If you lose your voice, how can you speak?

Simon King

Centre for Speech Technology Research
University of Edinburgh

If you lose your voice, how can you speak?

In the first part of this talk, I'll give an easy-to-understand, non-technical overview of the SpeakUnique project, in which we are providing personalised speech communication aids to people who are losing their own voice due to Motor Neurone Disease or other progressive conditions. We are currently conducting trials, to measure the improvement to quality-of-life that these communication aids give.

The second part of the talk will get a little more technical, where I will describe how the technology works. Using powerful statistical models, and a large database of donated speech from thousands of people, we create accent- and gender-specific "Average Voice Models". These are then further modified to produce speech that sounds like a particular person.

A unique capability of our approach is that it only needs a small sample of that person's speech and this sample may be disordered: the person is already becoming hard to understand. We are able to "repair" the voice by interchanging or interpolating parts of the Average Voice Model into a model learned from the person's own speech. This results in a computer-generated voice that sounds like a normal, intelligible version of the person. This is finally installed on a mobile device, such as an iPad, for the person to use in daily life.

Project website: www.speakunique.org

Part 1: easy-to-understand overview of SpeakUnique

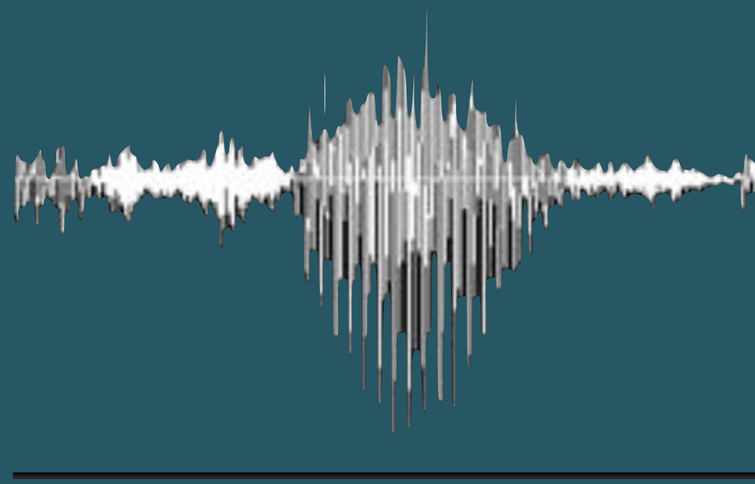
a project that provides personalised voice communication aids

