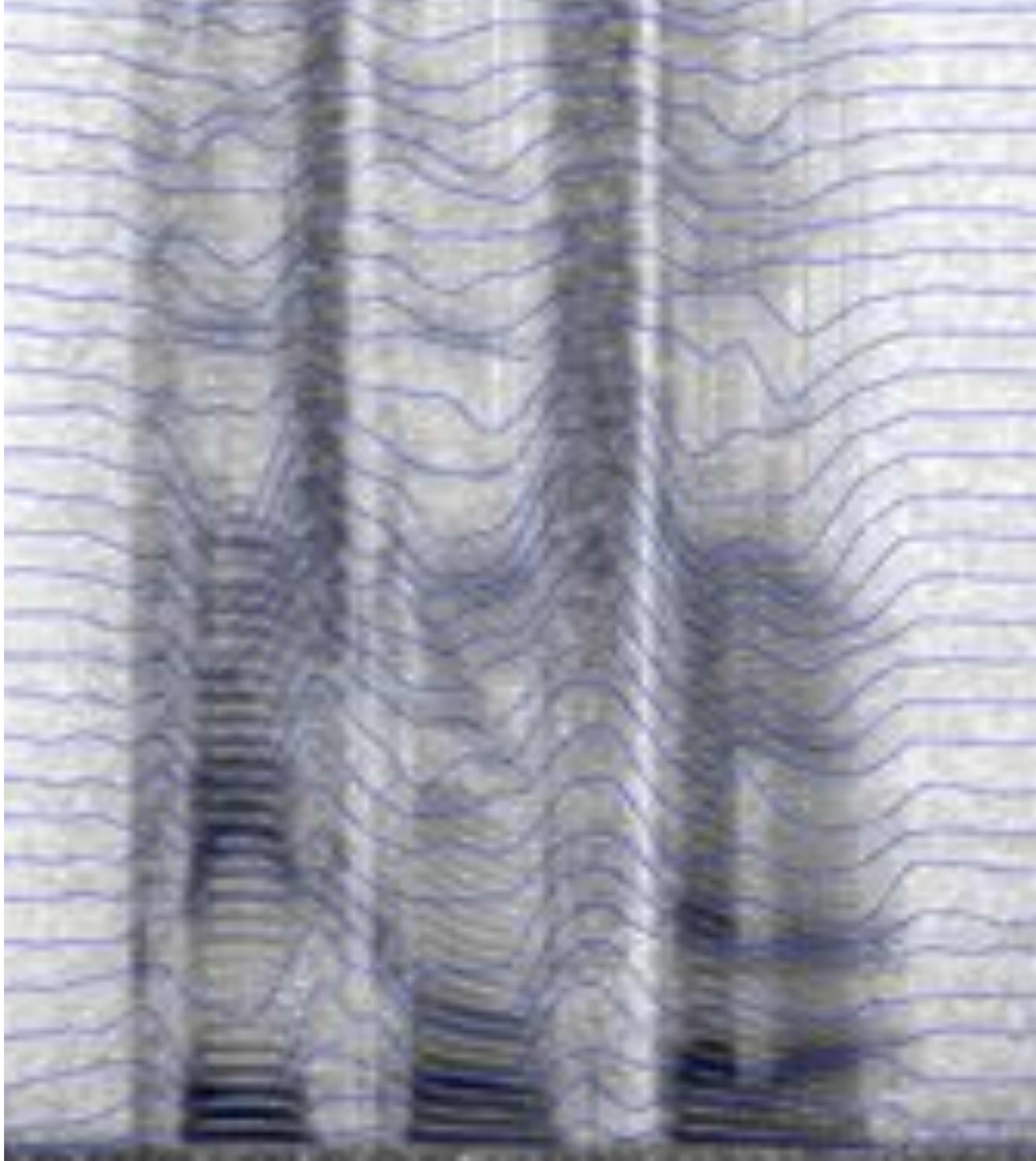
Speech Synthesis

Simon King
University of Edinburgh



Unit selection

Independent Feature Formulation (IFF) target cost function

What you should already know

- selecting waveform fragments from a database of natural speech
- target cost
- join cost
- search

What you should already know

- selecting waveform fragments from a database of natural speech
- target cost
- join cost
- search

the target cost
measures **mismatch**
between

a target unit
and
a candidate unit

A target cost function based only on linguistic features

The independent feature formulation (IFF)

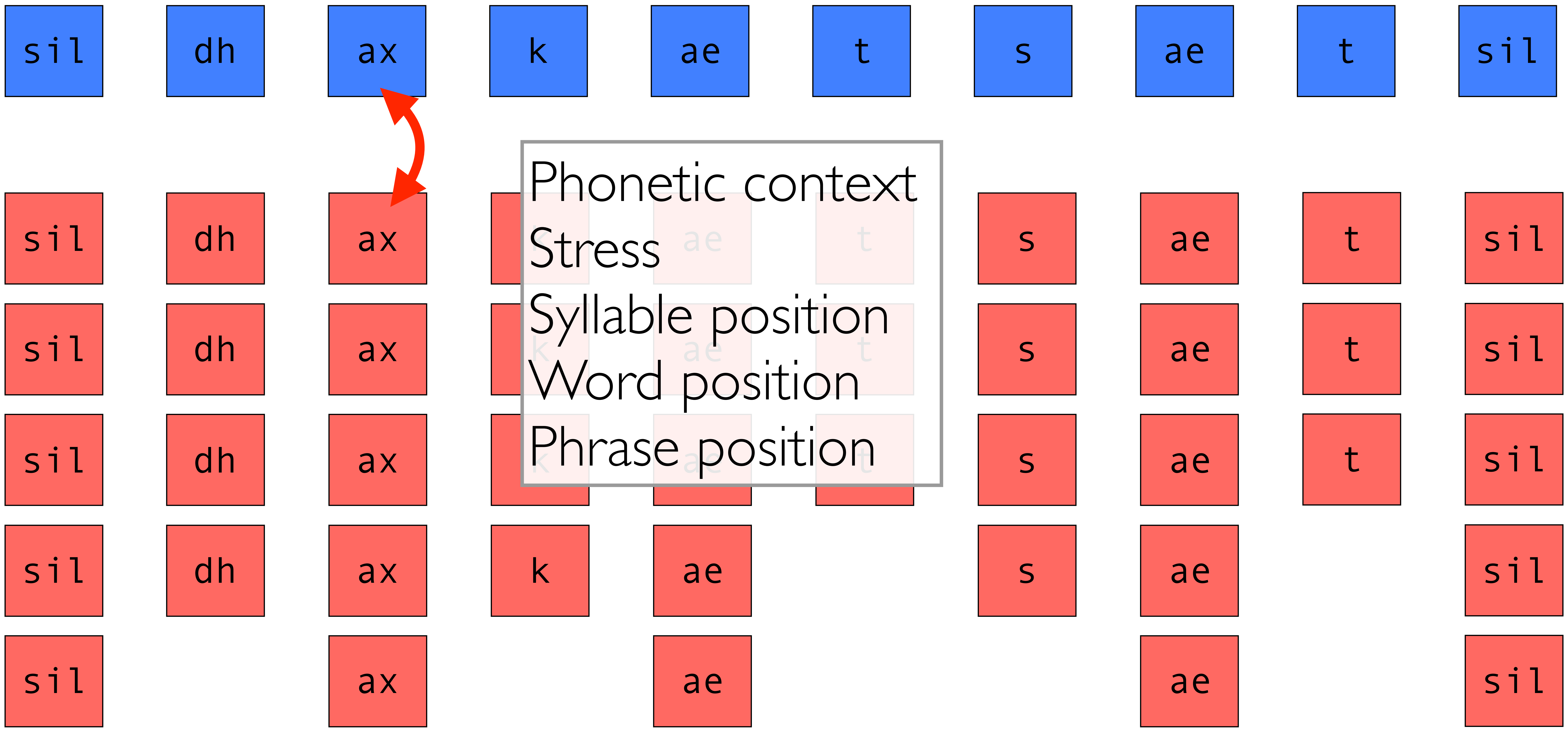
- Let's start with the simplest form of target cost function

