Unit selection - target cost

- discussion points

Class plan

- **Recap** of what you should know at this point
  - Q&A (*you ask, I answer*) on video content for module 3 (and module 2 if you wish)
  - Discussion (*I ask, you answer*) about IFF and ASF target cost functions
- **Readings**
  - Paper discussion: **Hunt & Black**
  - ~~Reading Q&A: **Taylor Chapter 16**~~
- Additional optional points (time permitting) for **discussion**

- A look forward to **neural approaches** and how they relate to unit selection
  - many design choices are the same (because we are still doing TTS)

Discussion about IFF and ASF target cost functions

- What linguistic features would you use in an IFF target cost?

- What acoustic features would you use in an ASF target cost?
- What predictive model might be good for predicting target acoustic features for an ASF?
  - what timescale would you make predictions on?
  - how would you compare a target's predicted features to a candidate's features?

- How would you combine IFF and ASF?
  - why would you want to do that?
  - any potential pitfalls?

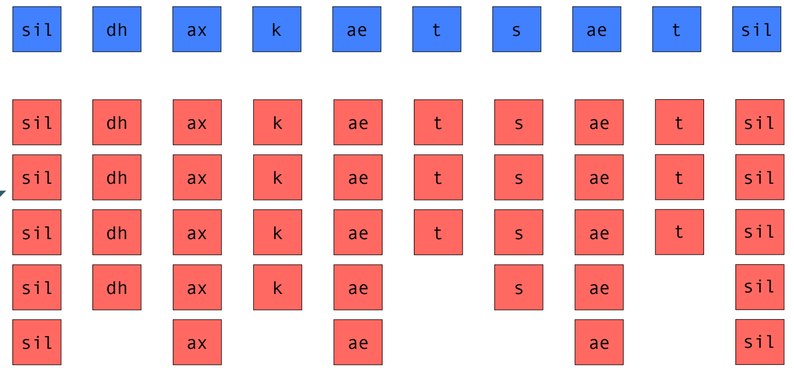
Readings

- **Hunt & Black paper**
  - structured reading and discussion

- ~~Taylor Chapter 16~~
  - ~~unstructured Q&A~~

Hunt & Black

- What size is the unit?
- What style of target cost is used?
  - What are the features used? How many?
- Re-draw Figure 2 so that it looks like this



- Relate Figure 1 to your new version of Figure 2
- "*The important distinction is that Markov models are probabilistic, whereas the current work uses cost functions*" - is this really important?
- How many types of pruning are applied, and what are they?
- Where in the paper is the "zero join cost trick" ?
- What evaluation was conducted?