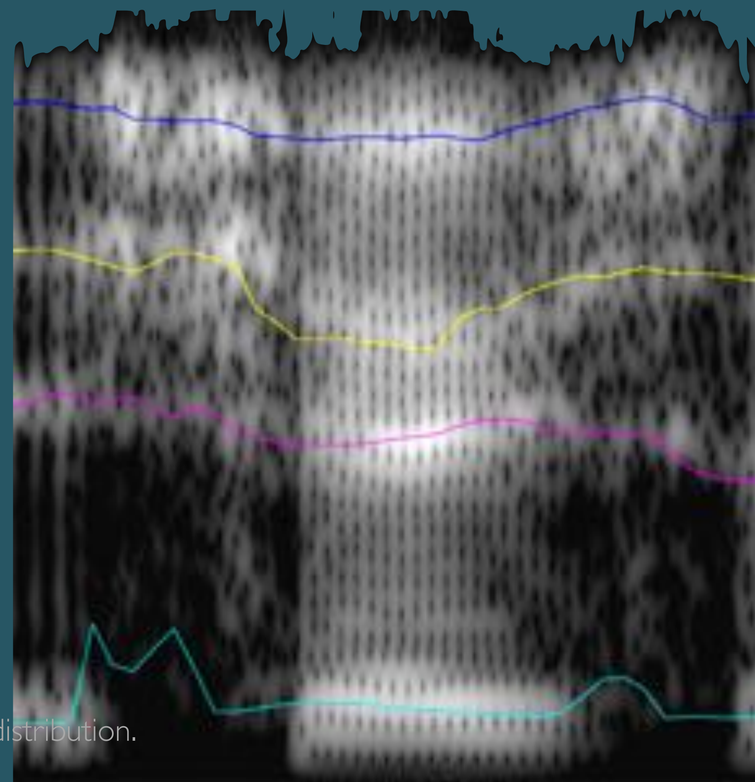
Speech

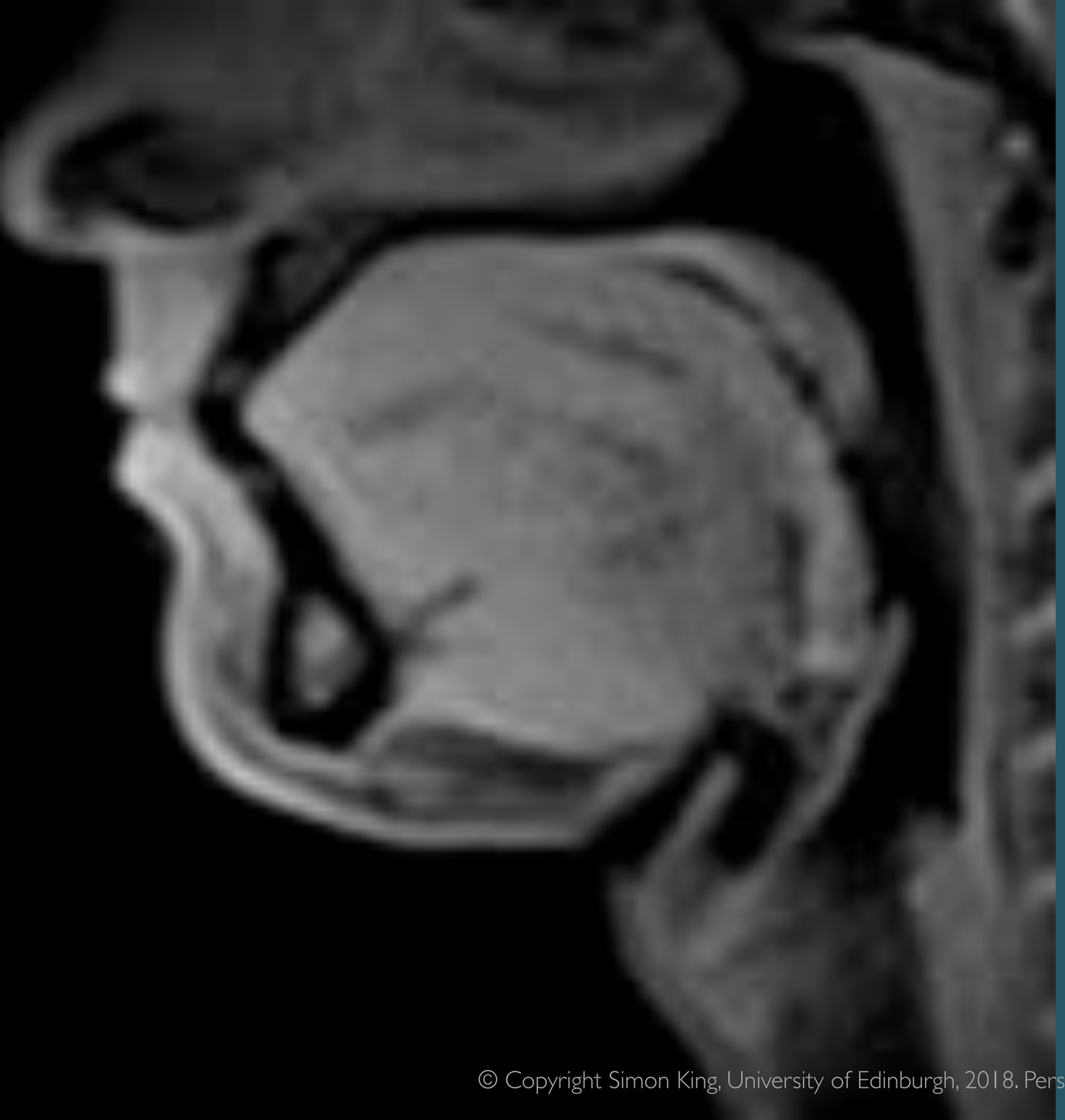
How it is made

Breaking it down

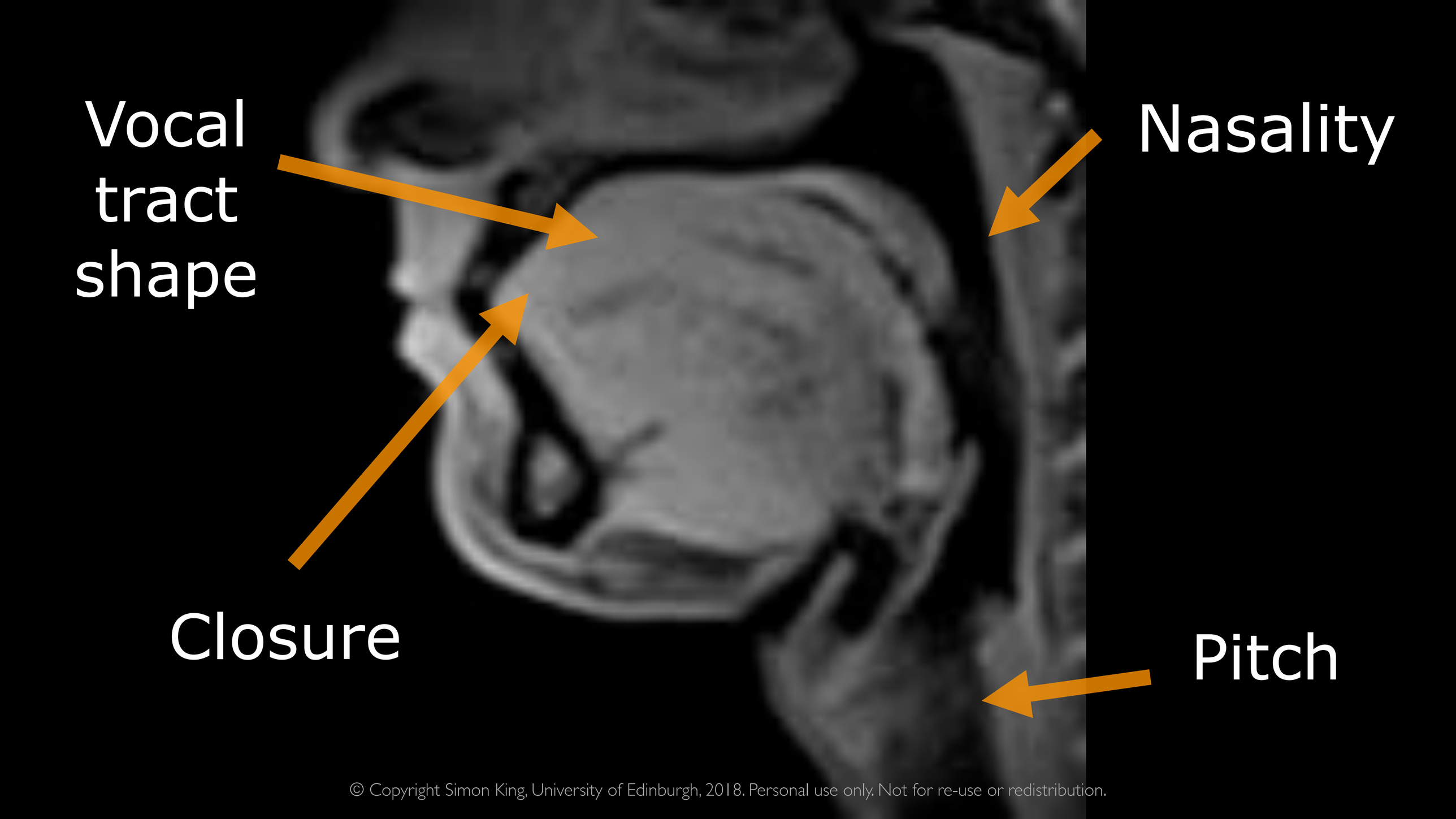
Recreating it with a computer





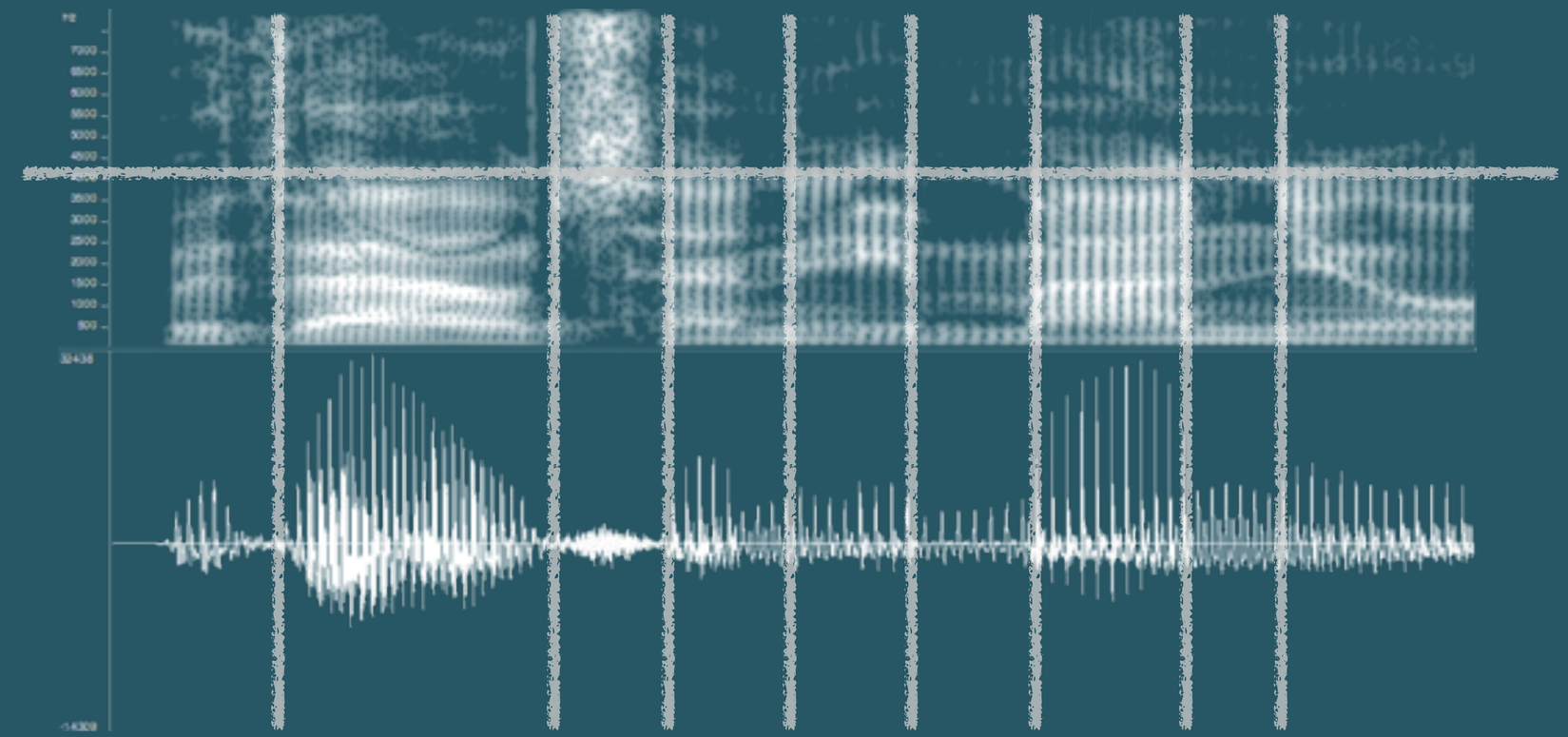


How speech is made



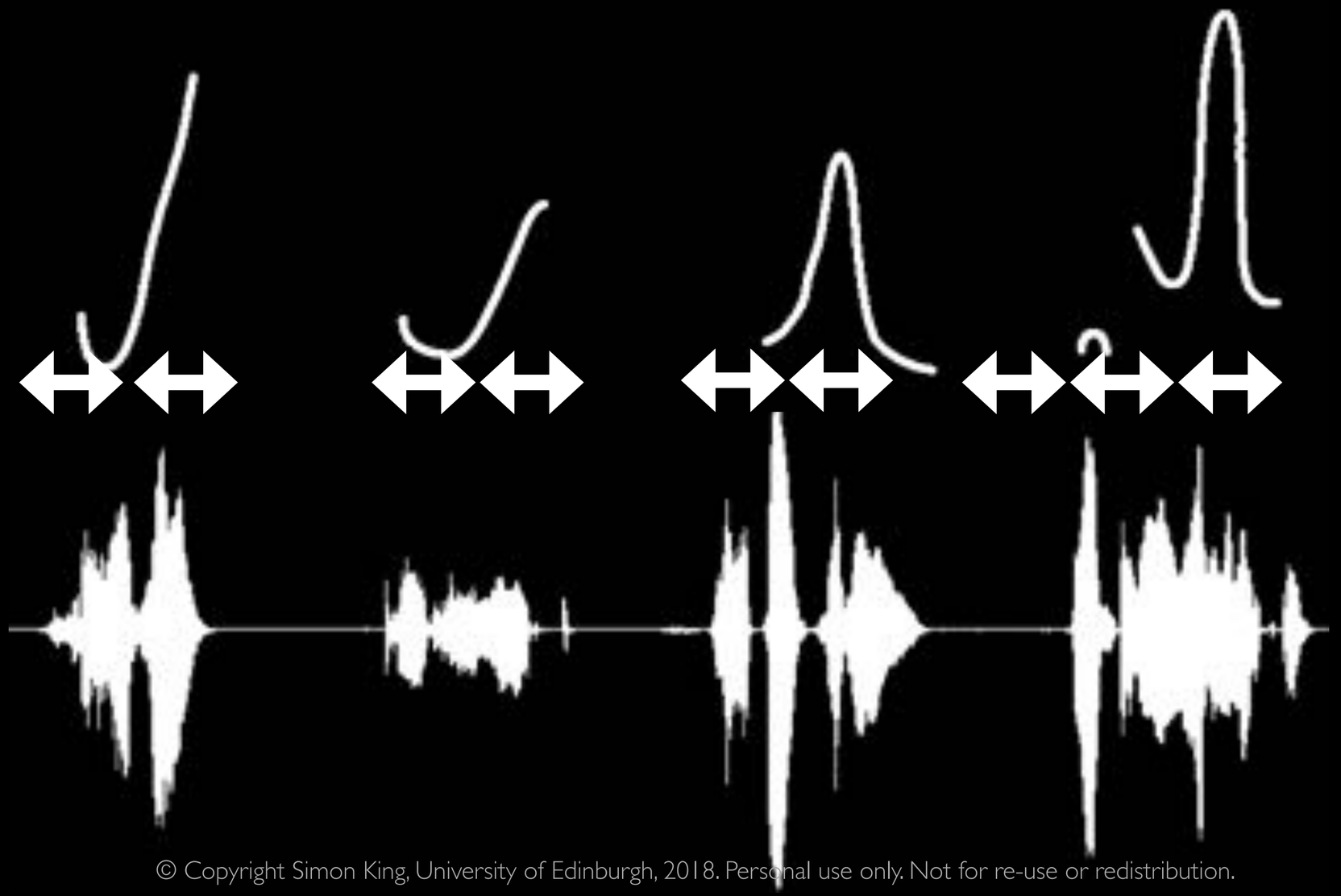
Analysing speech - and breaking it into pieces

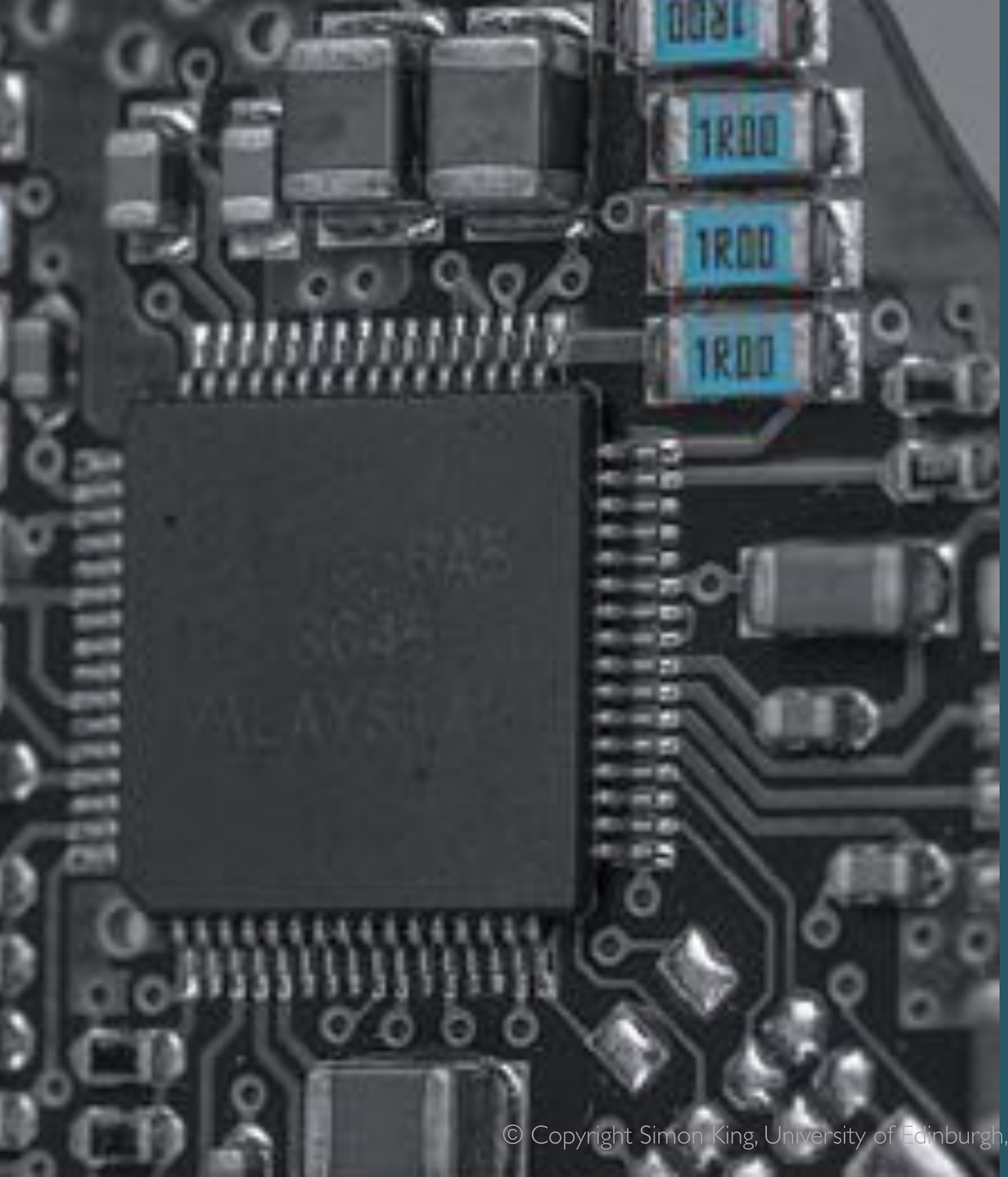
frequency content



categories of speech sounds: phonemes

Prosody

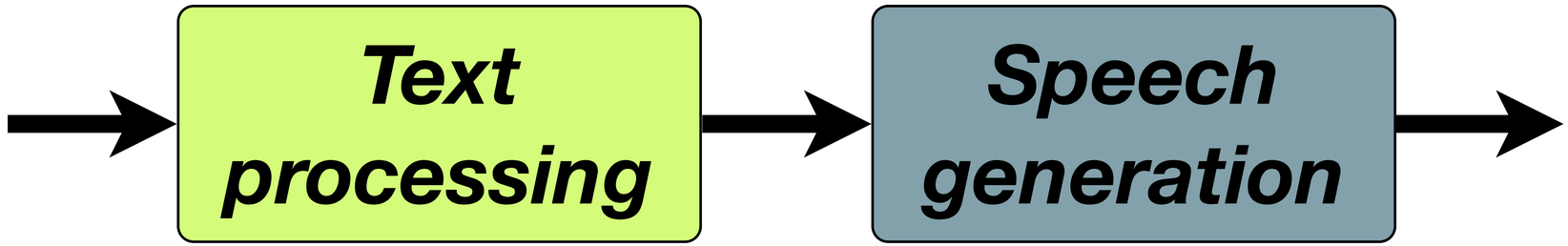




Creating speech with a computer

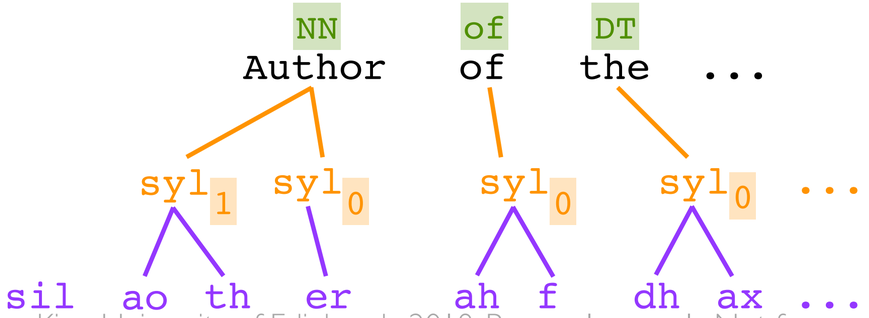
we call this “speech synthesis”

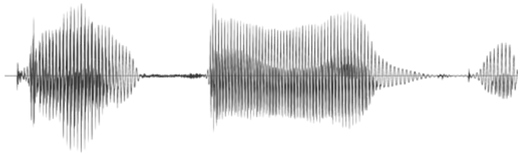
Not one, but **two** hard problems



text *"how to say this text"* *speech*

`Author of the…`





Dr. Smith lives at 123 Orchard Dr.

Buy me an IPA and we'll be BFF.

Did you see the meme about geotagging on your staycation?