- It will simply **count** the number of **linguistic features** in the context of the candidate that **do not match** those of the corresponding target unit

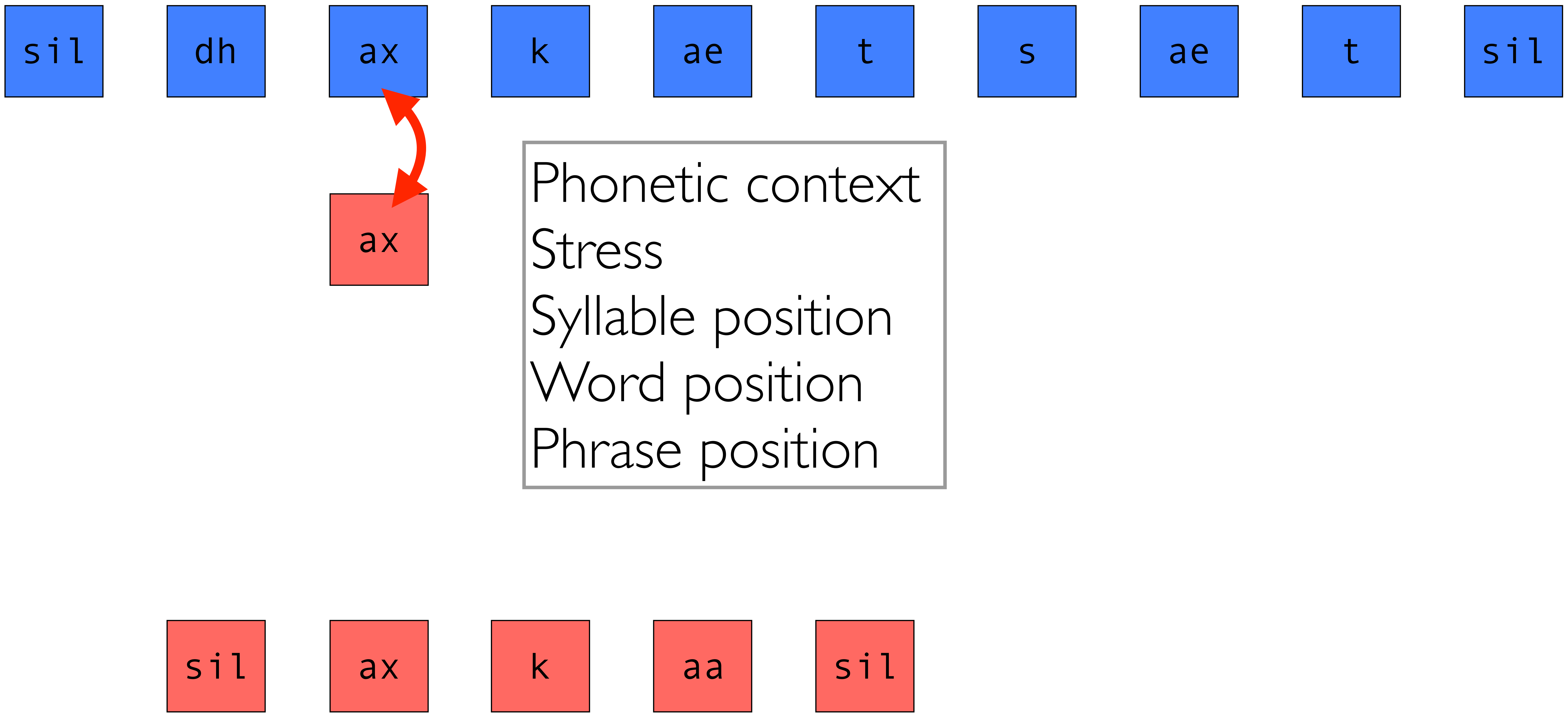
- Motivation is simple
  - An exactly-matching candidate will have a cost of zero (= no mismatch)
  - The more mismatched the context is between candidate and target, the higher the cost

- The cost is a prediction of 'how bad' the candidate would sound, if used here

The IFF target cost function



The IFF target cost function



In the database, we have a recording of the sentence "A car."

Festival's *multisyn* IFF target cost

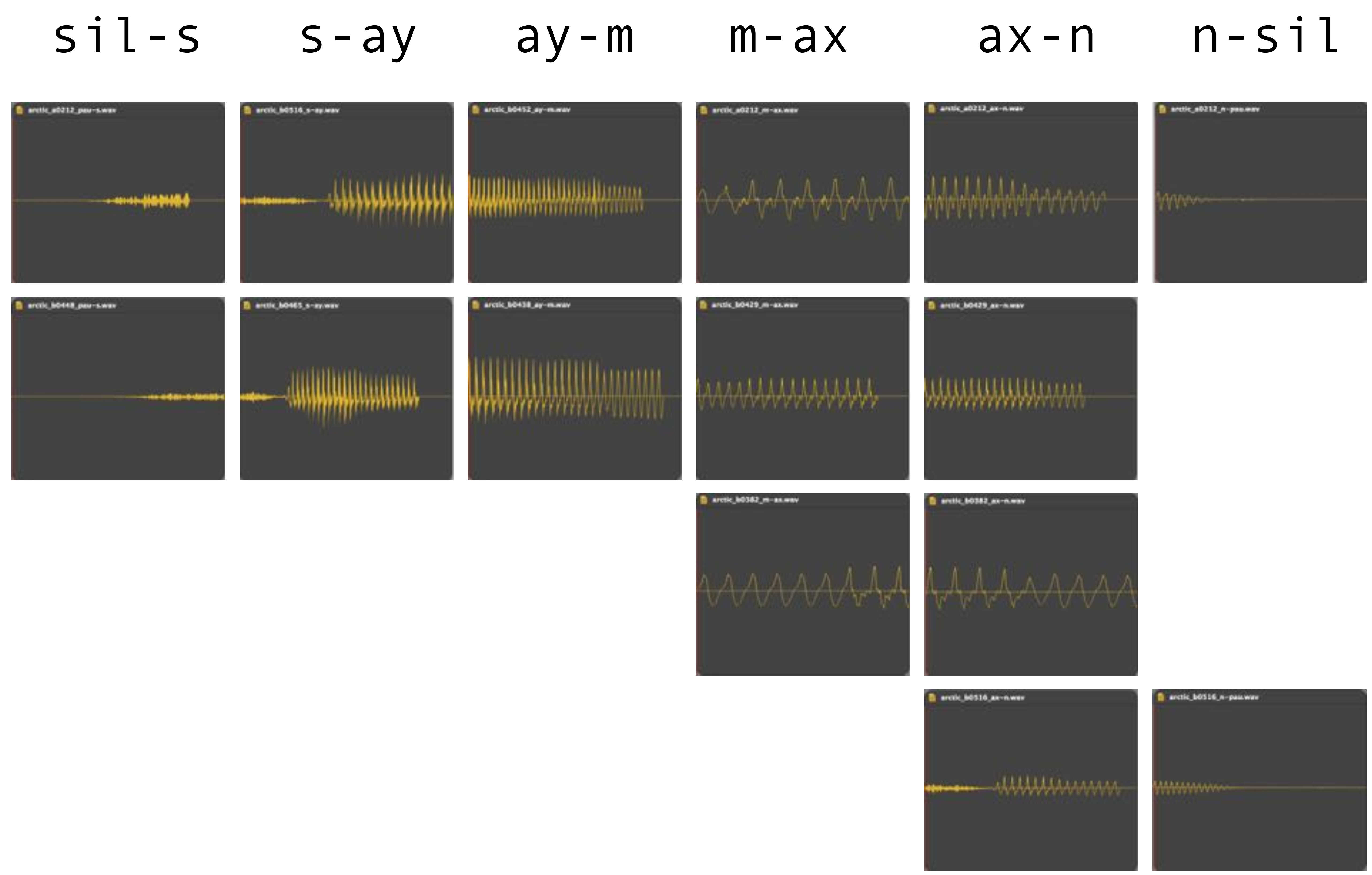
feature	weight
stress	10
syllable position	5
word position	5
POS	6
phrase position	7
left phonetic context	4
right phonetic context	3
bad F0	25
duration outlier	10

