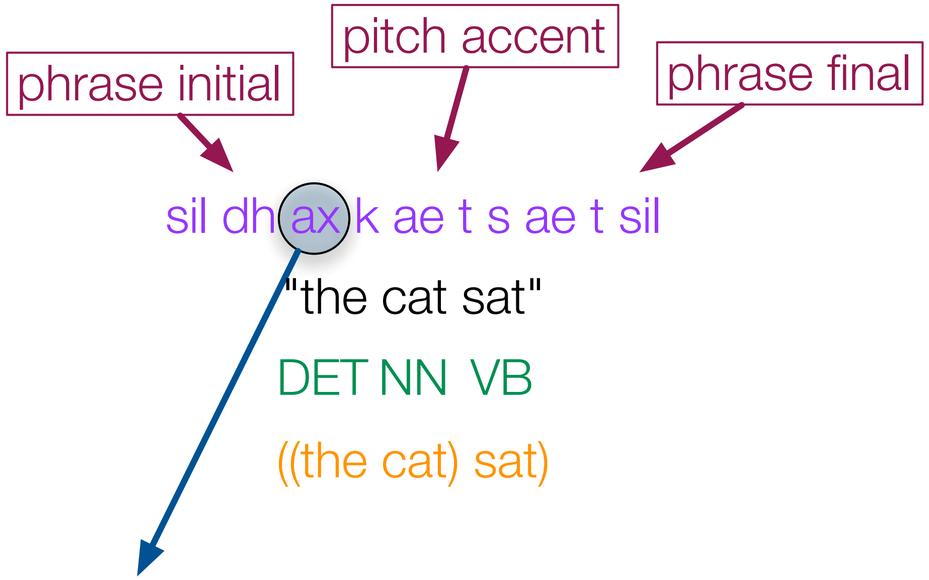
Orientation



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

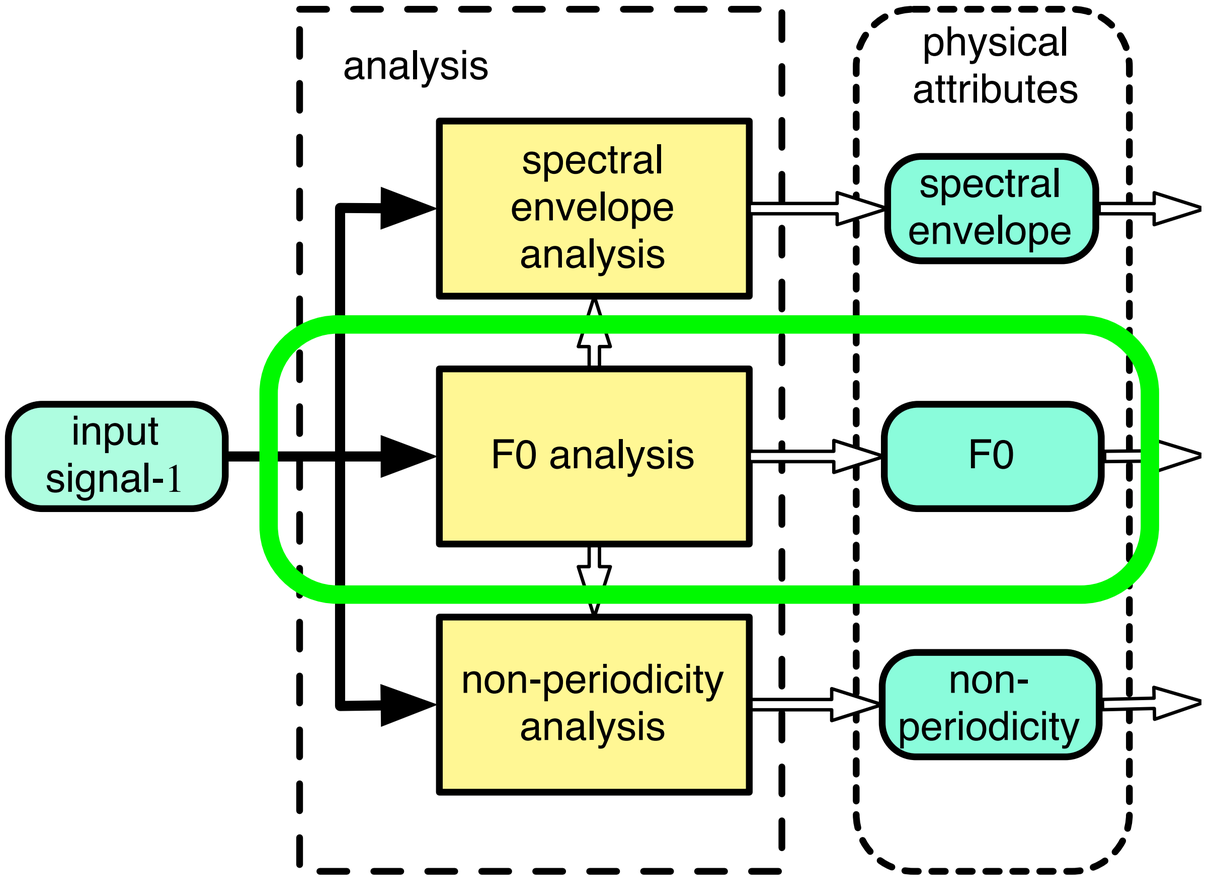
What are the input features ?