played
at risks,
cut into

risky,

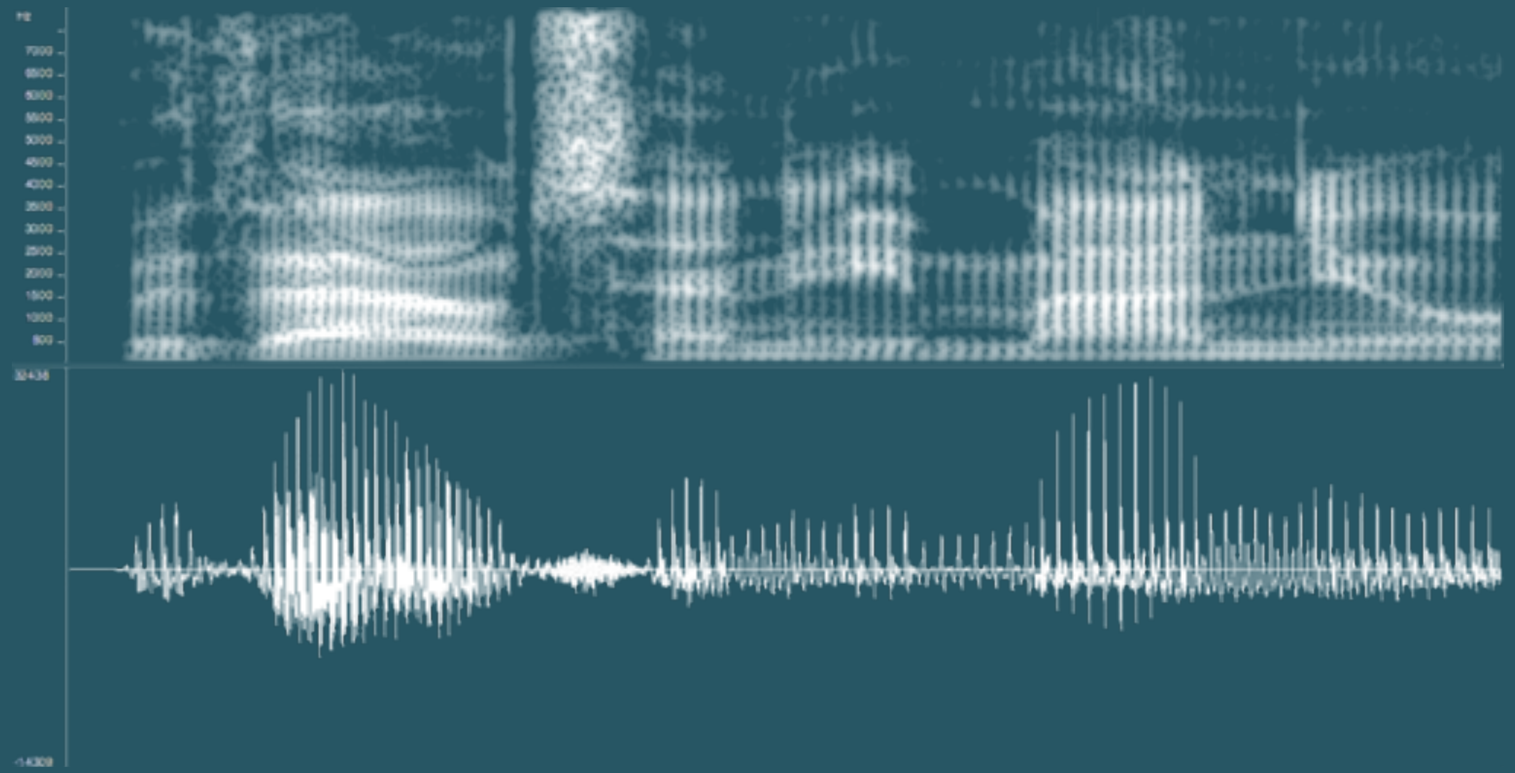
es) di-
efined
ed to

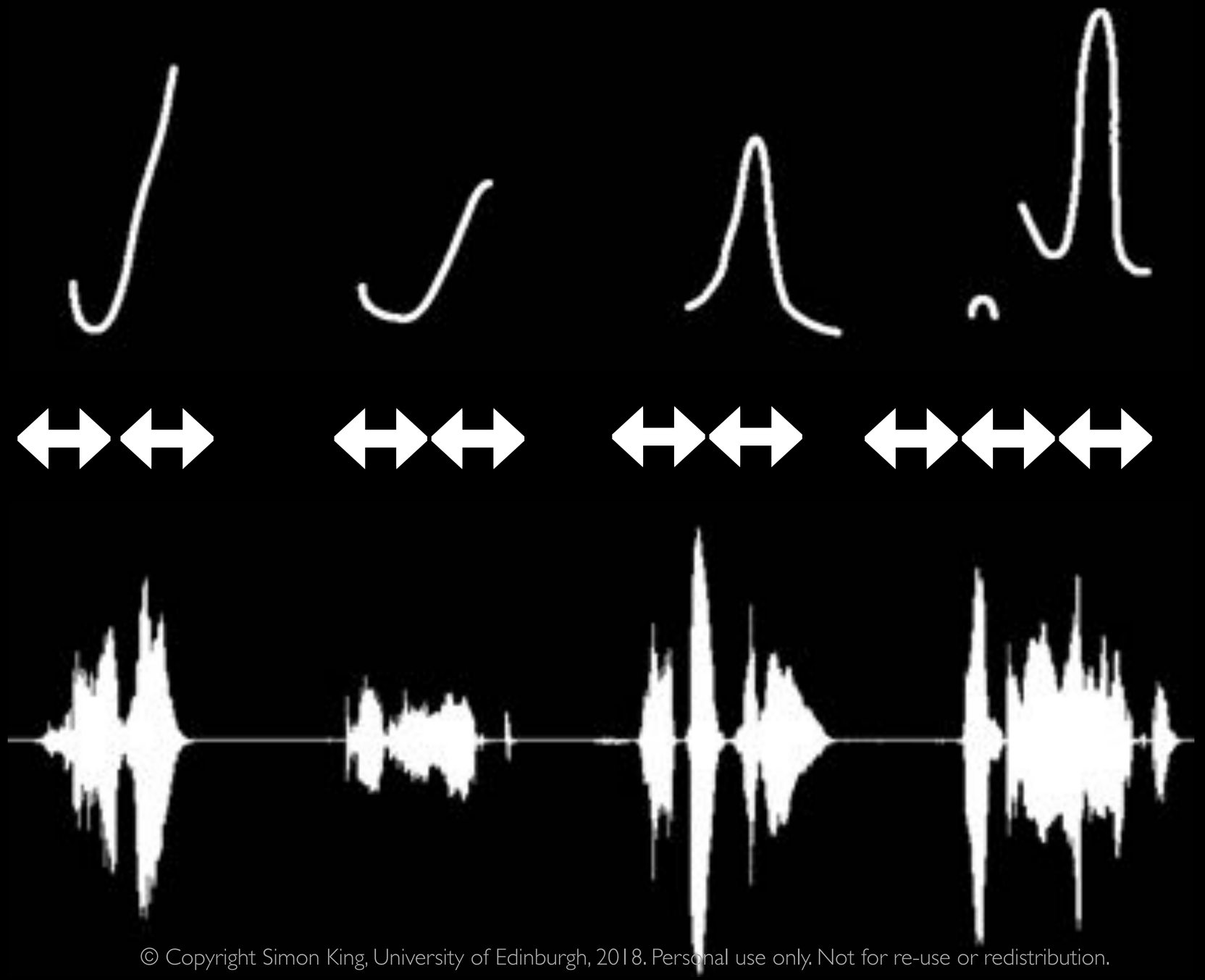
like a dictator. 2 overbearing
orially *adv.* [Latin: related
TATOR]

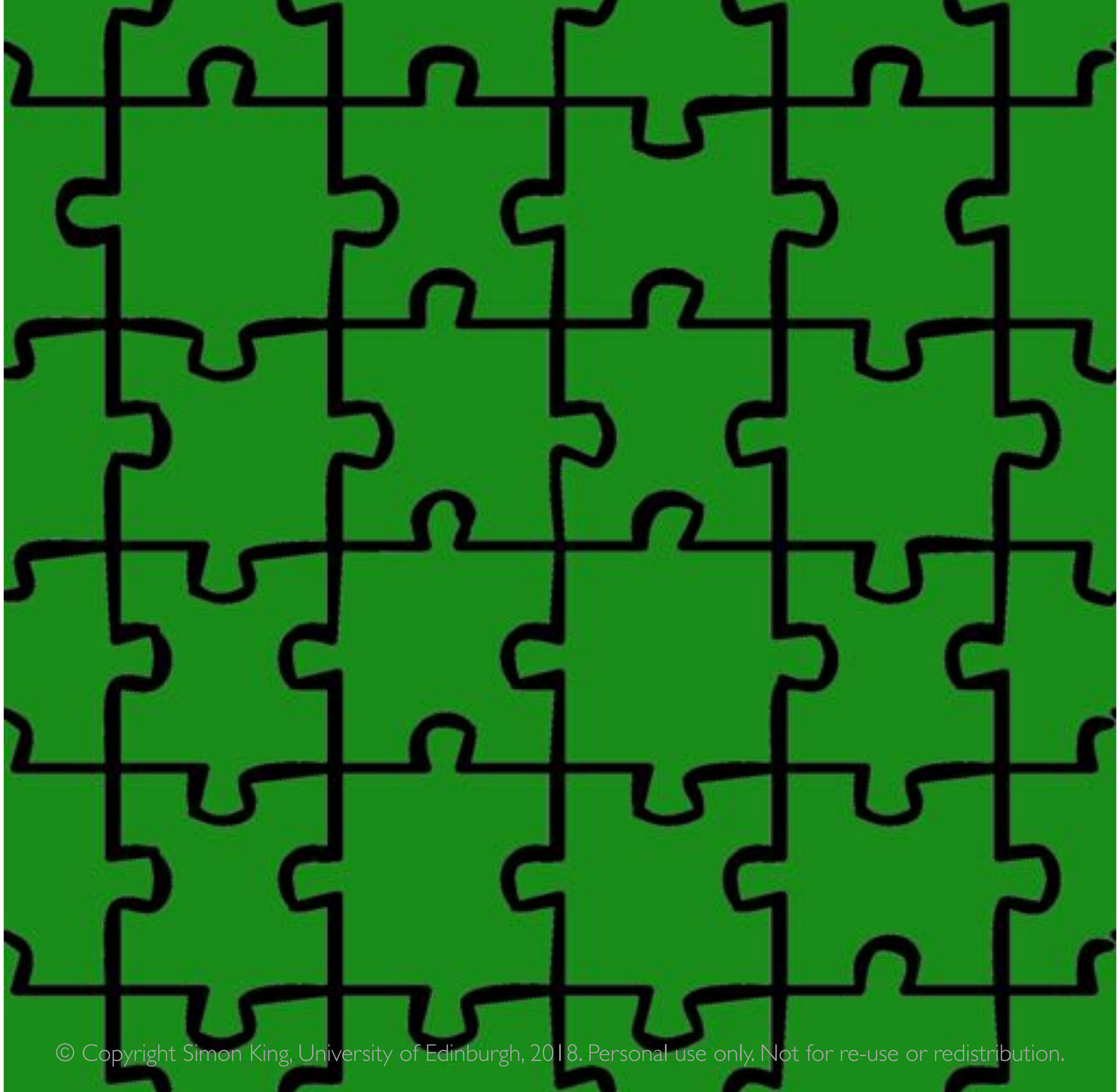
diction /'dɪkʃ(ə)n/ *n.* manner
ciation in speaking or singing
dictio from *dico dict-* say]

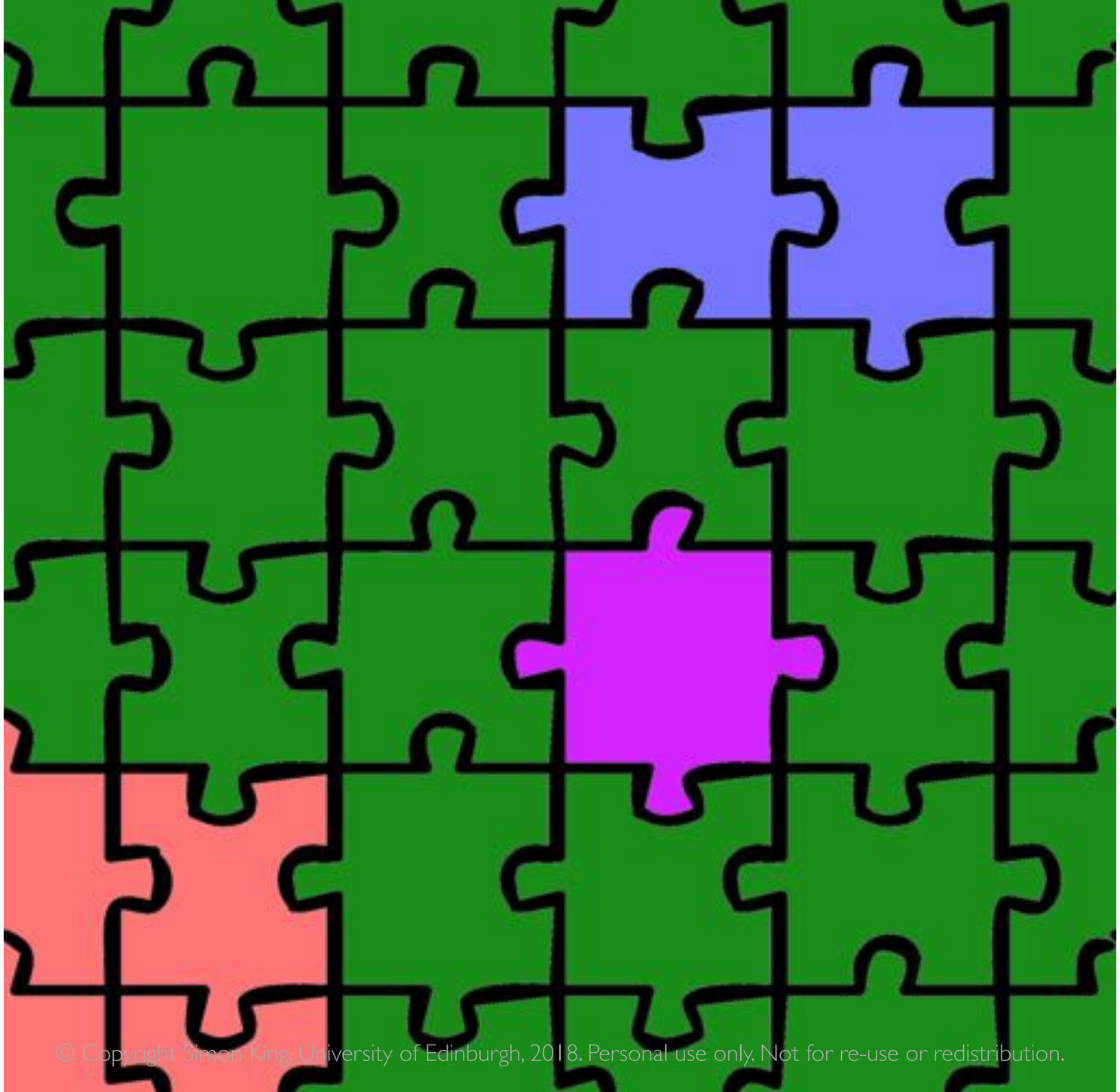
dictionary /'dɪkʃənərɪ/ *n.* (p
book listing (usu. alphabetic
explaining the words of a lan
giving corresponding words i
language. 2 reference book e
the terms of a partic