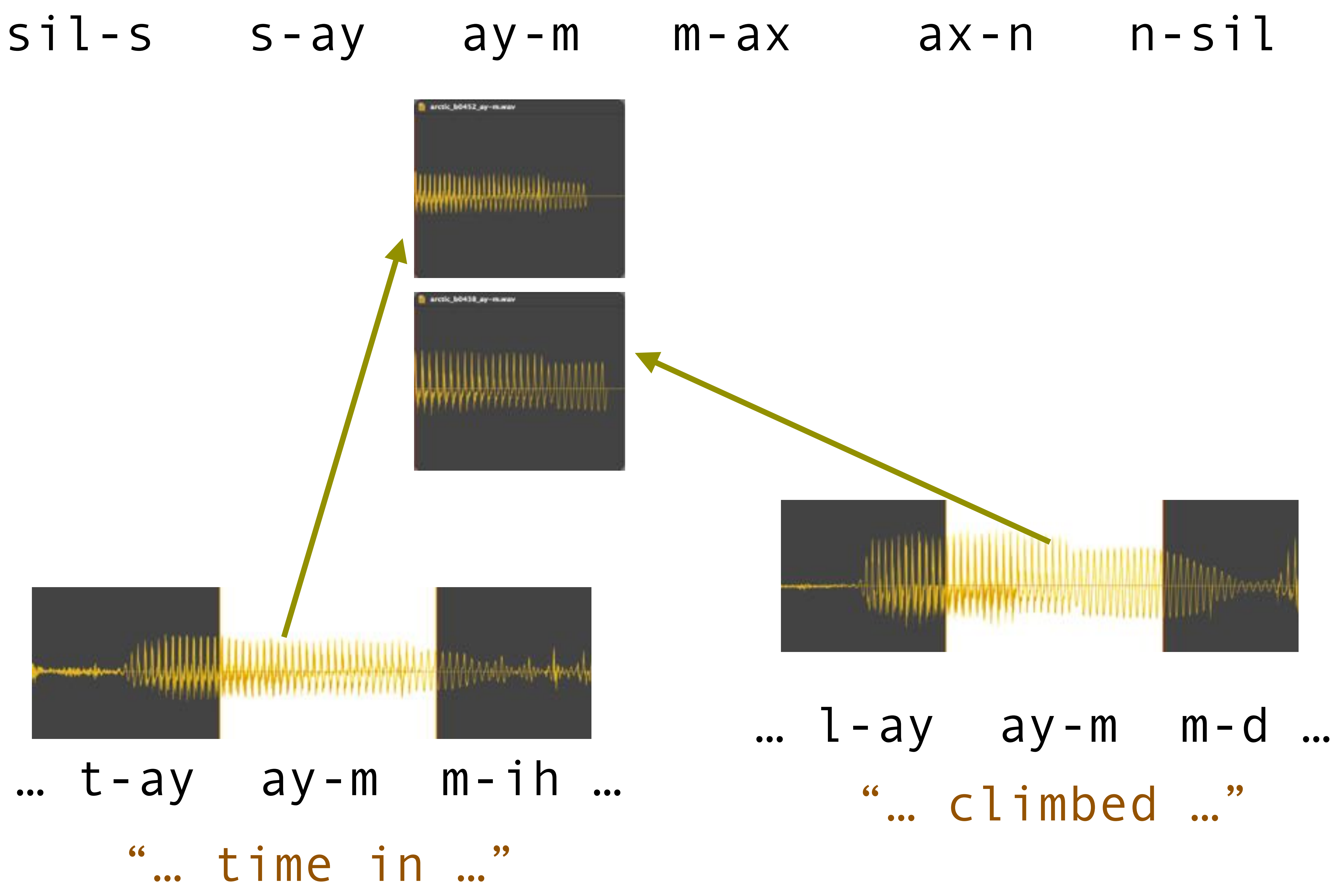
Example calculation of IFF target cost for two competing candidates

feature	weight	*target*	candidate 1	candidate 2
stress	10	*primary*	primary	none
syllable position	5	*coda*	onset	coda
word position	5	*final*	final	final
POS	6	*noun*	noun	verb
phrase position	7	*initial*	*initial*	*initial*
left context	4	*[b]*	[b]	[v]
right context	3	*[s]*	[w]	[s]
		target cost =		

Another example, this time for **diphone** units

"Simon"




sil-s s-ay ay-m m-ax ax-n n-sil
… t-ay ay-m m-ih …
“… time in …”
… l-ay ay-m m-d …
“… climbed …”

Wait … how is prosody "created" using an IFF target cost function ?

- With **no** explicit predictions of **any** acoustic properties, this is a reasonable question
- Answer:
  - candidates from appropriate contexts, when selected, will have appropriate prosody
  - the join cost will ensure that F0 is continuous

- So, we simply need to make sure the **linguistic features** capture sufficient contextual information that is relevant to prosody
  - e.g., stress status, position in phrase
- *Optional*: if our front end predicts **symbolic prosodic features** (e.g., ToBI accents and boundary tones), then we can use them in the target cost function

Unit selection

Acoustic Space Formulation (ASF) target cost function

Orientation

- Unit selection as we understand it so far
  - run text processor (front end)
  - construct target sequence
  - retrieve candidates from database
    - compute IFF target costs
    - compute join costs
    - perform search

- Now, a more sophisticated target cost
  - predict **acoustic properties** of target units
  - compare these with actual acoustic properties of candidates

Orientation

- Unit selection as we understand it so far
  - run text processor (front end)
  - construct target sequence
  - retrieve candidates from database
    - compute IFF target costs
    - compute join costs
    - perform search

by comparing linguistic features

weakness: it is possible for two units with differing (mismatched) features to **sound very similar**