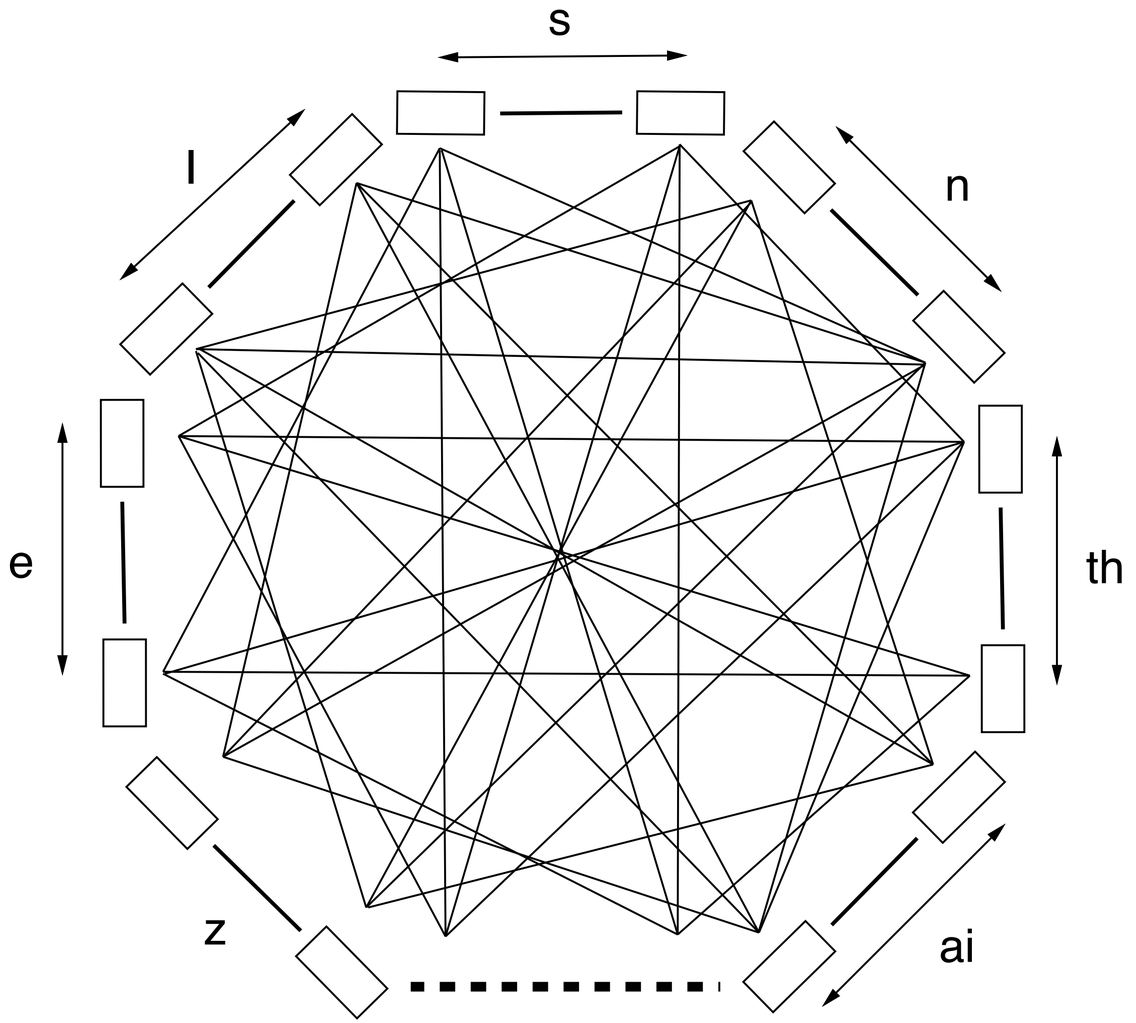
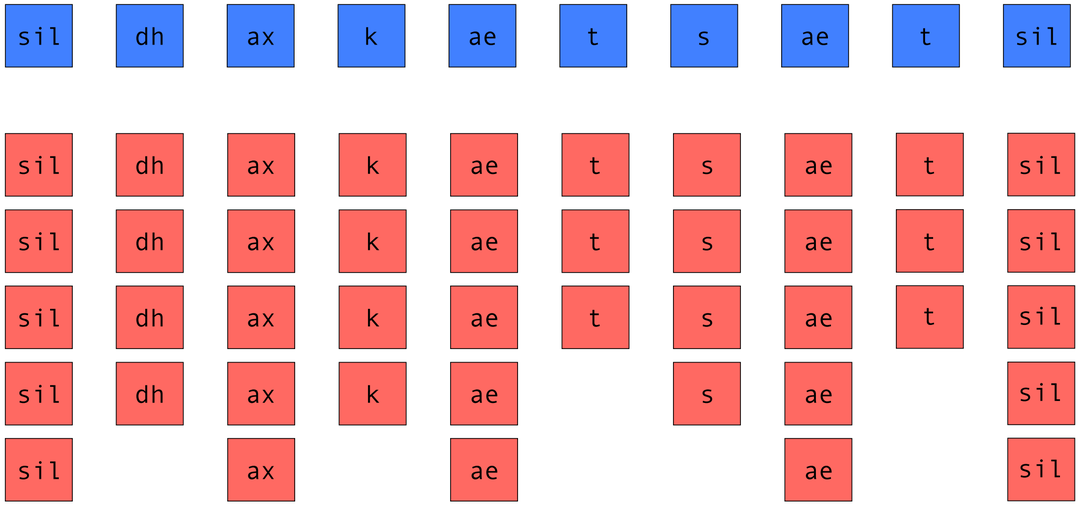


Figure 2. Phoneme Network for a Database

a join between consecutive units (u_{i-1} and u_i). Section 2.1 describes how the database units can be treated as a *state transition network* which is decoded by these costs. Section 2.2 descibes how the costs are calculated and Section 3 describes the training of the costs.