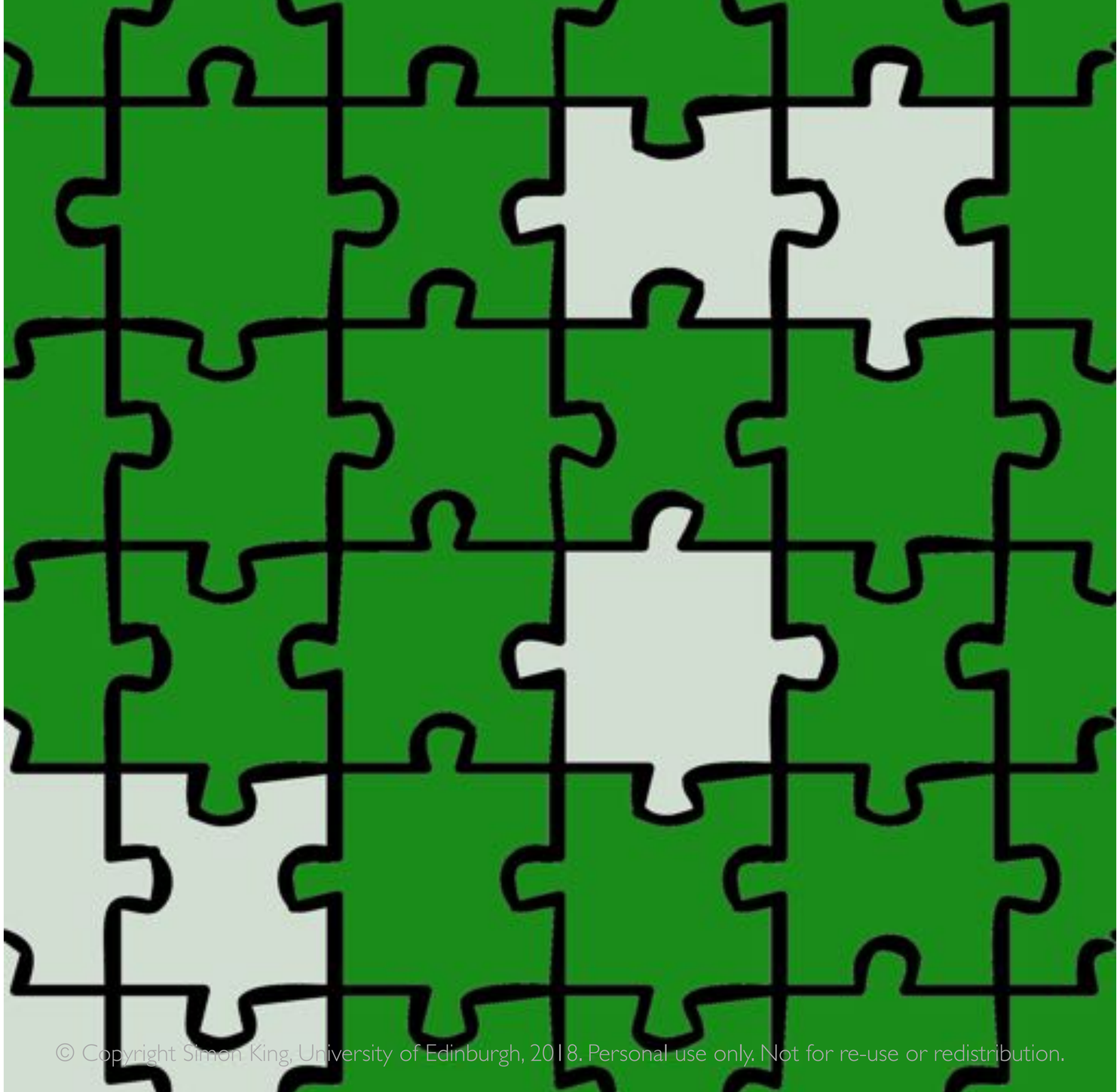
Speech also carries your **identity**

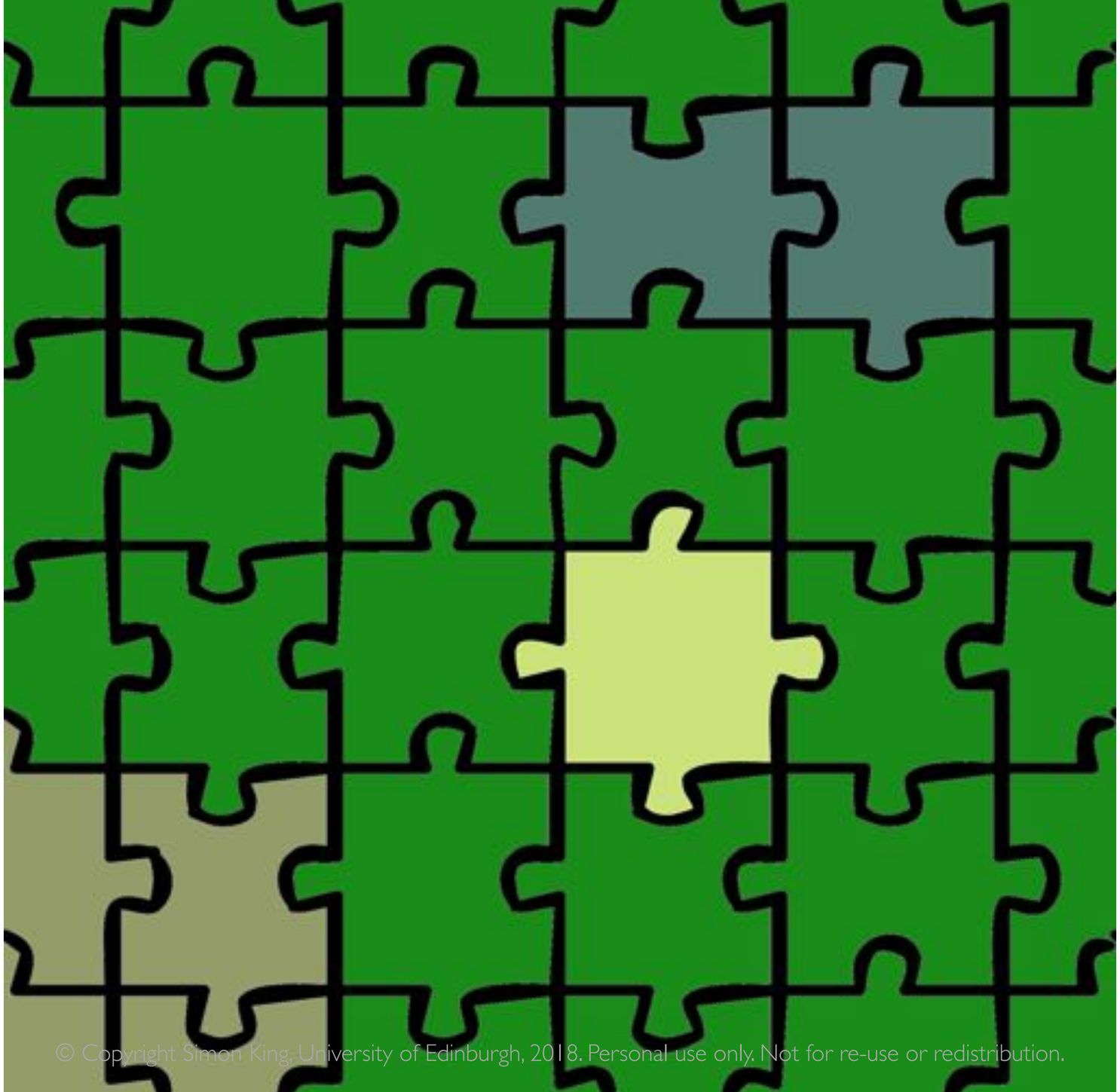




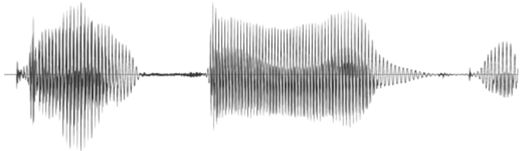








Synthetic speech for people who have damaged voices



lots of healthy donor voices

personalised synthetic speech

The voicebank

Image copyright: Keppie Design, architects

Part 2: how SpeakUnique works

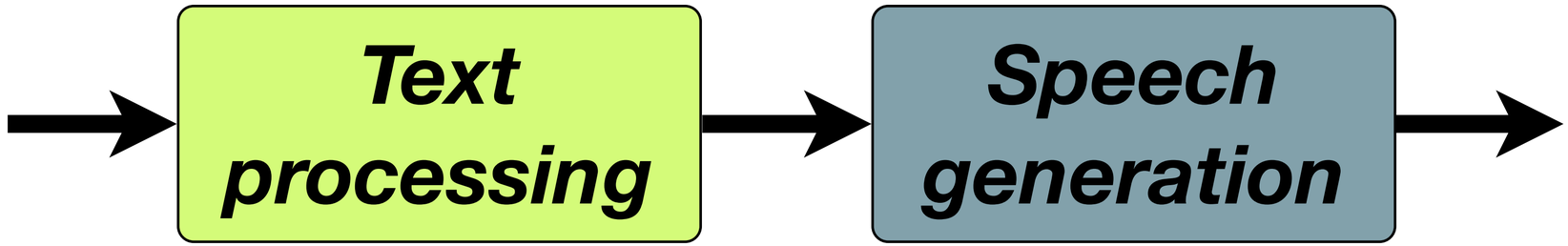
How speech synthesis works

Repairing voices

Open questions & challenges

How speech synthesis works

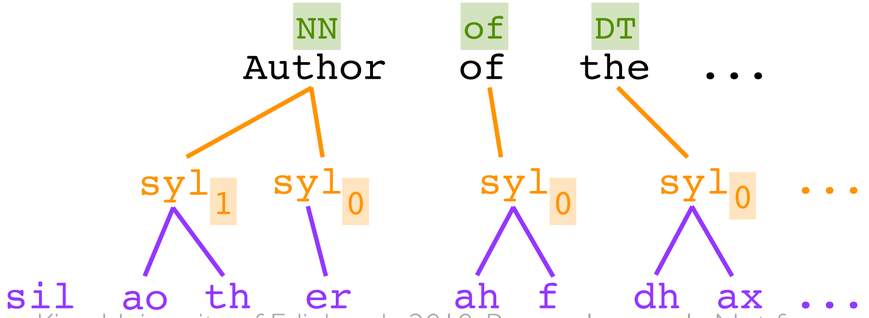
Two hard problems



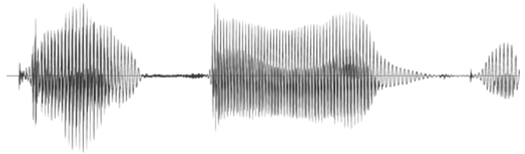
text

Author of the…

linguistic specification



waveform

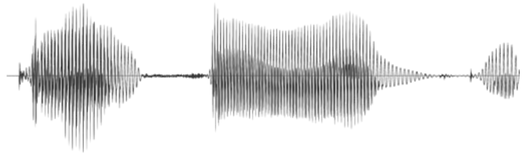


The end-to-end problem we want to solve

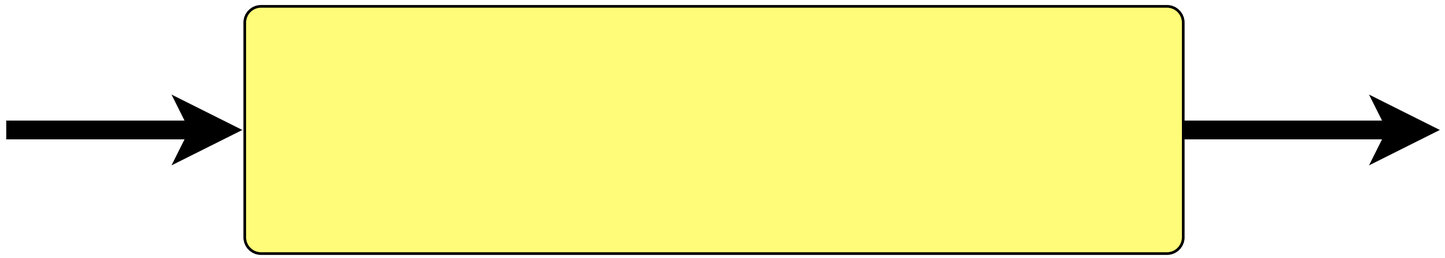
text

```
Author of the…
```

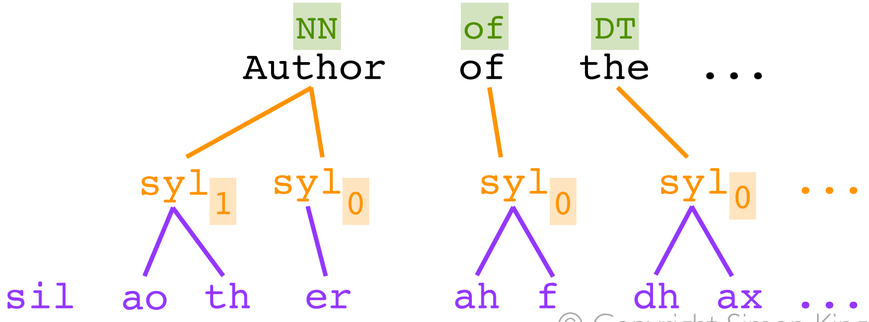
waveform



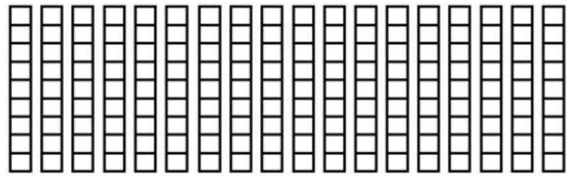
A problem we can actually solve with machine learning