Just the linguistic features !



input feature vector

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

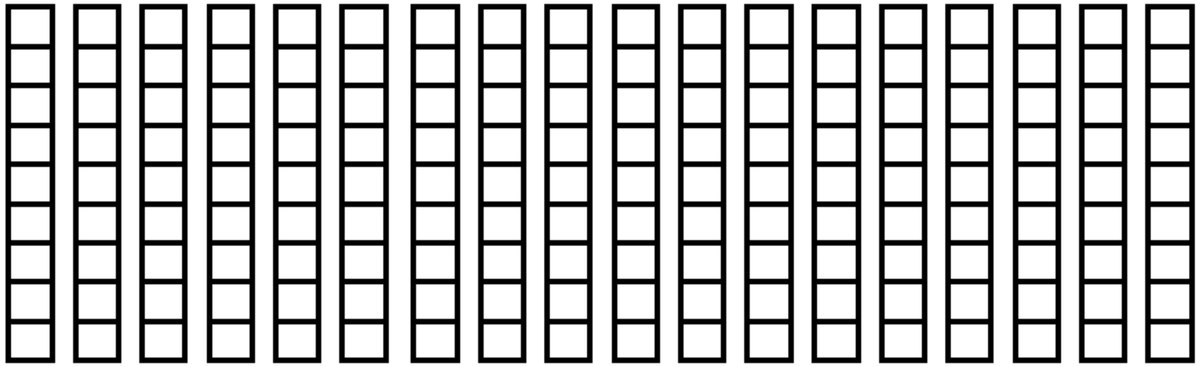


speech parameters

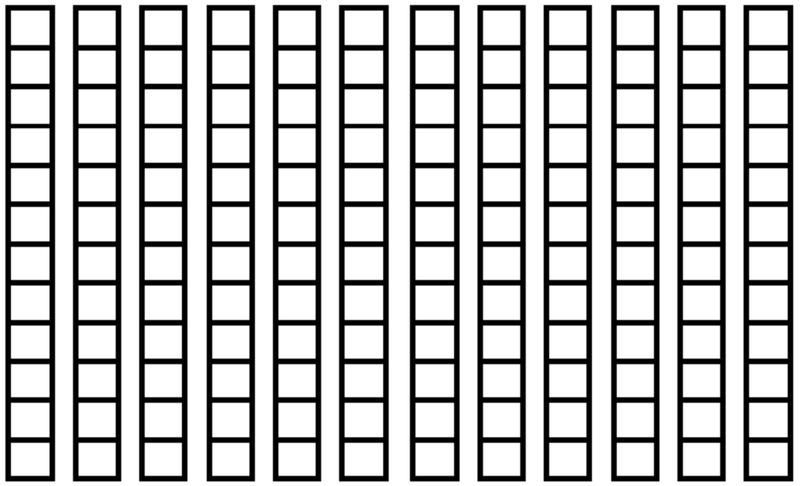
output feature vector

The **sequence-to-sequence** regression problem

output sequence



input sequence



Summary: characterising the speech synthesis problem

- Input = linguistic features (phone identities, neighbours, + other context features)
  - can be long strings for model names
  - can be flattened into a single (sparse) vector
- Output = vocoder parameters (like last week's material)
- Synthesis is then a **sequence to sequence regression problem**, with 2 hard bits:
  - regression from one feature set to the other
  - different "clock" rates of input to output features
- SPSS looks to 2 complementary models to address this (HMM-based; NN-based)

Summary: HMM-based Synthesis

- two views: regression v's context-dependent modelling (which are the same thing really!)
- interpretation via regression view:
    - Sequencing (how long to spend in each part) = HMMs
      (state transition probabilities)
    - Regression (map input features -> output features) = Regression tree
      (to give HMM state parameters (mean/variance/alphas of multivariate Gaussians))
- interpretation via context-dependent modelling:
    - construct a (v. large!) number of model names, based on linguistic features
    - many, many models only seen once or never in training data, so need to have tying…
    - …tying tree *is* the regression tree above (ties/shares HMM state parameters)
    - tying based on linguistic "questions", splitting nodes (model+data) to improve likelihood