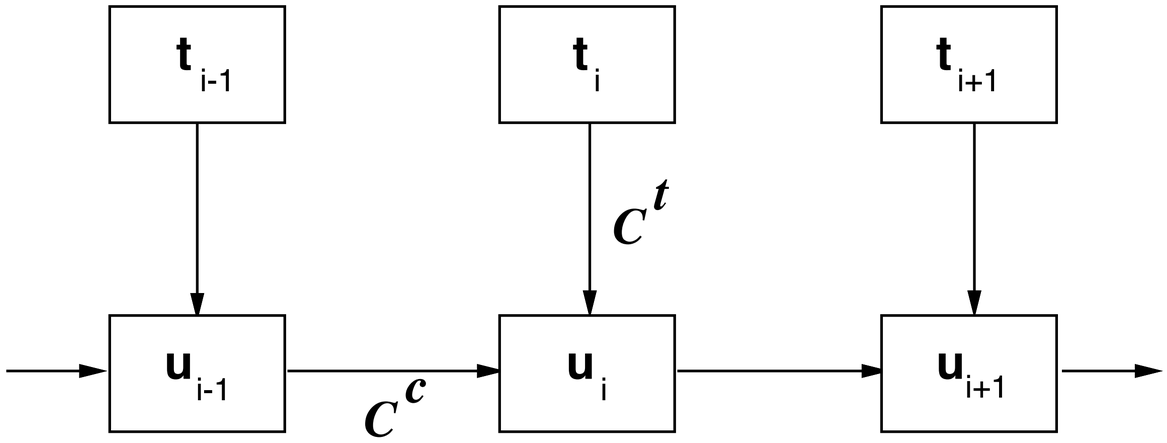
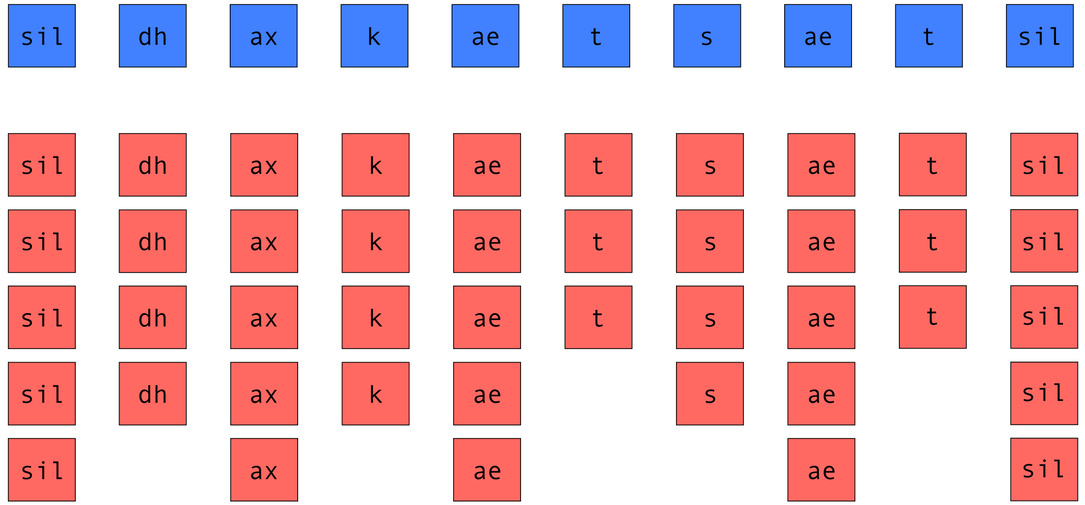


Figure 1. Unit Selection Costs



2. UNIT SELECTION

The input to CHATR is typically text, though this may be augmented with structural and discourse information. The first stages of synthesis transform this input into a *target specification* (or simply *target*). The target for an utterance defines the string of phonemes required to synthesize the text, and is annotated with prosodic features (pitch, duration and power) which specify the desired speech output in more detail. This paper is not concerned with the procedures required to produce the target specification, but instead focuses on the selection of appropriate units from a database to synthesize the target.