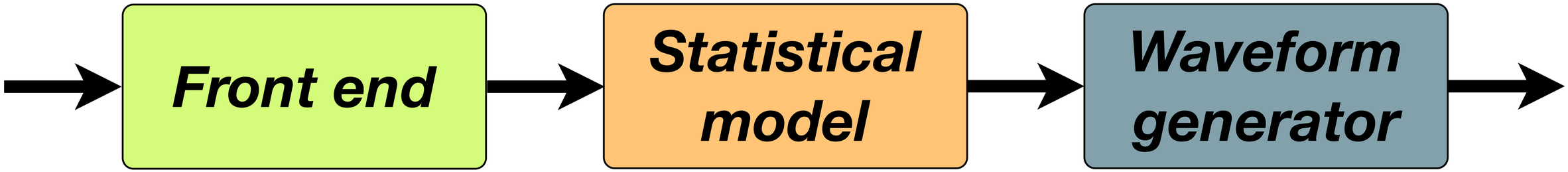
linguistic specification



acoustic features



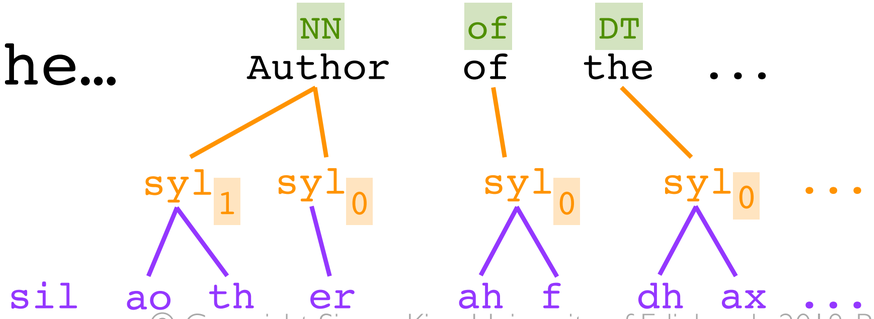
Statistical parametric speech synthesis



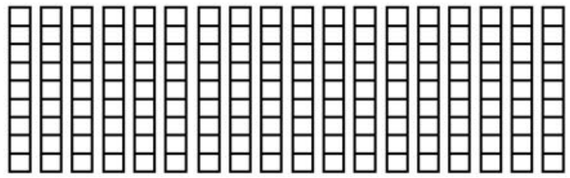
text

Author of the…

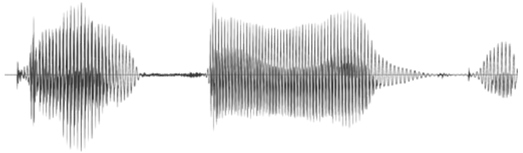
linguistic specification



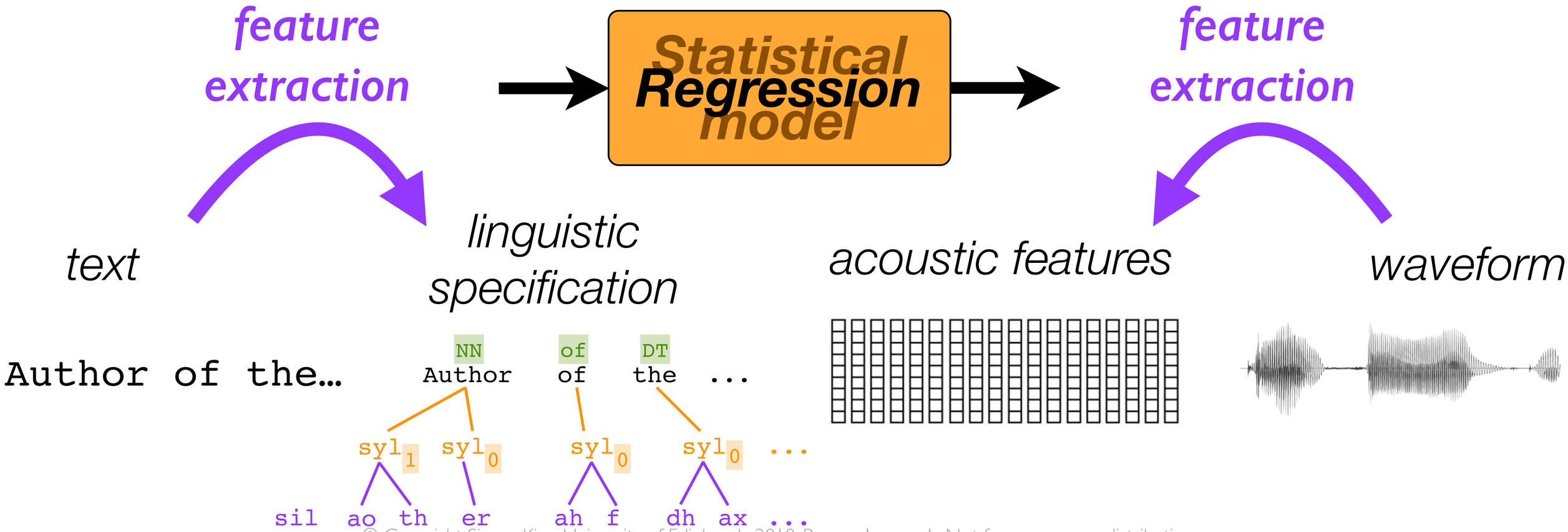
acoustic features



waveform

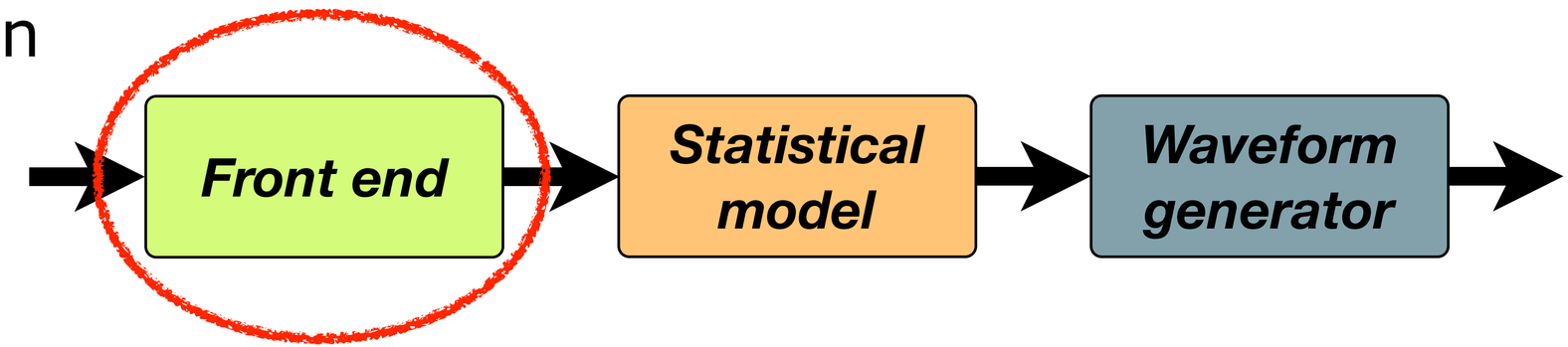


Statistical parametric speech synthesis

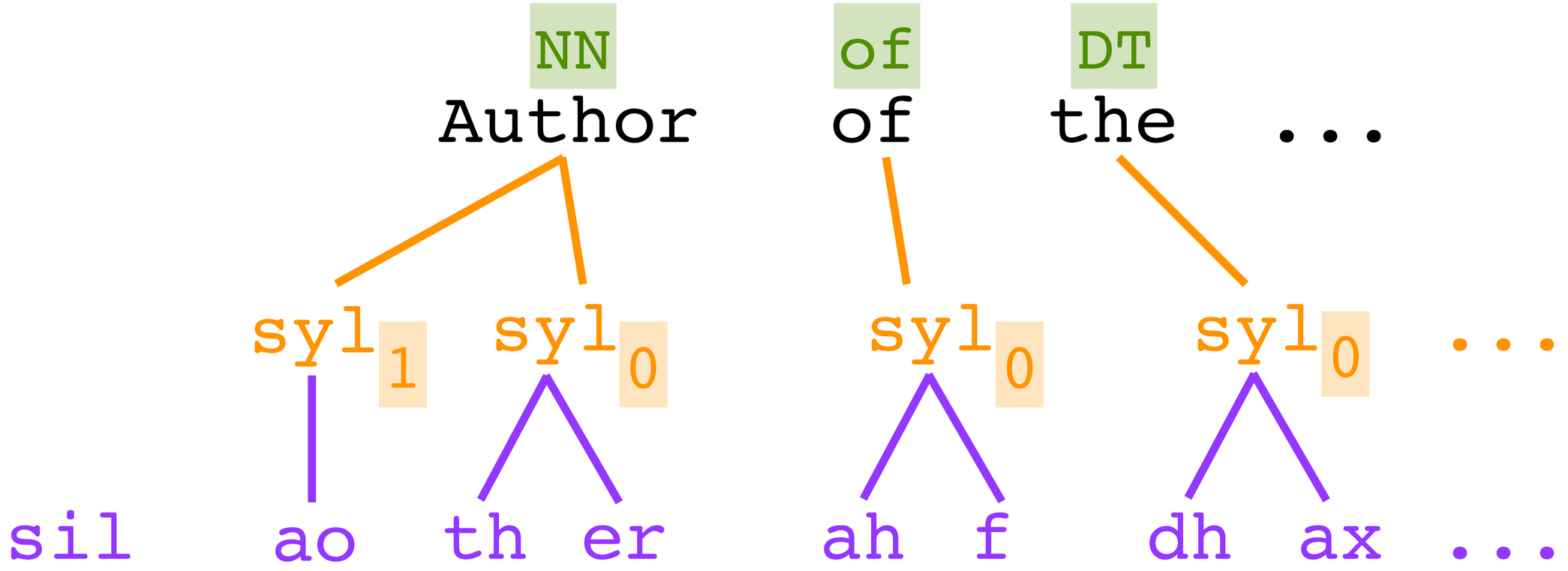


From text to speech

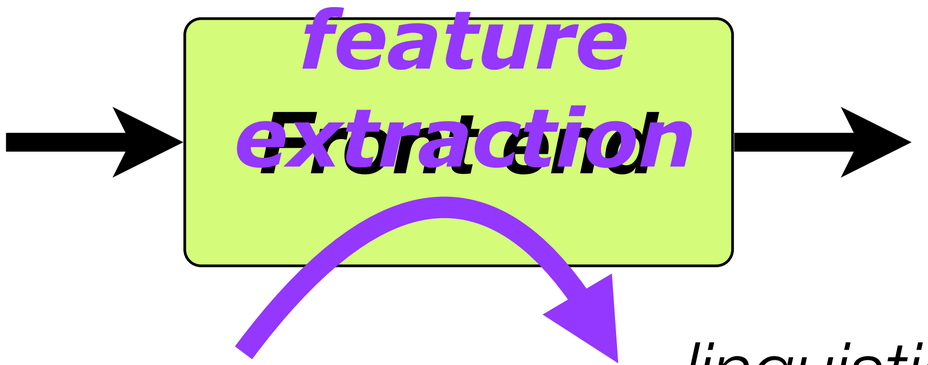
- Text processing
  - pipeline architecture
  - linguistic specification
- Modelling
  - duration model
  - acoustic model
- Waveform generation
  - acoustic features
  - signal processing



The linguistic specification



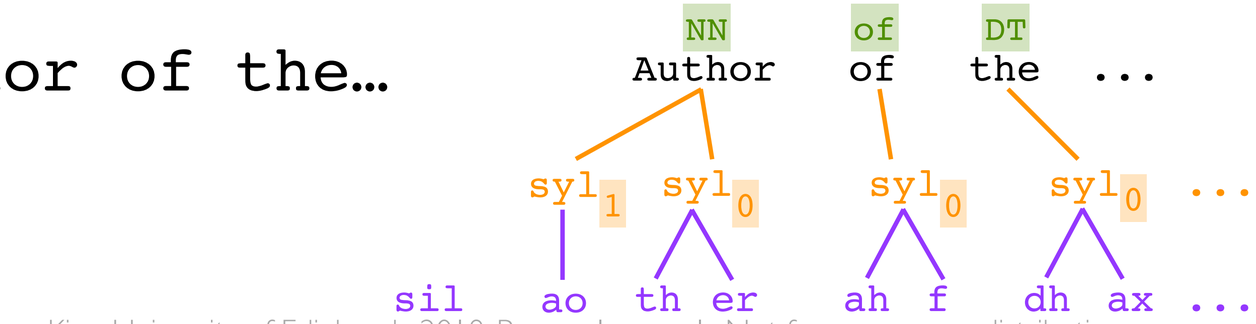
Extracting features from text using the front end



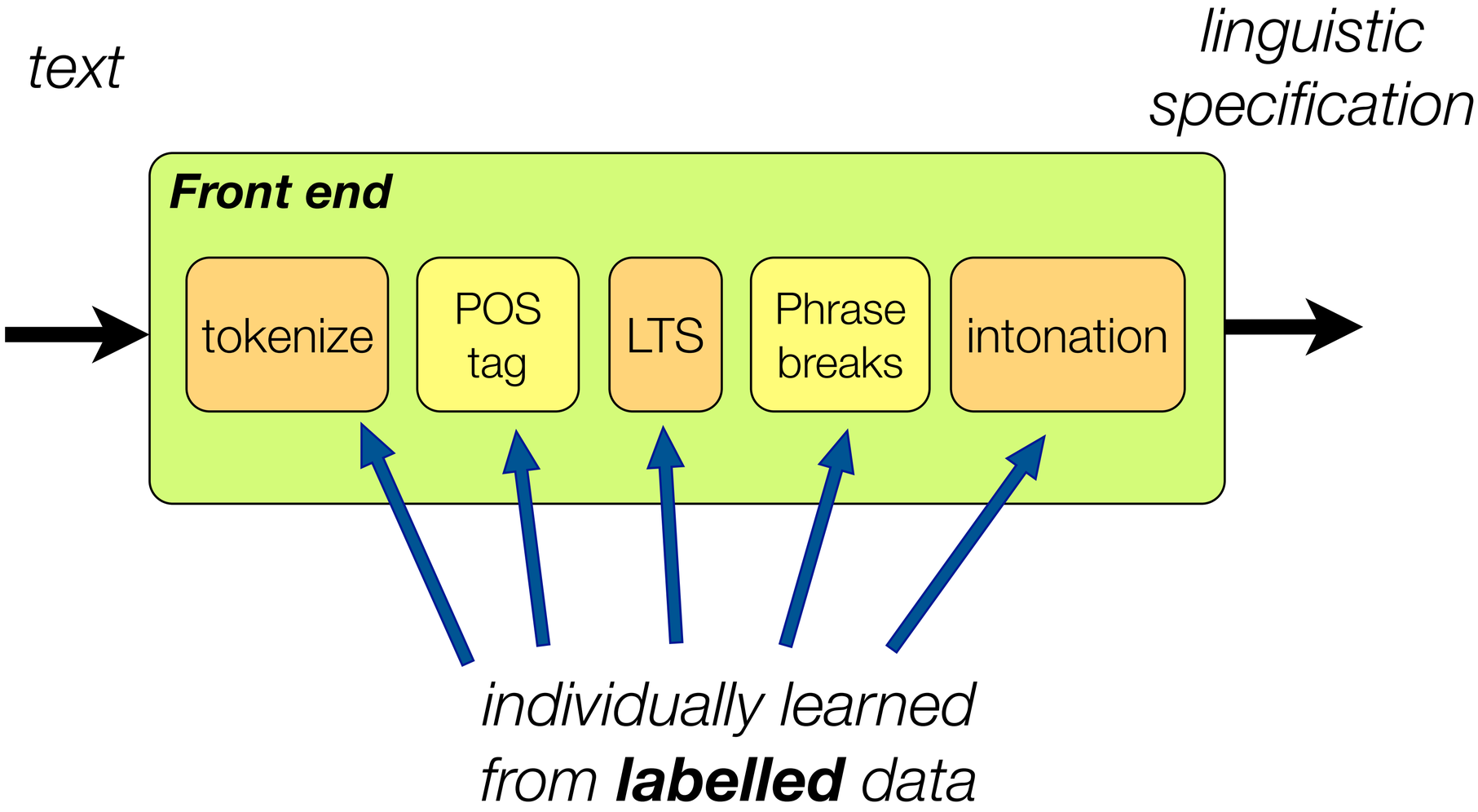
text

linguistic specification

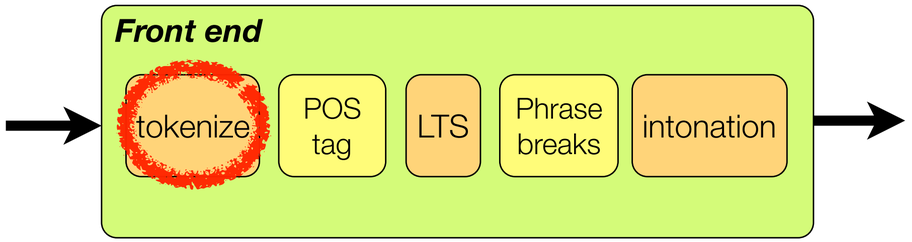
`Author of the…`



Text processing pipeline



Tokenize & Normalize

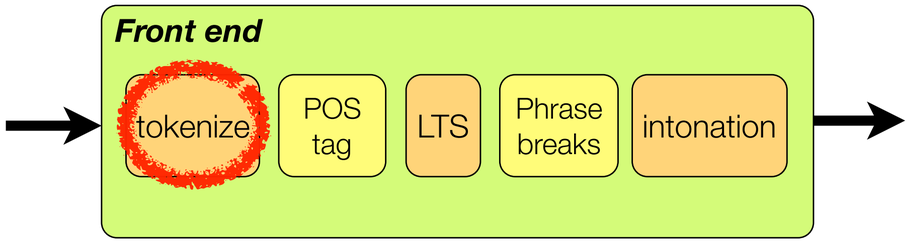


- Step 1: divide input stream into tokens, which are potential words

- For English and many other languages
  - rule based
  - whitespace and punctuation are good features