- Now, a more sophisticated target cost
  - predict **acoustic properties** of target units
  - compare these with actual acoustic properties of candidates

solution: compare how units sound

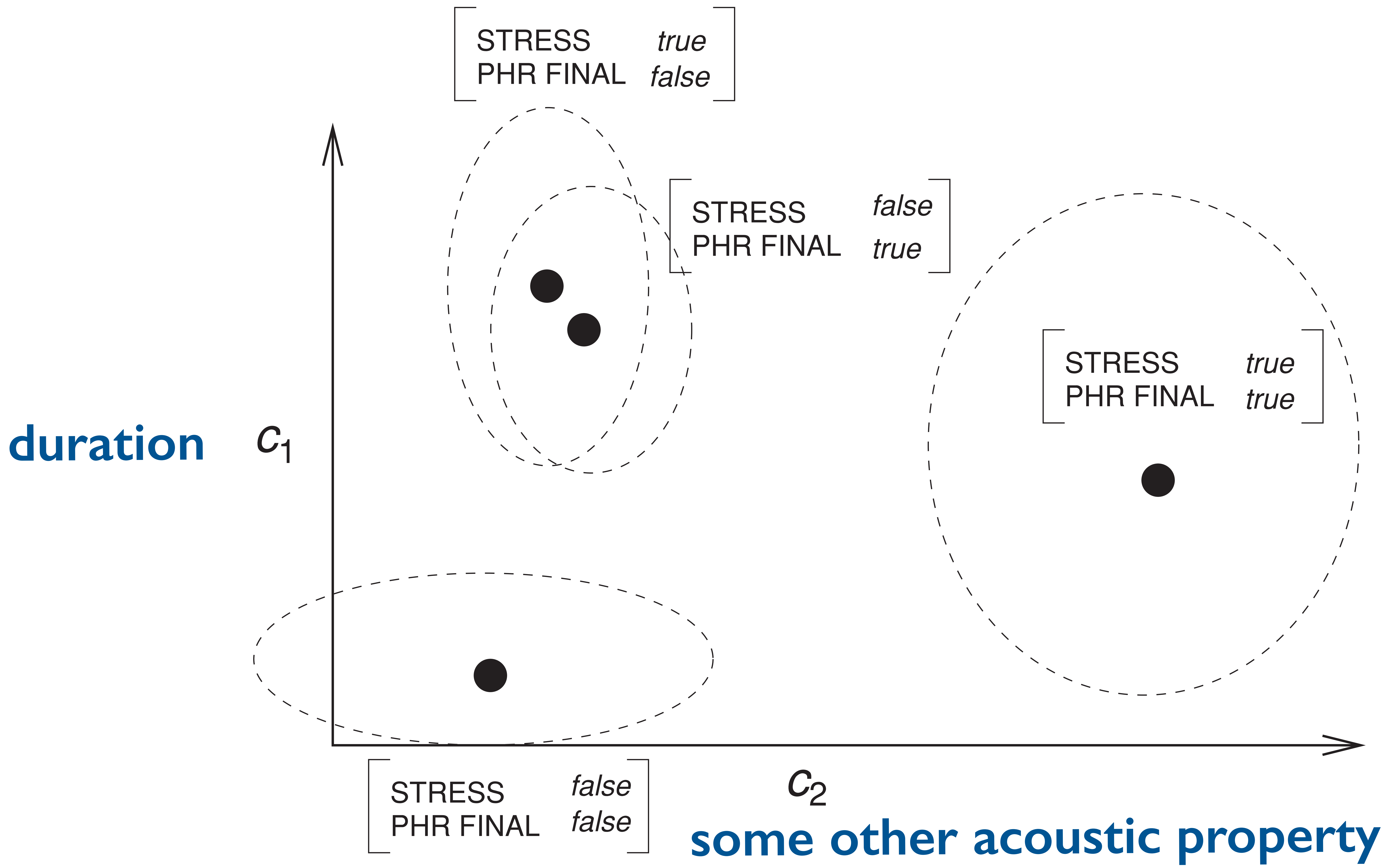
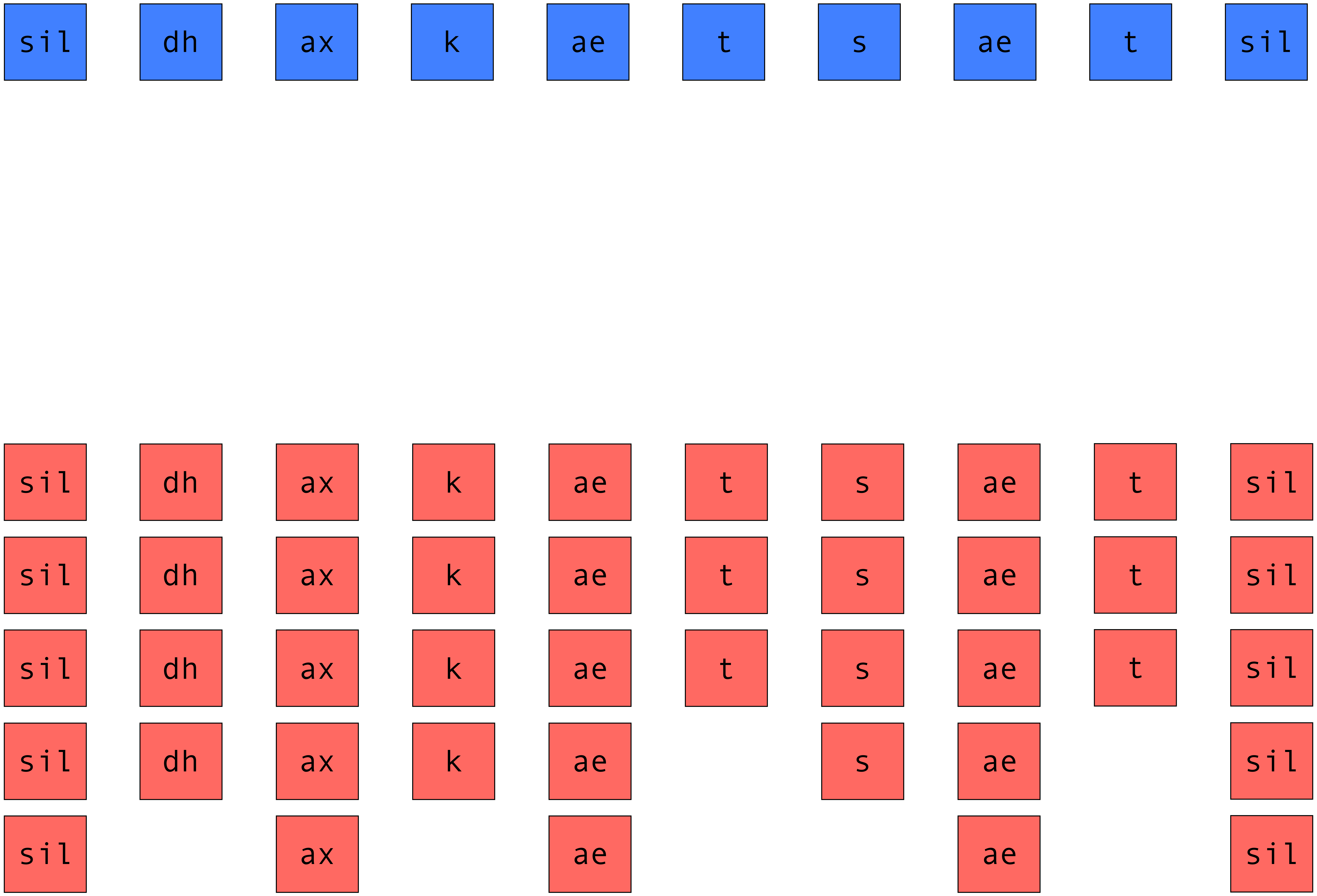


Figure 16.6 from Paul Taylor "Text-to-speech synthesis", 2009, Cambridge University Press, Cambridge, ISBN 0521899273

Predicting acoustic properties of the target units

- Think of this as ‘partial synthesis’
  - *do not* need to predict **all** acoustic properties
  - *do not* need to actually generate a speech **waveform**
  - just need to predict **sufficient** properties to allow **comparison** with candidate units