Summary: HMM-based Synthesis - generating new speech

- front end linguistic analysis
- flatten that to work out model + state sequence (incl. duration model - HSMM)
- MLPG algorithm to generate frames of speech parameters (MCCs, BAPs, F0)
  - fancy, but basically just smoothing
- vocoder converts those to a speech waveform

What we will cover in this class

- Brief recap of video content and Q&A
- **Discussion points**

Comparison of some unit selection & SPSS synthesis samples

> Mini listening quiz: which is which? (and how to tell?!)

From text to speech with HMMs

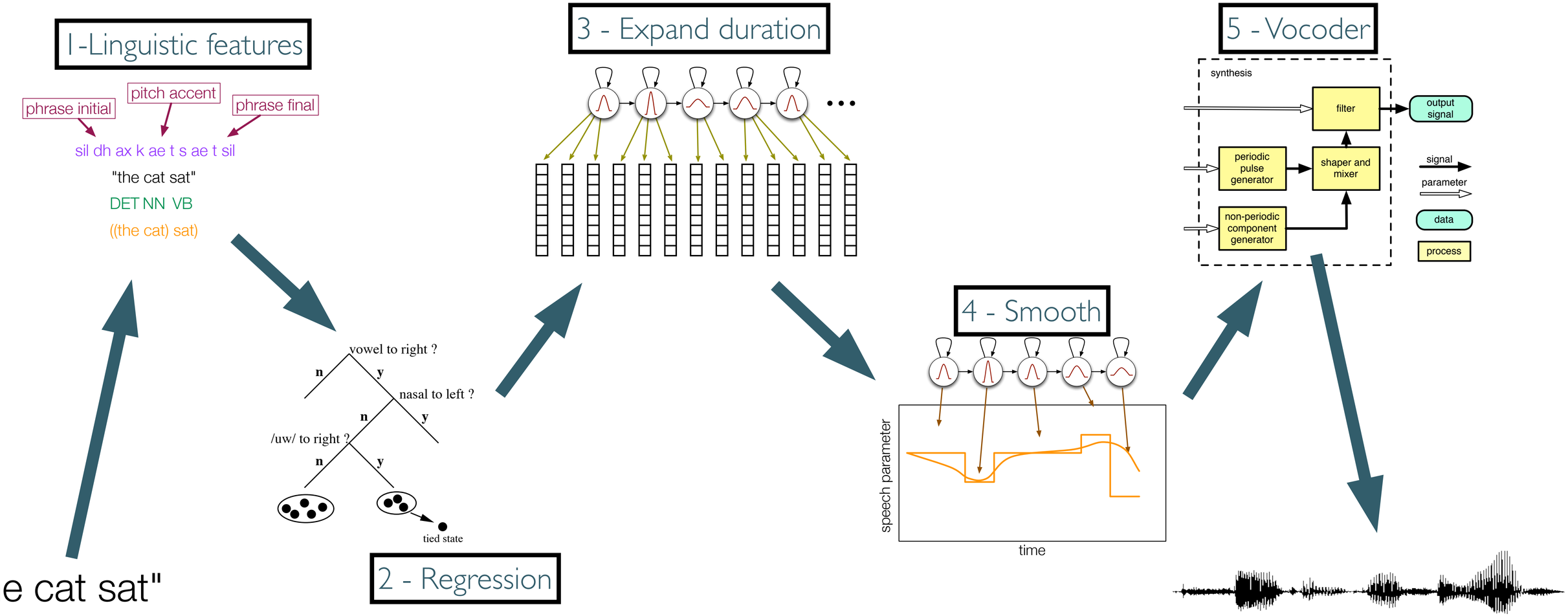
Q: What is the full sequence of steps from text to speech in HMM-based synthesis?

"the cat sat"

???

From text to speech with HMMs

Q: What is the full sequence of steps from text to speech in HMM-based synthesis?



The important role of context in TTS

We've talked a bit about context features, but let's think more about what their role is…

- Q: What would happen if we used no context feature in unit selection? (i.e. only phone identity?)

- Q: And the same for HMM-based synthesis -what if we used few or no context features?

Controllability

Unit selection versus HMM-based synthesis

- Compare and contrast how the following could be realised in each method:

- Q: Make the voice speak faster or slower?
- Q: Make the voice speak in 5 different emotions?
- Q: Make the voice sound like a new person?