- For some other languages, especially those that don't use whitespace
  - may be more difficult
  - other techniques required

Tokenize & Normalize

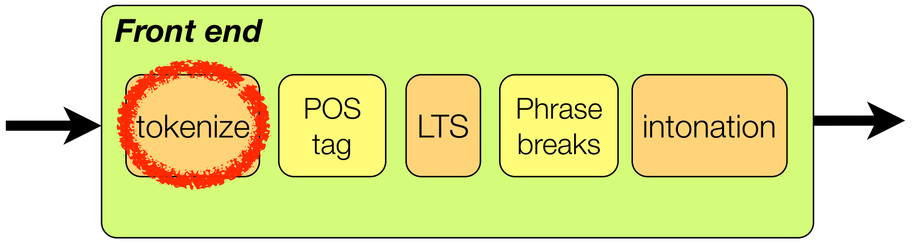


- Step 2: classify every token, finding **Non-Standard Words** that need further processing

```
In 2011, I spent £100 at IKEA on 100 DVD holders.

   NYER                MONEY   ASWD    NUM LSEQ
```

Tokenize & Normalize



- Step 3: a set of specialised modules to process NSWs of a each type

2011 ⇨ NYER ⇨ twenty eleven

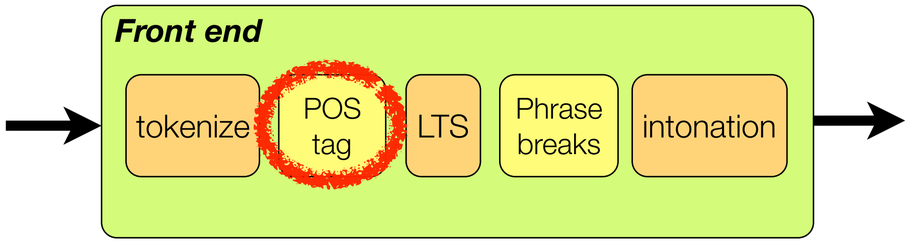
£100 ⇨ MONEY ⇨ one hundred pounds

IKEA ⇨ ASWD ⇨ *apply letter-to-sound*

100 ⇨ NUM ⇨ one hundred

DVD ⇨ LSEQ ⇨ D. V. D. ⇨ dee vee dee

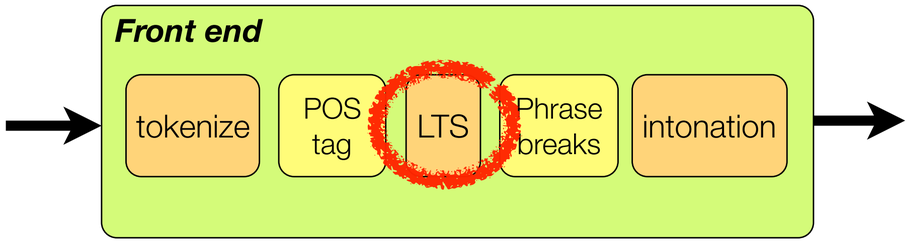
POS tagging



- Part-of-speech tagger
- Accuracy can be very high
- Trained on **annotated** text data
- **Categories** are designed for text, not speech

```
NN  Director
IN  of
DT  the
NP  McCormick
NP  Public
NPS Affairs
NP  Institute
IN  at
NP  U-Mass
NP  Boston,
NP  Doctor
NP  Ed
NP  Beard,
VBZ says
DT  the
NN  push
IN  for
VBP do
PP  it
PP  yourself
NN  lawmaking
```

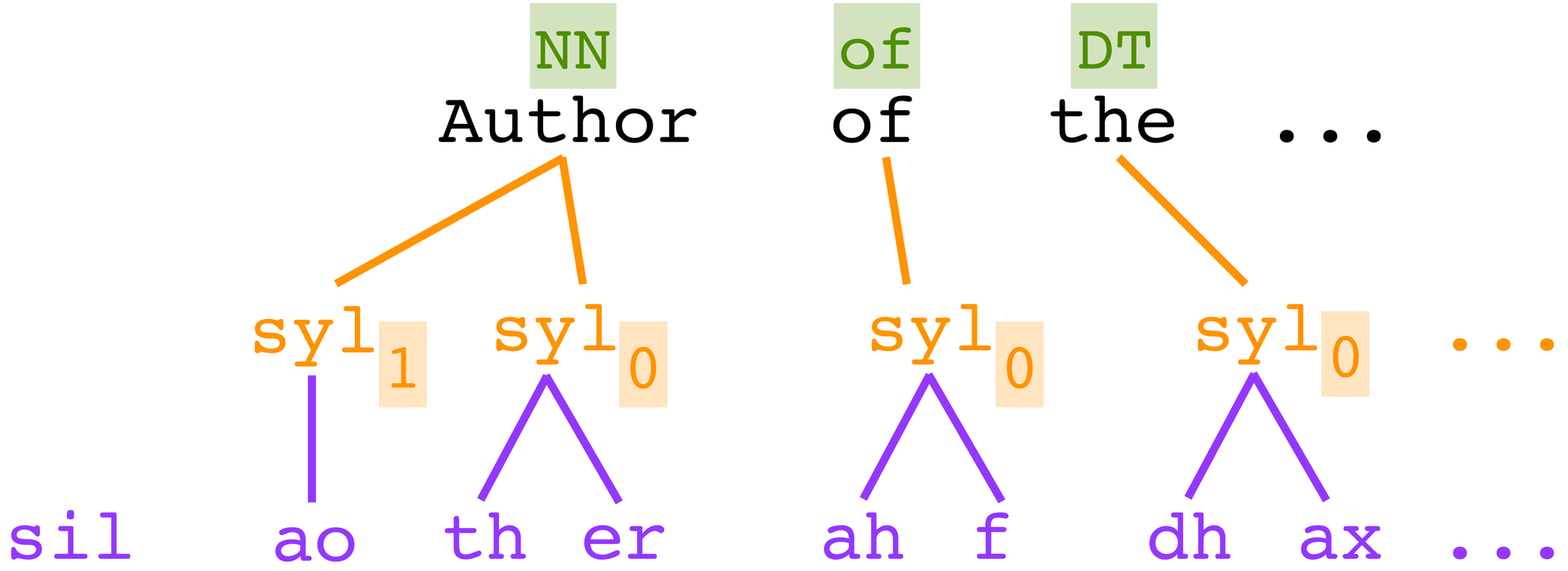
Pronunciation



- Pronunciation model
  - dictionary look-up, *plus*
  - letter-to-sound model
- But
  - need deep **knowledge** of the language to design the phoneme set
  - human **expert** must write dictionary

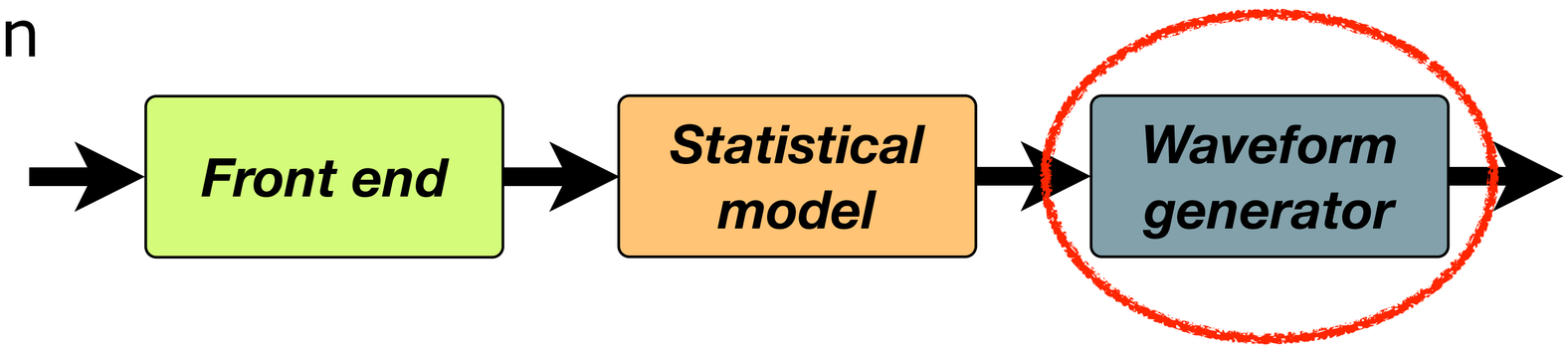
```
ADVOCATING  AE1 D V AH0 K EY2 T IH0 NG
ADVOCATION  AE2 D V AH0 K EY1 SH AH0 N
ADWEEK  AE1 D W IY0 K
ADWELL  AH0 D W EH1 L
ADY  EY1 D IY0
ADZ  AE1 D Z
AE  EY1
AEGEAN  IH0 JH IY1 AH0 N
AEGIS  IY1 JH AH0 S
AEGON  EY1 G AA0 N
AELTUS  AE1 L T AH0 S
AENEAS  AE1 N IY0 AH0 S
AENEID  AH0 N IY1 IH0 D
AEQUITRON  EY1 K W IH0 T R AA0 N
AER  EH1 R
AERIAL  EH1 R IY0 AH0 L
AERIALS  EH1 R IY0 AH0 L Z
AERIE  EH1 R IY0
AERIEN  EH1 R IY0 AH0 N
AERIENS  EH1 R IY0 AH0 N Z
AERITALIA  EH2 R IH0 T AE1 L Y AH0
AERO  EH1 R OW0
```

The linguistic specification

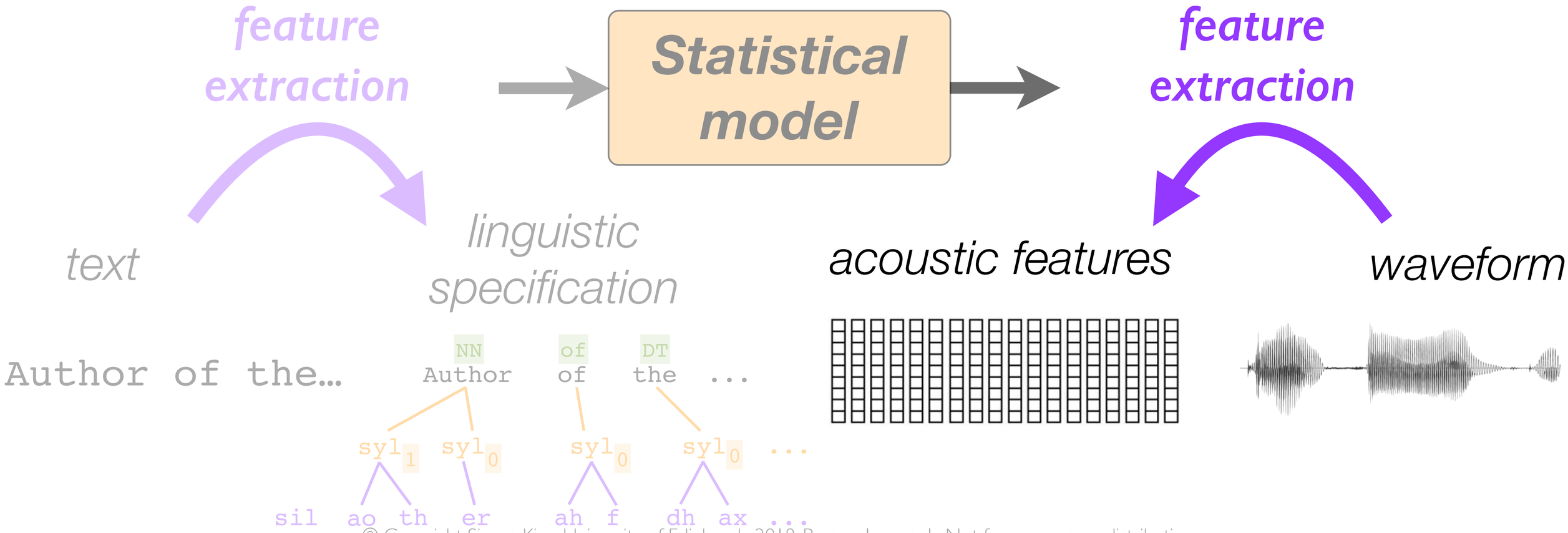


From text to speech

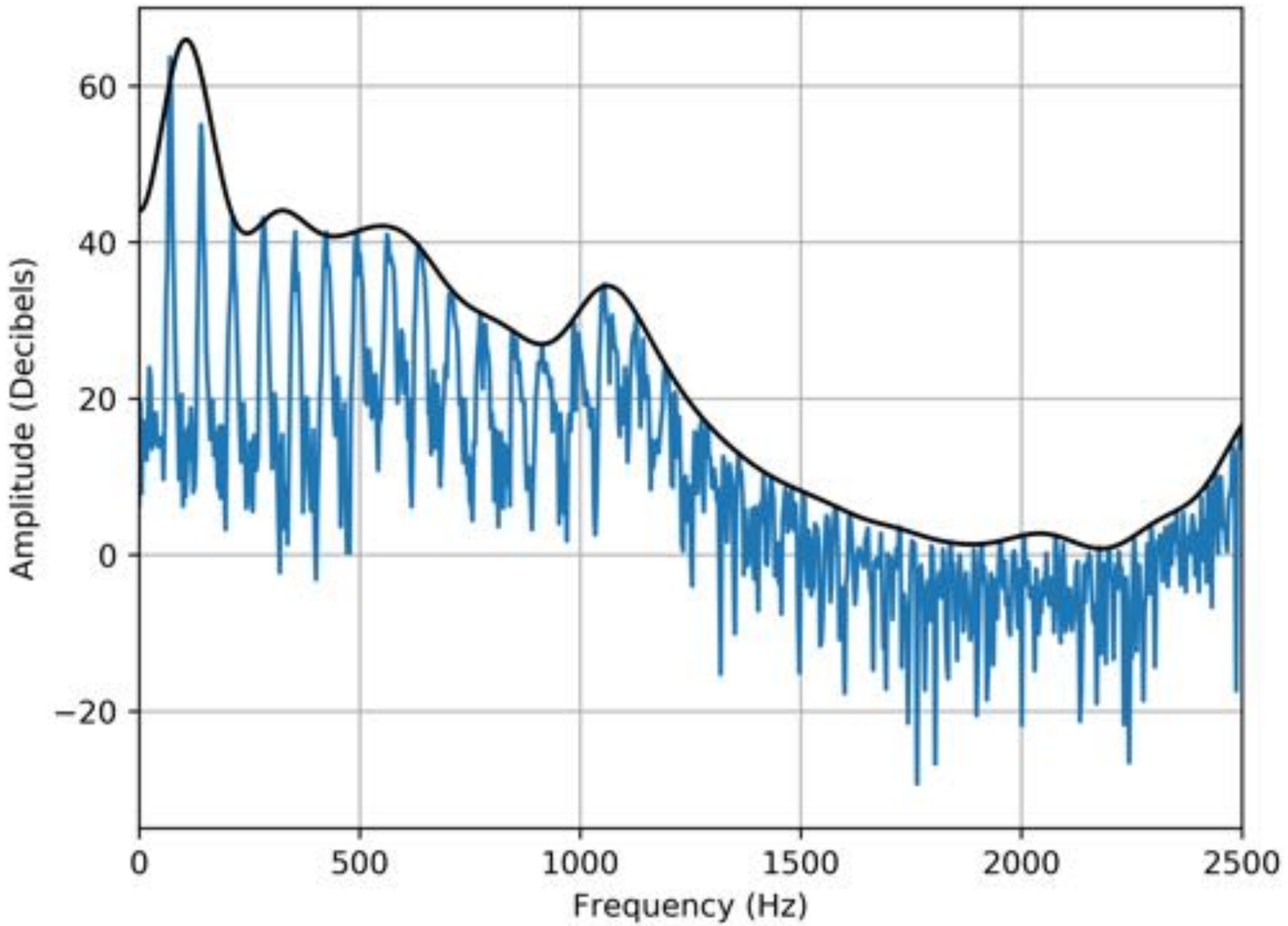
- Text processing
  - pipeline architecture
  - linguistic specification
- Modelling
  - duration model
  - acoustic model
- Waveform generation
  - acoustic features
  - signal processing