The target cost is calculated as the weighted sum of the differences between the elements of the target and candidate feature vectors: these differences are the p *target sub-costs*, $C_j^t(t_i, u_i)$ $(j = 1, ..., p)$. In the current implementations p varies between 20 and 30. The target cost, given weights w_j^t for the sub-costs, is calculated as follows:

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) \qquad [1]$$

The *concatenation cost*, $C^c(u_{i-1}, u_i)$, is also determined by the weighted sum of q *concatenation sub-costs*, $C_j^c(u_{i-1}, u_i)$ $(j = 1, ..., q)$. The sub-costs can be determined from the unit characterisations of u_{i-1} and u_i (as with the target cost), but may additionally be derived from signal processing of the units. Three sub-costs were used in the current work (i.e. $q = 3$): cepstral distance at the point of concatenation and the absolute differences in log power and pitch. The concatenation cost, given weights w_j^c, is

Optimal unit selection can be performed with a Viterbi search. However, to obtain near real-time synthesis on large speech databases, containing tens of thousands of units, the search space must be pruned. This has been implemented by multiple pruning steps. Initially, units with phonetic contexts similar to the target are identified. Next, the remaining units are pruned with the target cost and finally with the concatenation cost [5]. With a beam width of 10-20 units, the search can be performed in near real-time on a database with around 100,000 units (on a Sun SPARC-Station 20). Synthesis is faster than real time for smaller databases (less than 50,000 units). Pruning appears to have little effect on the output quality.

Readings

- Hunt & Black paper
  - structured reading and discussion

- ~~**Taylor Chapter 16**~~
  - ~~unstructured Q&A~~

How do cost functions look **forwards** *and* **backwards**?

- we know that various **connected speech processes** operate in both directions
  - anticipatory (depends on next sound) *vs.*
  - perseverative (depends on previous sound)
  - e.g., ***assimilation*** in the word "handbag"

- so, the selection of units needs to take this into account

- **is this achieved** in unit selection synthesis?
  - if so, how?

Prosody generation in unit selection

- Recall Taylor's two choices for the target function
  - independent feature formulation (IFF)
    - just compare **linguistic features** between target and candidates
    - the key question is: ***what linguistic features*** ?
  - acoustic-space formulation (ASF)
    - perform partial synthesis (e.g., F0 value prediction)
    - this provides us with acoustic features for the target
    - then compare **acoustic features** between target and candidates
    - the key questions are:
      - ***how to predict*** the acoustic features for the target?
      - ***how to compare*** them with the candidates?

Prosody generation in unit selection: IFF approach

- the key question is: ***what linguistic features*** should the target cost compare?

- well - they can be anything we can reliably predict from the text

- should that include **ToBI accents & boundary** tones, for example?
  - how would we predict these?
    - choose your classifier: ……
    - list available predictors: ……
    - obtain training data: ……
  - how accurate would those predictions be?