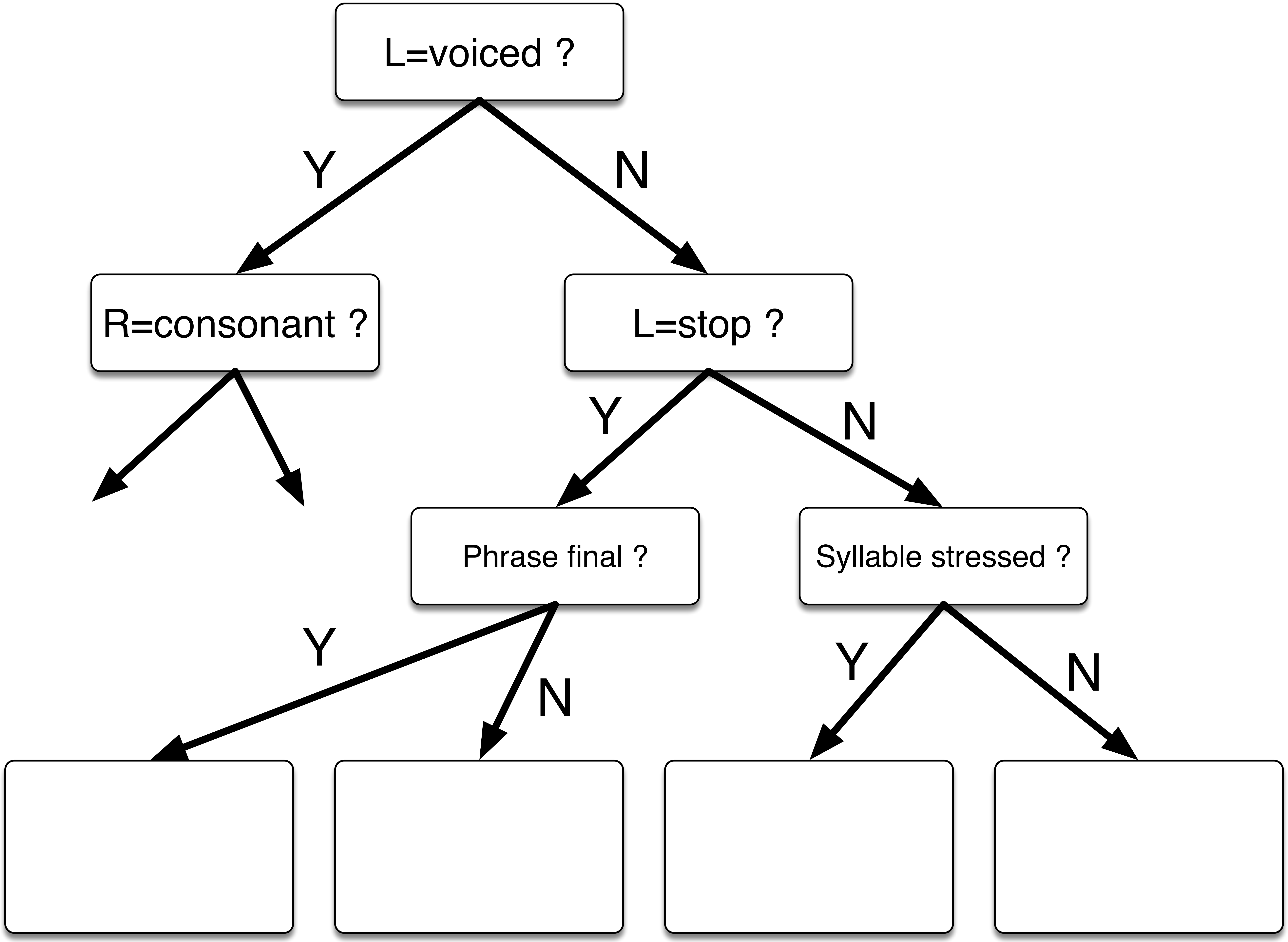


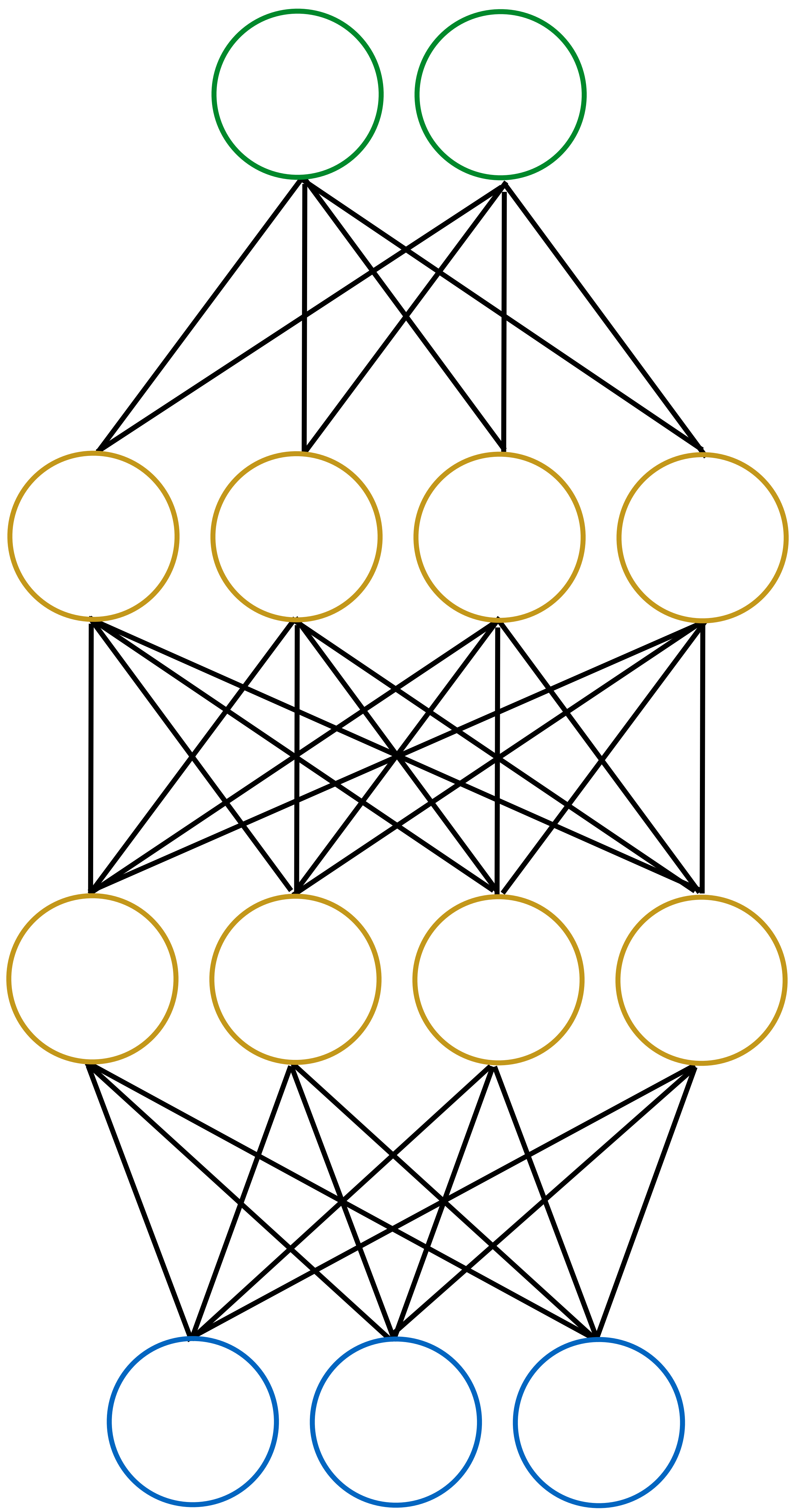
sil dh ax k ae t s ae t sil
sil dh ax k ae t s ae t sil
sil dh ax k ae t s ae t sil
sil dh ax k ae t s ae t sil
sil dh ax k ae s ae sil
sil ax ae ae sil

What exactly are the acoustic features?