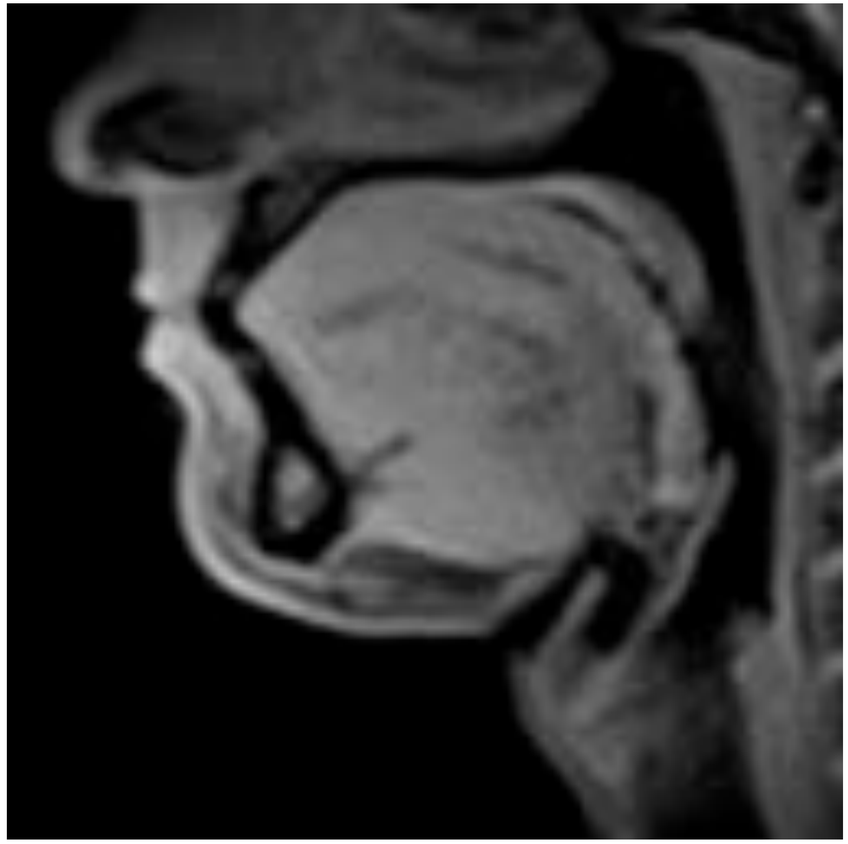


Acoustic feature extraction



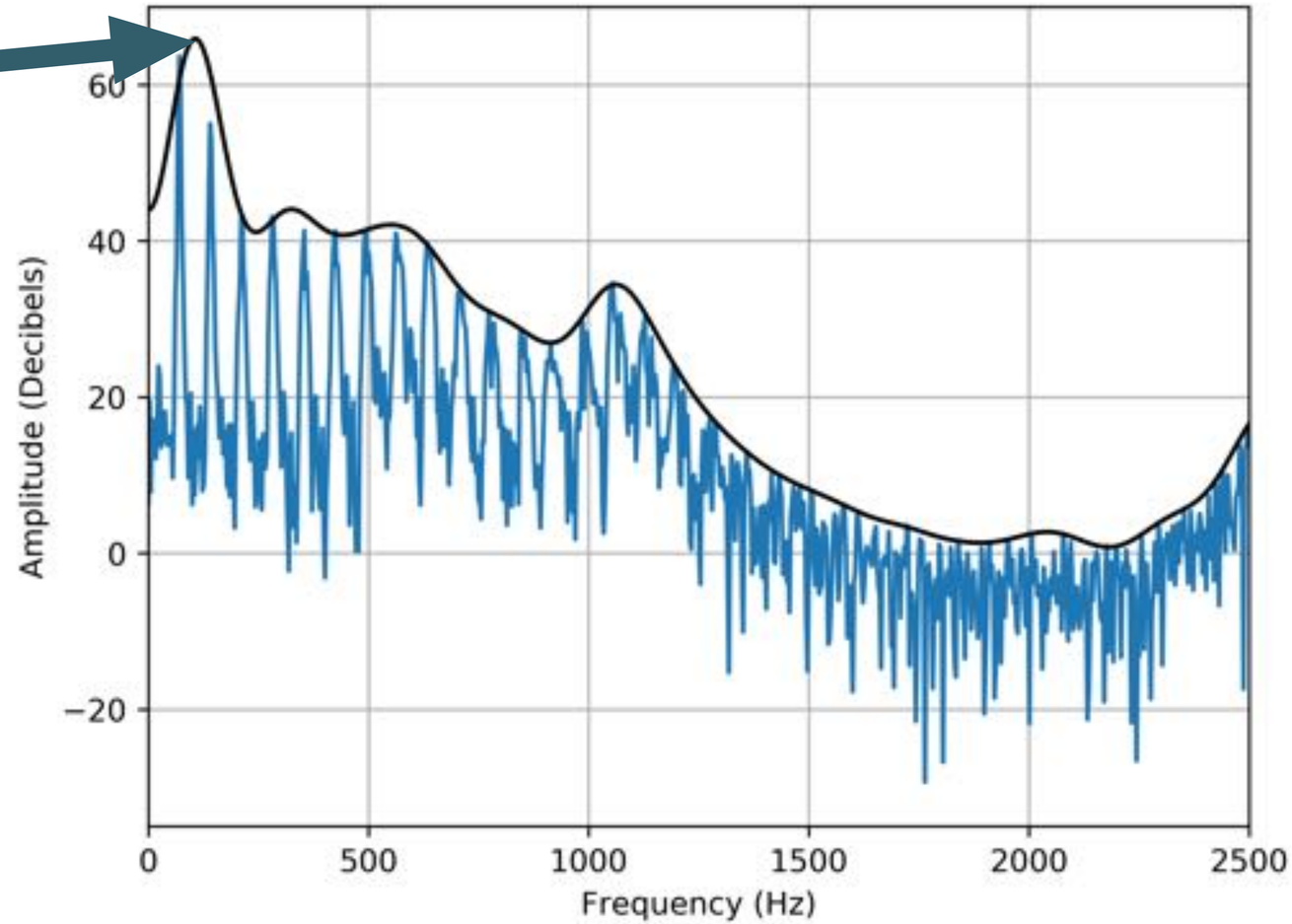
Acoustic features: motivated by speech production