Unifying framework - sequence to sequence regression

- Assertion - TTS is at heart a sequence-to-sequence regression problem, and ***all*** TTS methods are an instance of that
- Already covered: SPSS synthesis (HMM-based, DNN-based)
- Will cover: SOTA seq2seq neural models
- What about unit selection?…
  - Q: how does unit selection fit this regression view?

Input **representation**

- representing features as **binary**
  - can this be done for **any** feature at all?
  - does this place any limitation on performance?

- **how and why** might you encode the following linguistic structures
  - place & manner of articulation
  - position of phone in syllable ; position of syllable in word ; position of word in phrase

- **upsample** all features to the acoustic **frame rate**
    - is this reasonable?

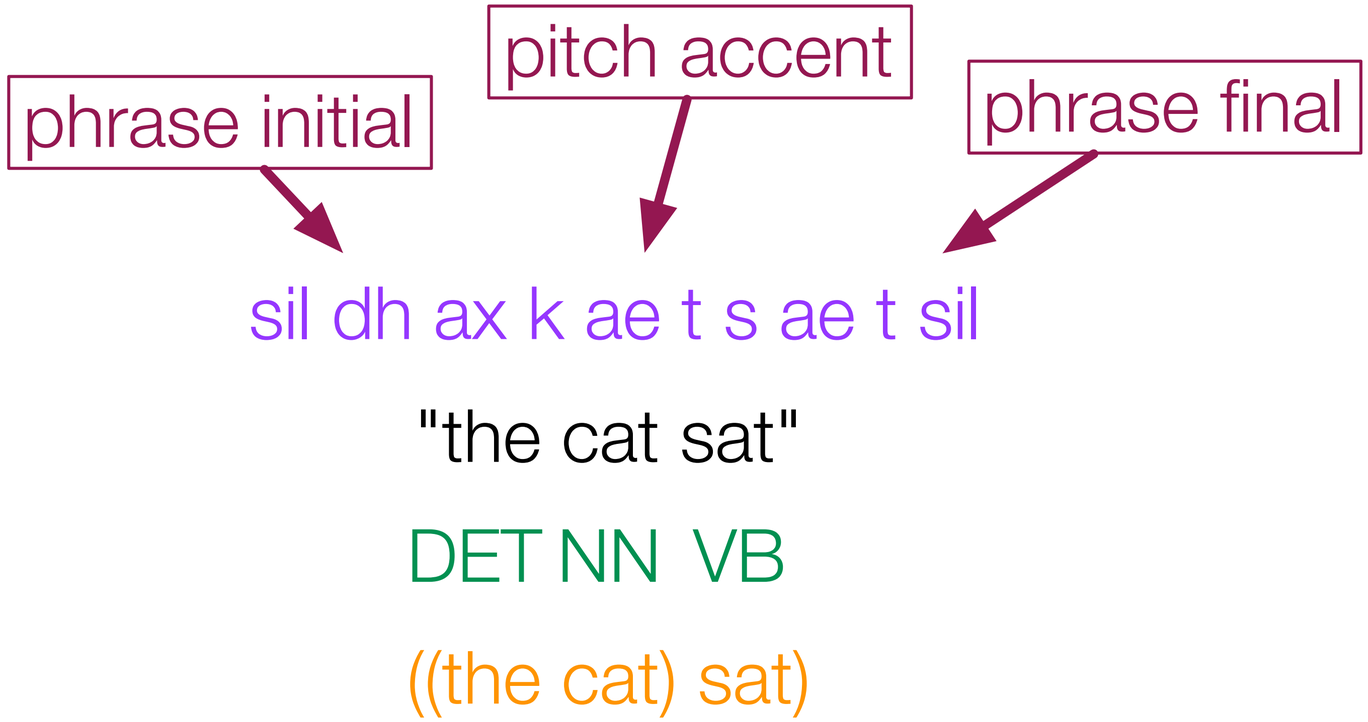
Exercise: a decision tree effectively treats the input features as "one hot"

- Draw a very simple decision tree that predicts the speech parameters for a phone
  - ignore duration for now - assume each phone has a duration of 1 frame
- Describe step-by-step how that can be used to predict a **sequence** of speech parameters
  - what are the **predictors** and what is the **predictee** ?

- List possible questions that could be asked in your decision tree
- Use your questions to rewrite the phone sequence as a sequence of one-hot vectors

- Draw a new decision tree that uses these vectors as the predictor

Exercise: a decision tree effectively treats the input features as "one hot"



What next?

- **Better regression model**
  - a Neural Network
  - input & output features essentially the same as regression tree + HMM

- Quality will still be limited by the **vocoder**

- Later, we will also address that problem
  - hybrid synthesis
  - direct waveform generation