- We have choices:
  - simple acoustic properties such as F0, duration and energy
  - a more detailed specification such as the spectral envelope (e.g., as cepstral coefficients)

- It will only work if we can **accurately predict** these properties from the linguistic features
  - how about predicting a *complete* acoustic specification?


L=voiced ?
Y
N
R=consonant ?
L=stop ?
Y
N
Phrase final ?
Syllable stressed ?
Y
N
Y
N



Combining IFF and ASF into a single target cost function

- Many actual systems actually use a mixed IFF + ASF target cost function
  - some sub-costs use linguistic features, others use acoustic features
  - each is weighted appropriately

- Why use **both types** of sub-cost?
  - ASF escapes **some of the sparsity problems** inherent in IFF
  - but our acoustic properties **do not capture all possible acoustic variation**
    - e.g., voice quality, such as phrase-final creaky voice
  - *and, of course, our predictions of acoustic properties will contain **errors***

Orientation

- Summary of unit selection design choices

  - Unit type
  - Target cost
  - Join cost
  - Search
  - Database

Orientation

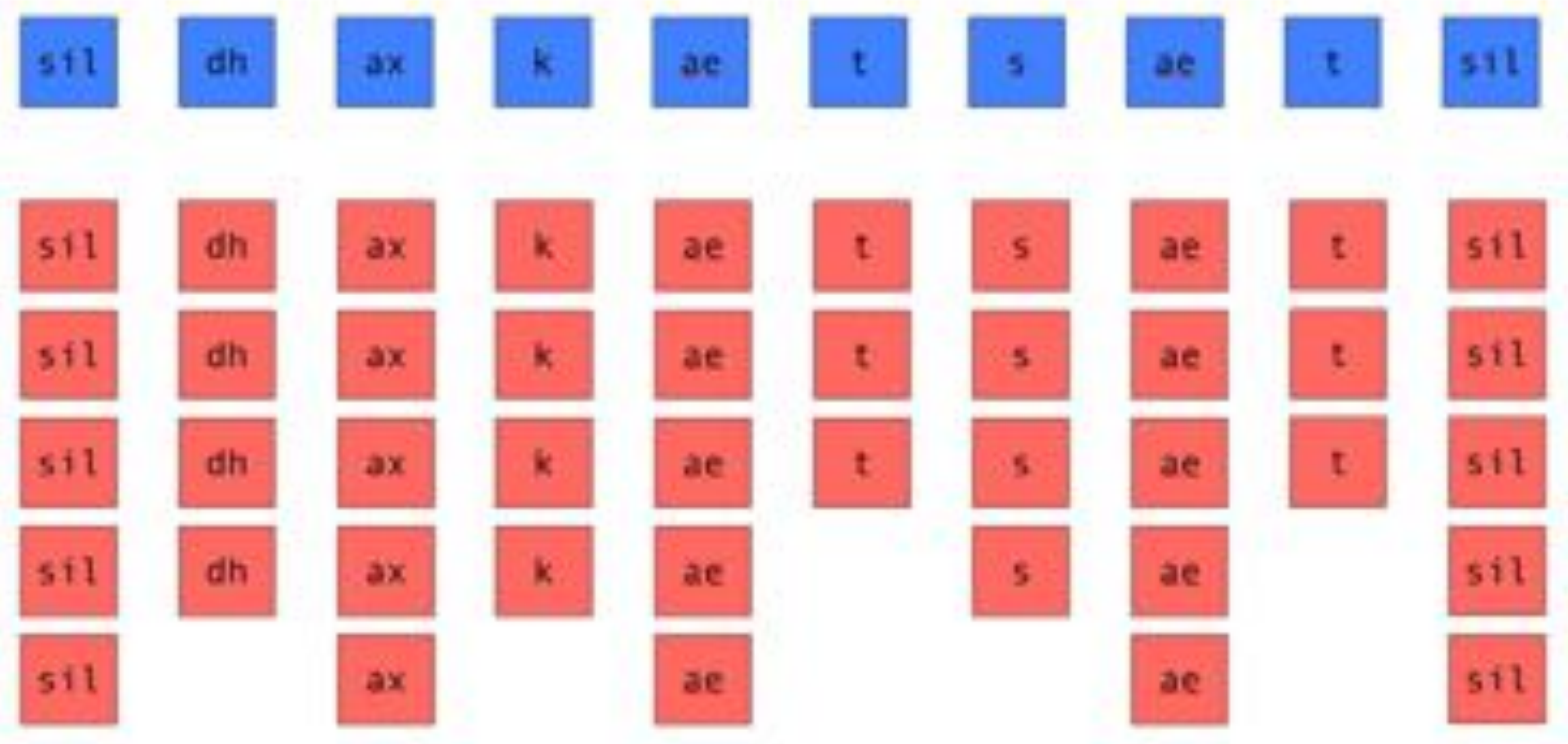
- Summary of unit selection design choices

  - Unit type
  - Target cost
  - Join cost
  - Search
  - Database

Orientation

- Summary of unit selection design choices
  - Unit type
  - Target cost
  - Join cost
  - Search
  - Database

Often diphones or half-phones.
Use the "zero join cost trick" to effectively use (much) larger units



Orientation

- Summary of unit selection design choices
  - Unit type
  - Target cost
  - Join cost
  - Search
  - Database