Prosody generation in unit selection: ASF approach

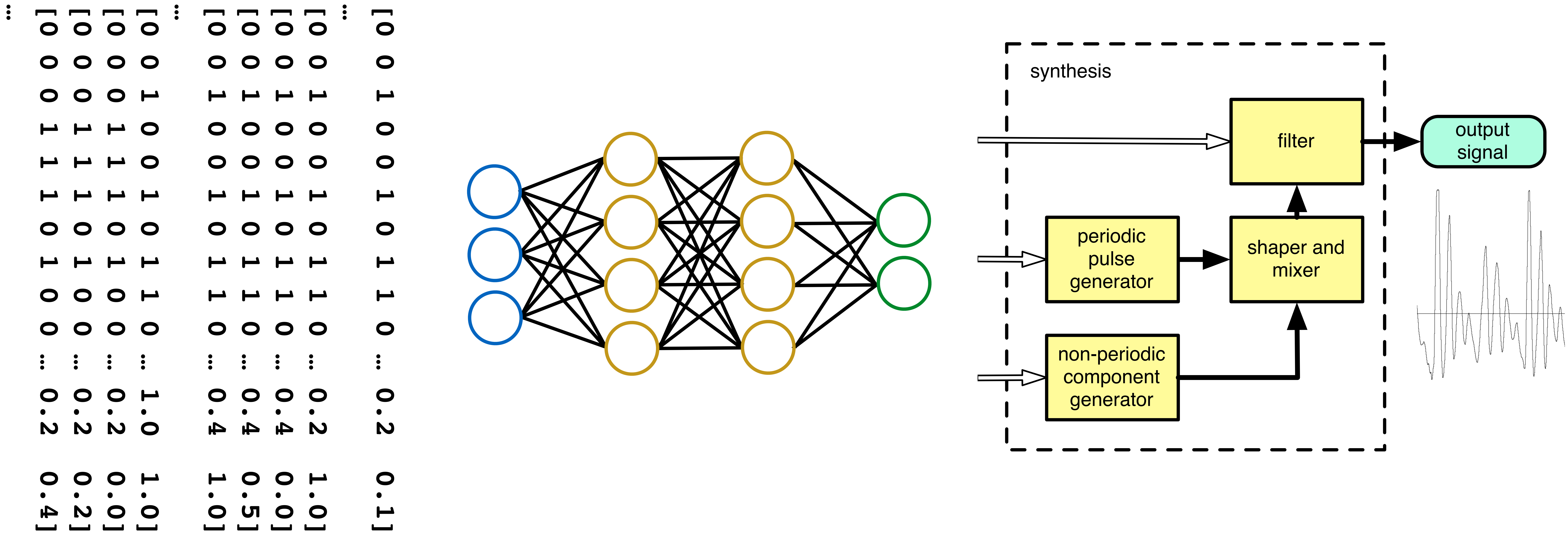
- ***how to predict*** the acoustic features for the target?
  - assume we will use ToBI as the symbolic representation of prosody
  - step 1: predict ToBI symbols from text
    - a classification task, as in the IFF approach
  - step 2: render ToBI symbols as an F0 contour
    - a regression task - will need training on data

- ***how to compare*** the acoustic features between target and candidate?
  - Euclidean distance between F0 contours?
  - is that perceptually relevant?