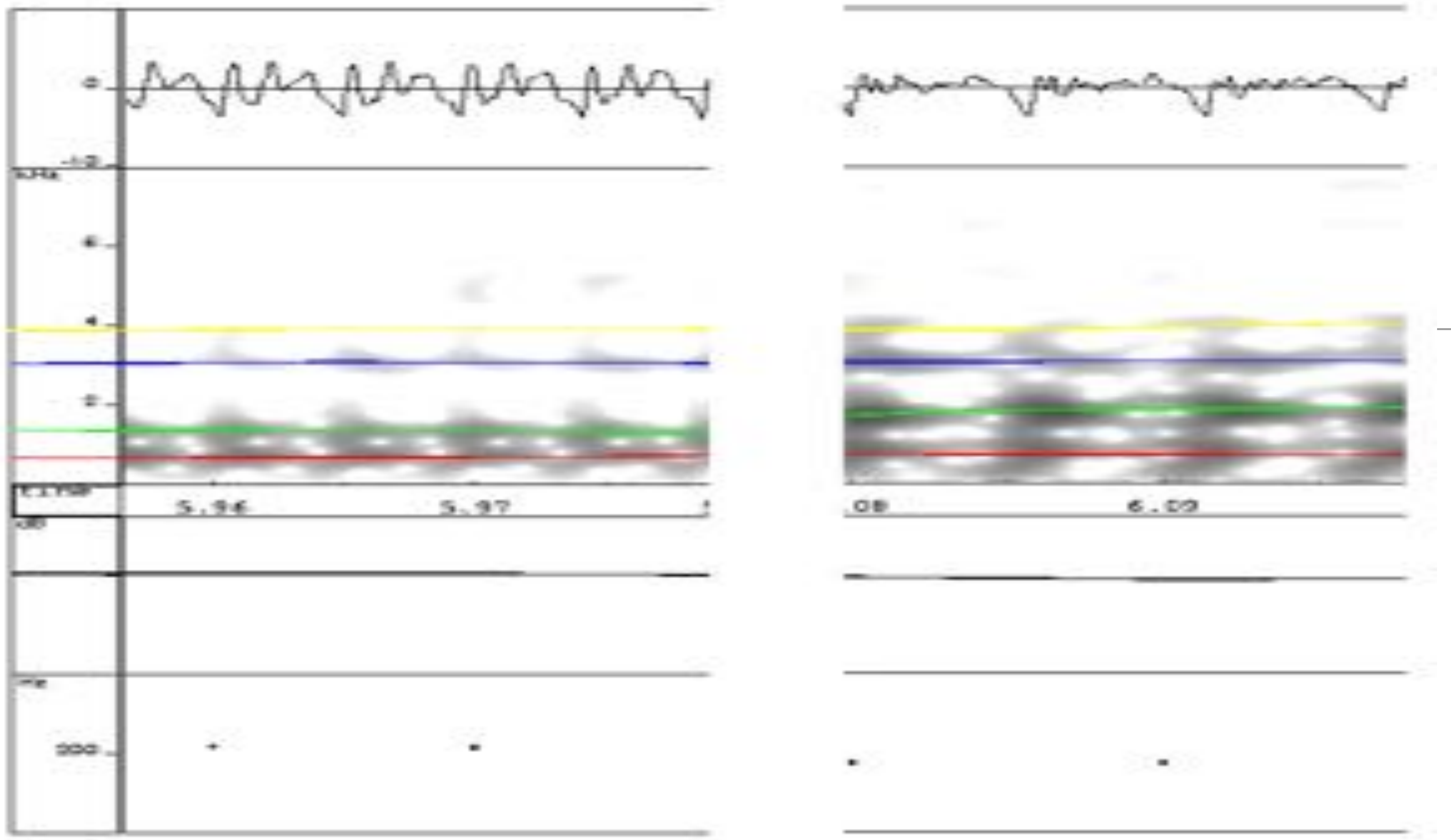
→

Pure IFF only using linguistic features

Pure ASF, involving 'partial synthesis'
(must decide which acoustic features to predict)

Mixed IFF + ASF

Orientation



- Summary of unit selection design choices

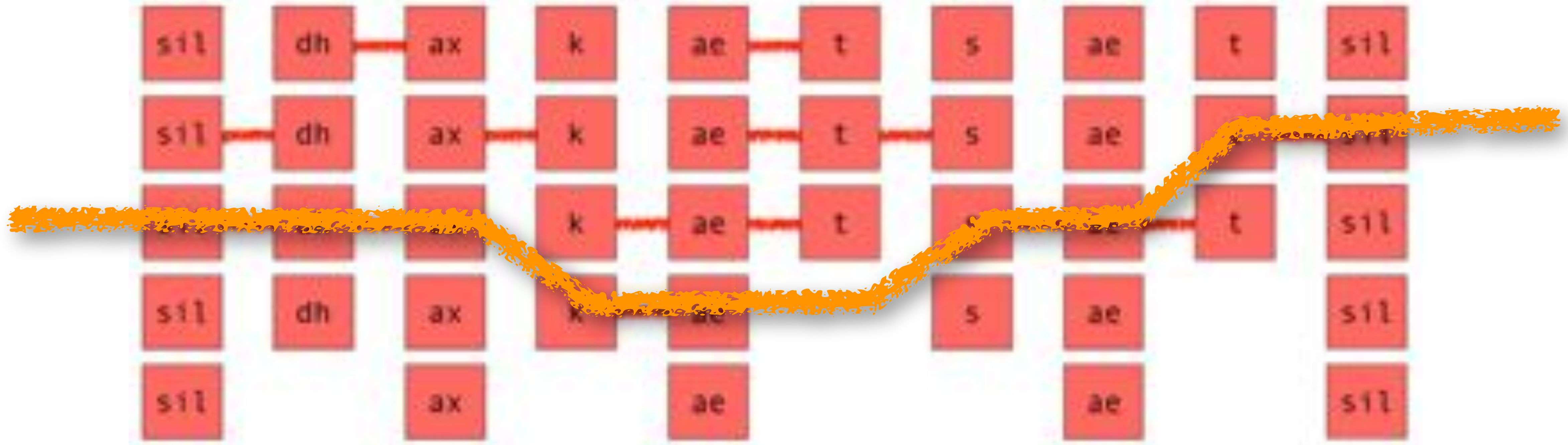
  - Unit type
  - Target cost
  - Join cost
  - Search
  - Database

Usually includes F0, energy and spectral envelope

We have not mentioned optional smoothing of joins using signal processing.

Orientation

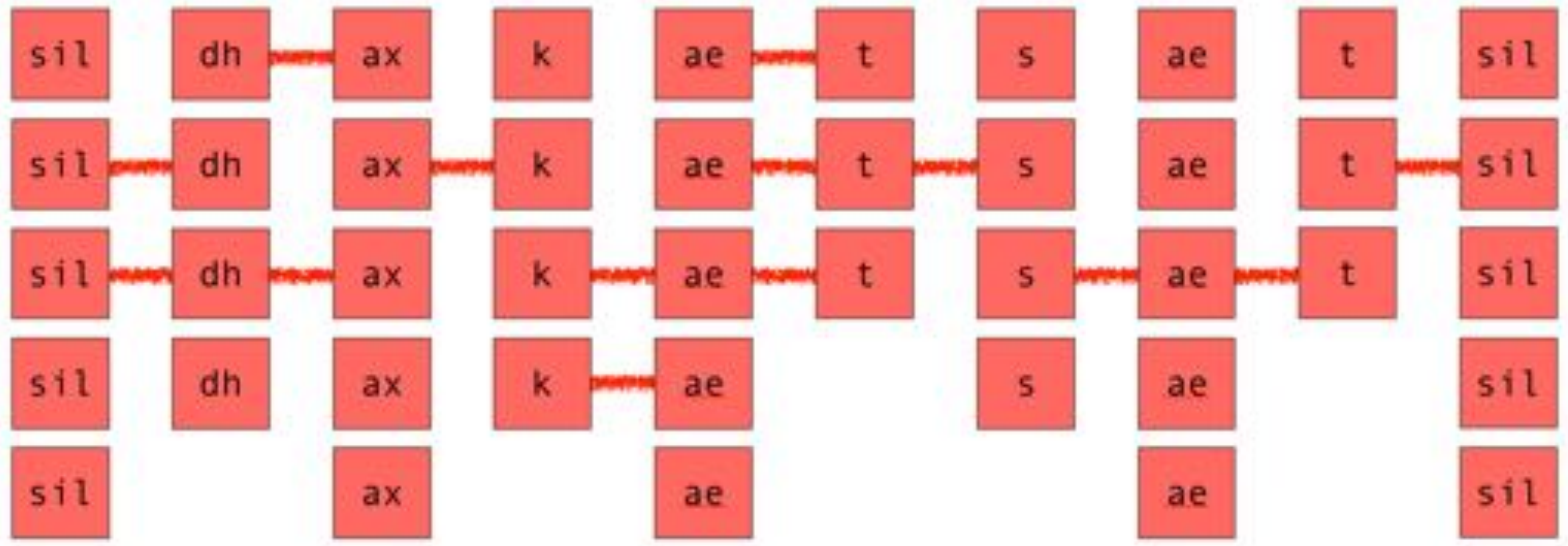
- Summary of unit selection design choices



- Unit type
- Target cost
- Join cost
- Search
- Database

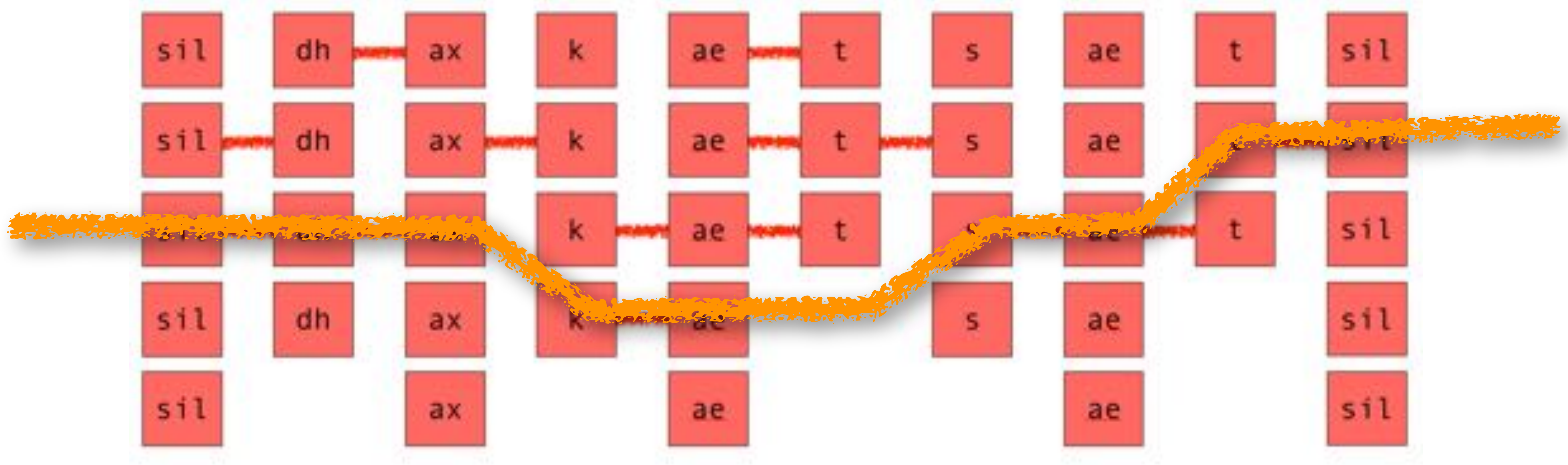
Efficient dynamic programming

As in Automatic Speech Recognition,
can use **pruning** to make it as fast as needed


sil
dh
ax
k
ae
t
s
ae
t
sil

Orientation

- Summary of unit selection design choices



  - Unit type
  - Target cost
  - Join cost
  - Search
  - Database

Efficient dynamic programming

As in Automatic Speech Recognition,
can use **pruning** to make it as fast as needed

Orientation

- Summary of unit selection design choices

  - Unit type
  - Target cost
  - Join cost
  - Search
  - Database

Coming next…

What next?

- How to create the **database**
  - what to record
  - how to record it
  - how to annotate it

- Later, *after* we learn about statistical parametric speech synthesis
  - we can use that statistical **model** in the ASF target cost function of a unit selection synthesiser
  - this is called **hybrid** synthesis

What next?

- How to create the **database**
  - what to record
  - how to record it
  - how to annotate it

Knowing what **features**
our target cost requires,
will help us design a suitable
database of recorded speech

- Later, *after* we learn about statistical parametric speech synthesis
  - we can use that statistical **model** in the ASF target cost function of a unit selection synthesiser
  - this is called **hybrid** synthesis

What next?

- How to create the **database**
  - what to record
  - how to record it
  - how to annotate it

We will have to annotate
the database with the **features**
that our target cost requires

- Later, *after* we learn about statistical parametric speech synthesis
  - we can use that statistical **model** in the ASF target cost function of a unit selection synthesiser
  - this is called **hybrid** synthesis

What next?

- How to create the **database**
  - what to record
  - how to record it
  - how to annotate it

- Later, *after* we learn about statistical parametric speech synthesis
  - we can use that statistical **model** in the ASF target cost function of a unit selection synthesiser
  - this is called **hybrid** synthesis

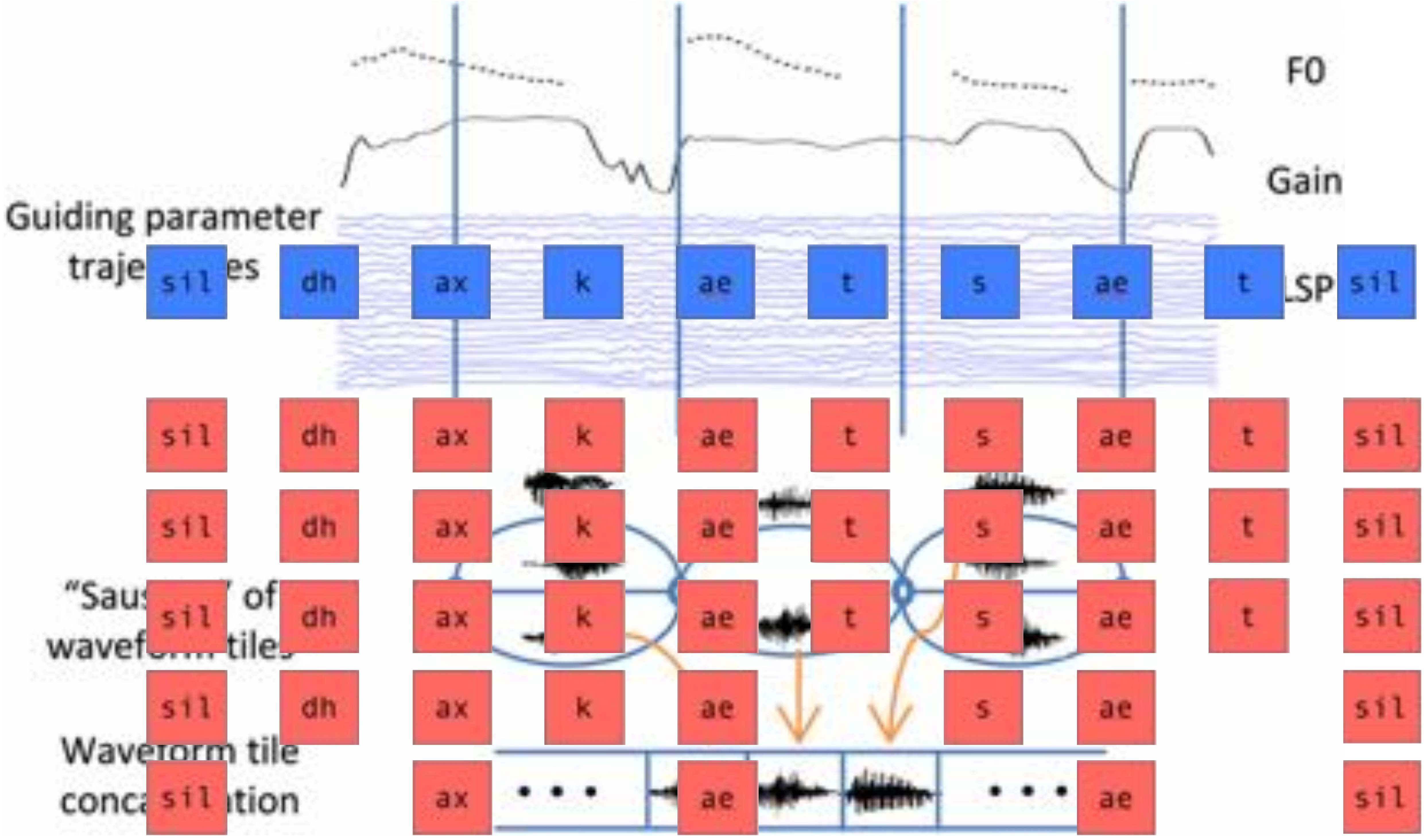


Figure 1 from Y. Qian, F. K. Soong and Z. J. Yan "A Unified Trajectory Tiling Approach to High Quality Speech Rendering" IEEE Trans. Audio, Speech, and Language Proc. 21 (2), pp. 280-290, 2013. DOI: 10.1109/TASL.2012.2221460