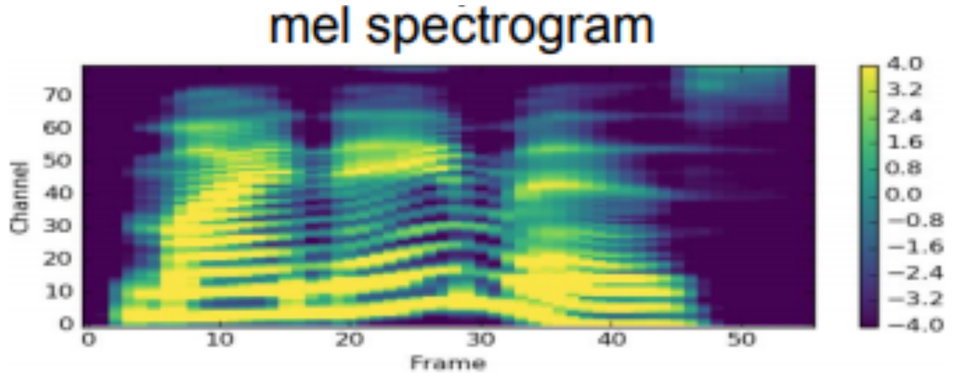
A look forward to neural approaches

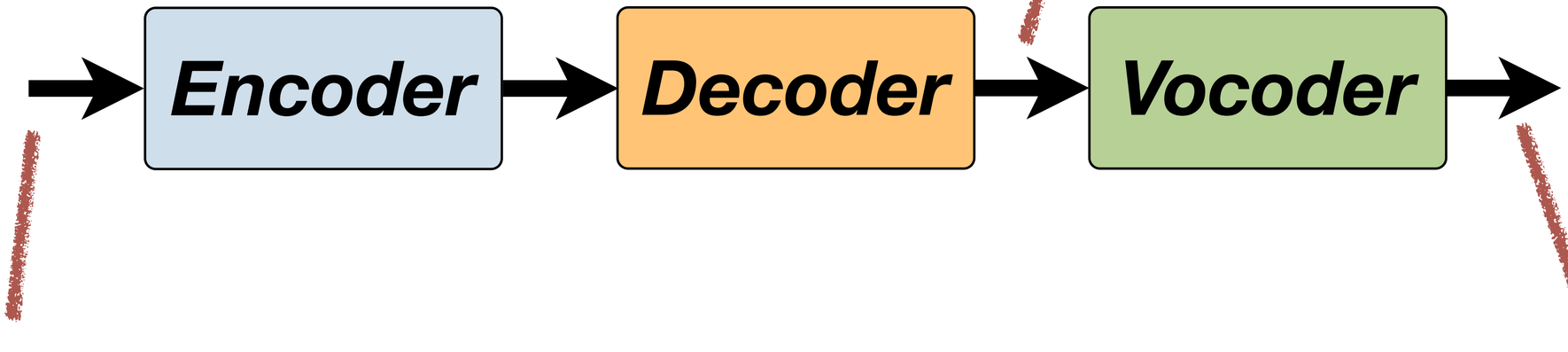
- finding the connections to unit selection

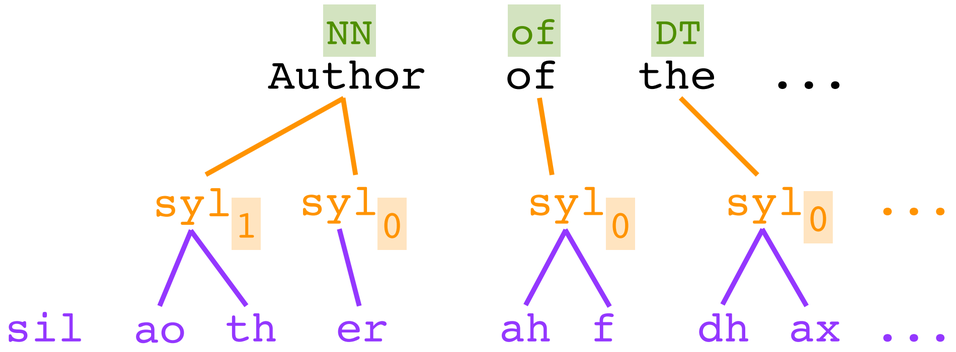
Module 8 - Deep Neural Networks (DNNs)

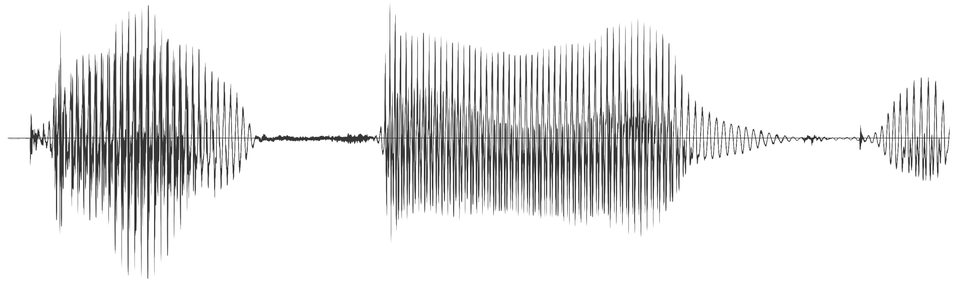


Module 9 - sequence-to-sequence models









The state of the art - e.g., FastPitch