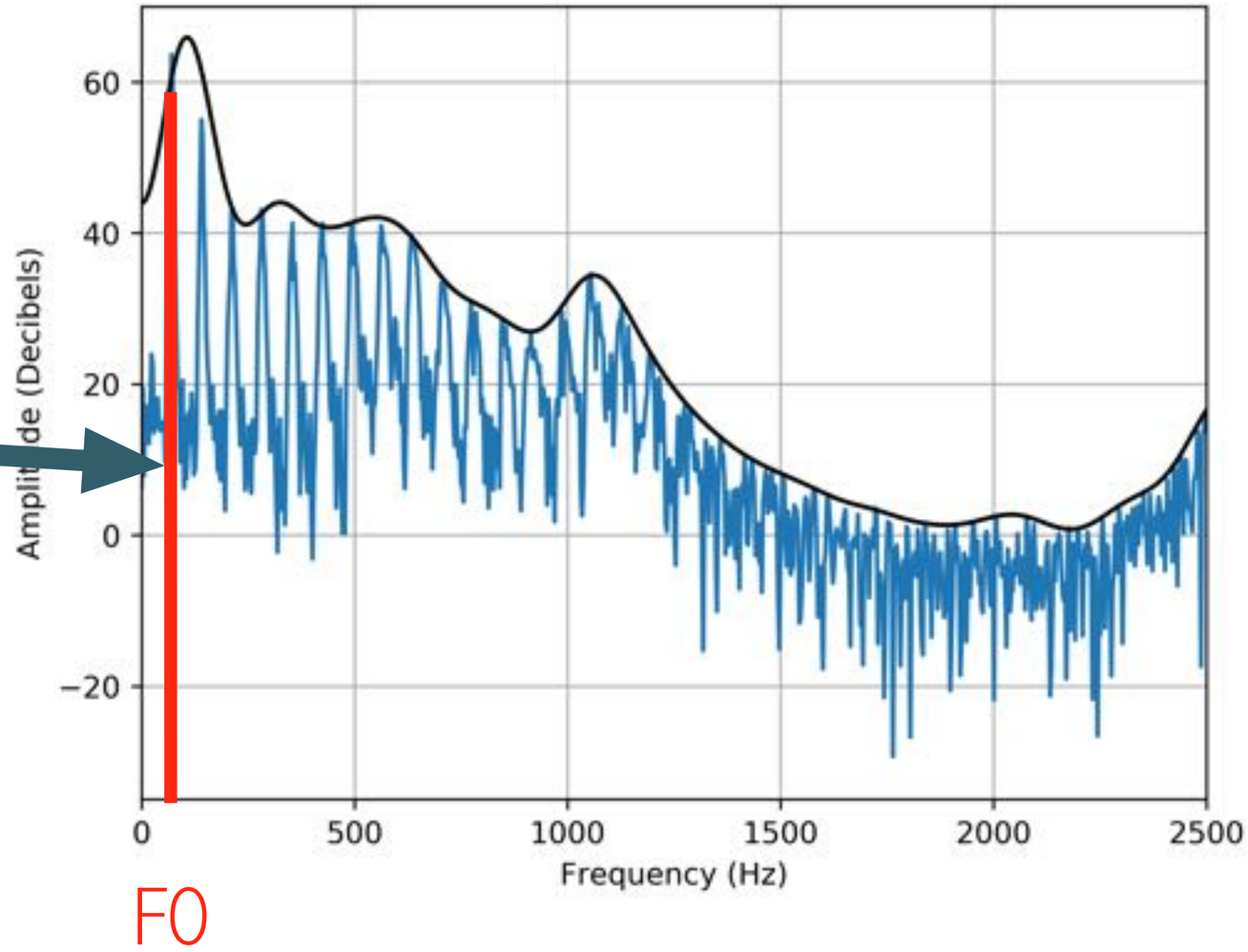
Acoustic features

- Spectral Envelope
- F0
- Aperiodic energy



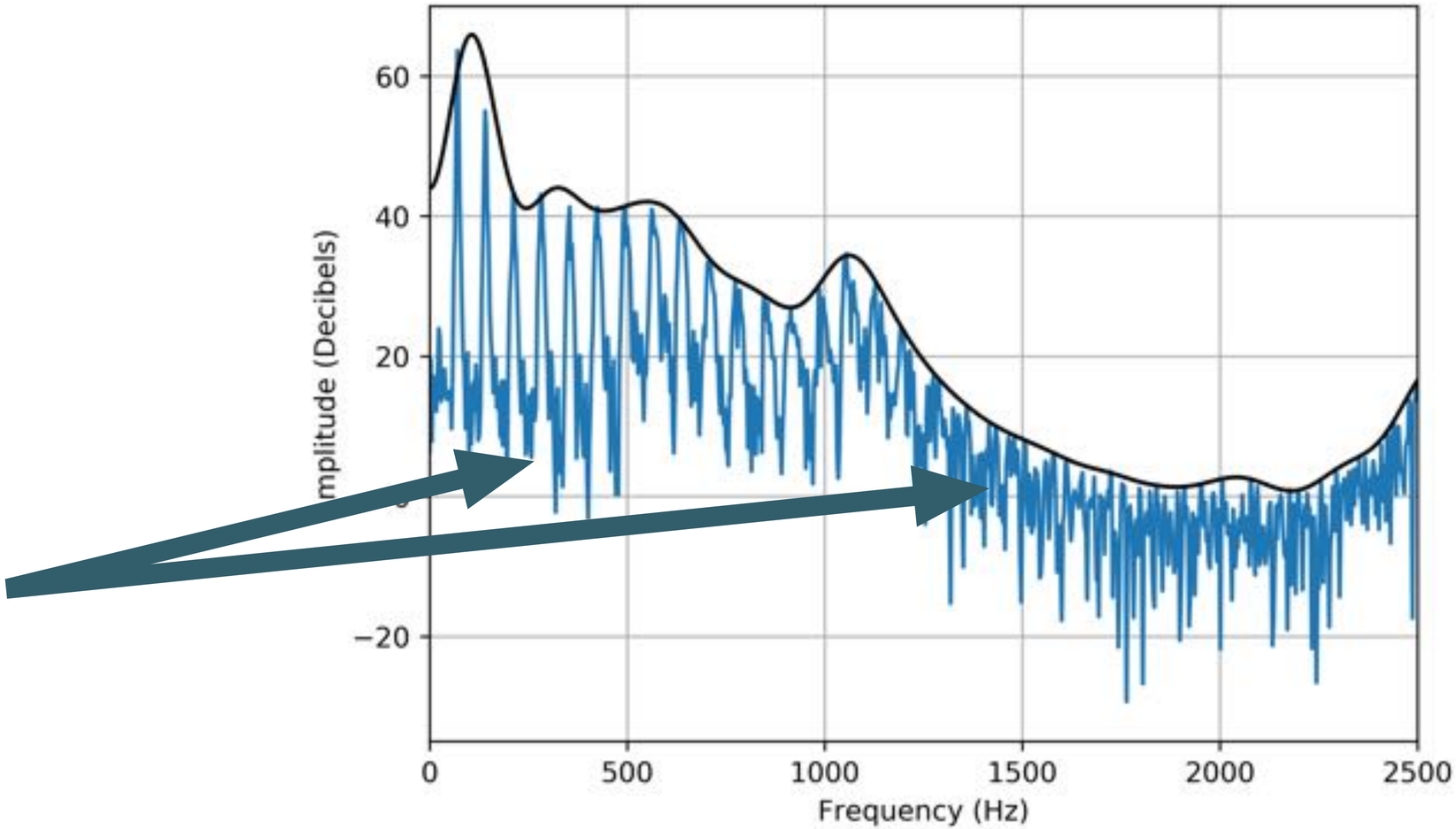
Acoustic features

- Spectral Envelope
- F0
- Aperiodic energy



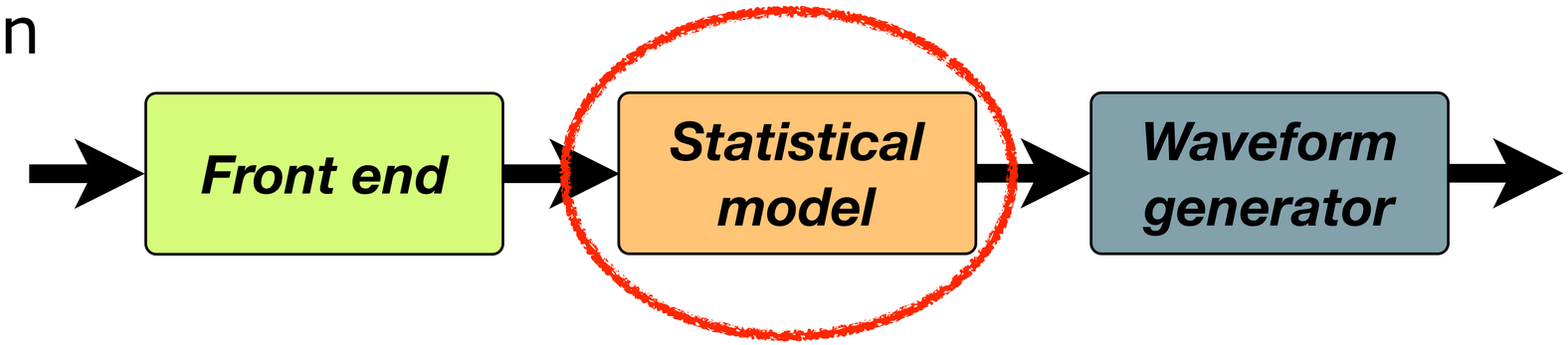
Acoustic features

- Spectral Envelope
- F0
- Aperiodic energy

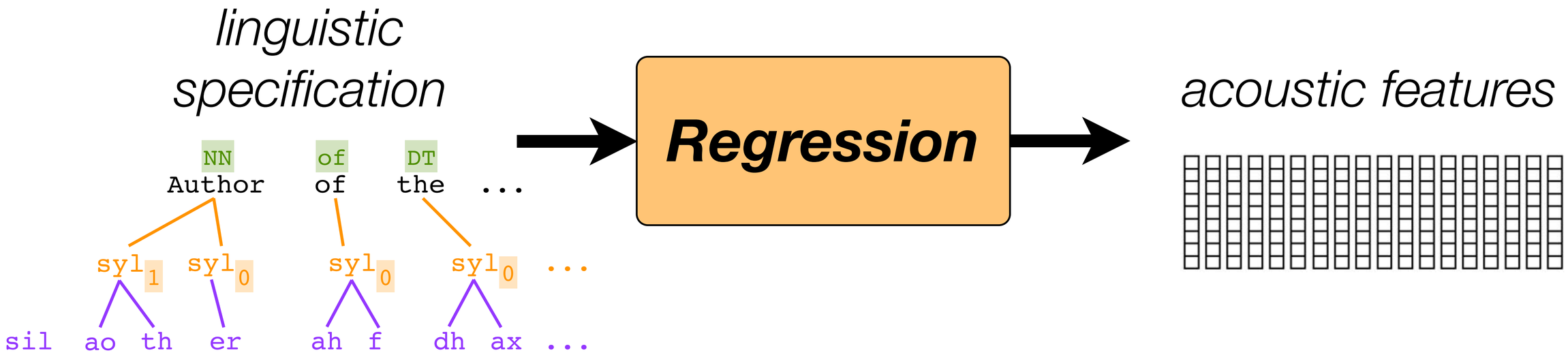


From text to speech

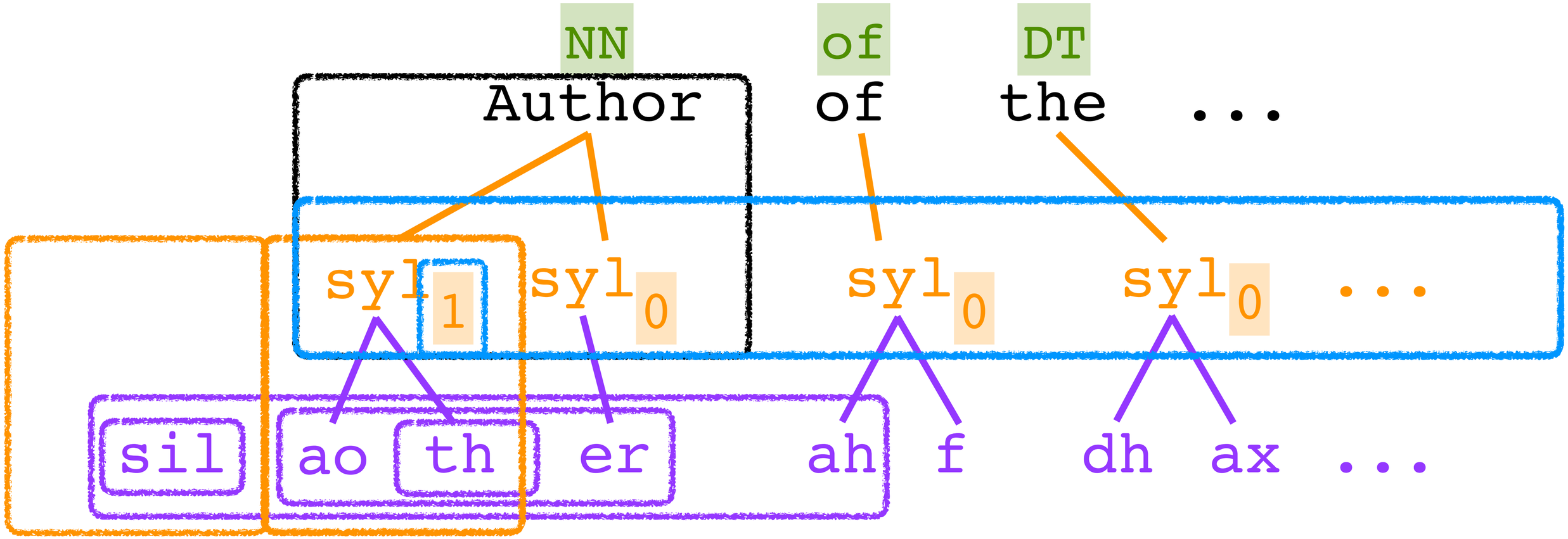
- Text processing
  - pipeline architecture
  - linguistic specification
- Modelling
  - duration model
  - acoustic model
- Waveform generation
  - acoustic features
  - signal processing



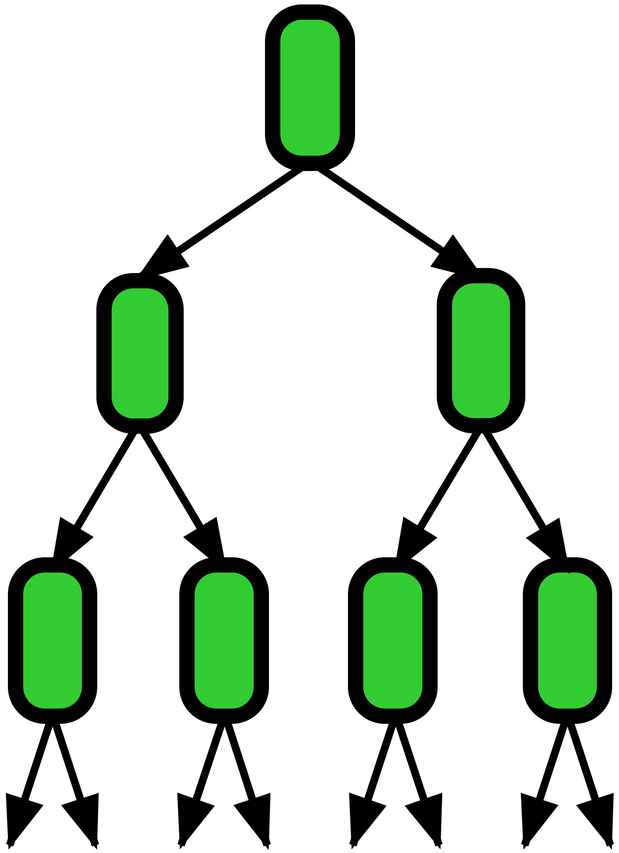
A regression model predicts the acoustic features

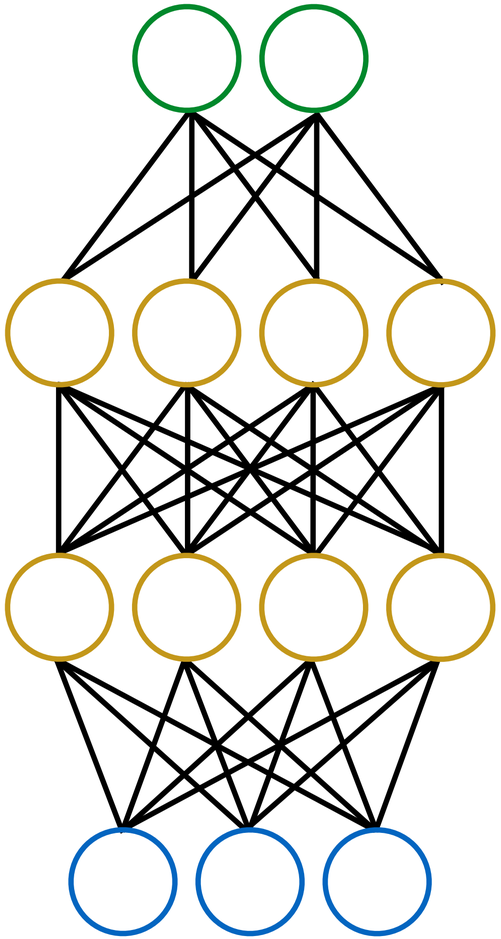


The input to the regression model

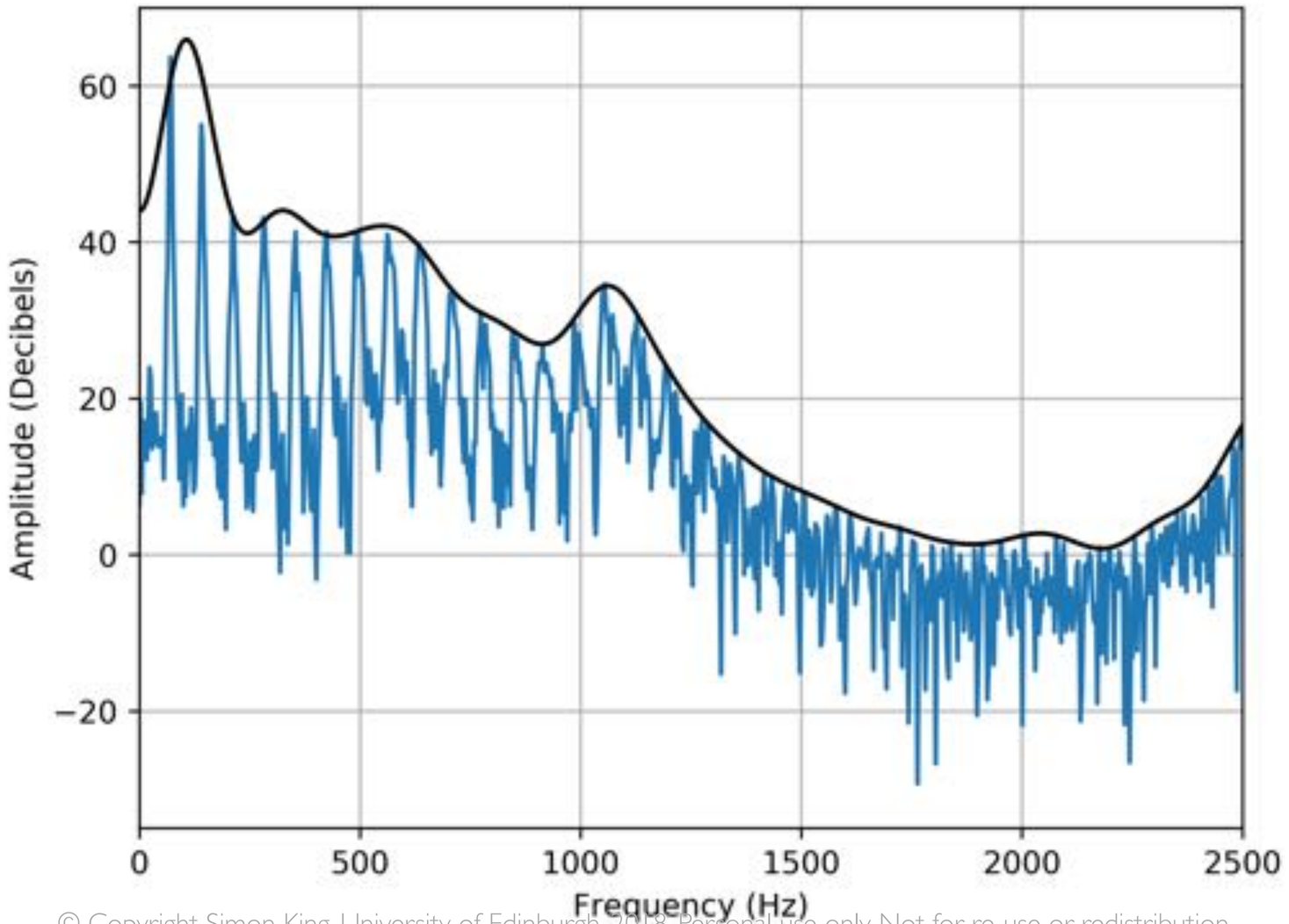


Choices for the regression model





Acoustic features can be modelled **separately**



Repairing voices

Identifying the problems

Borrowing from healthy voices

Identifying problems with disordered speech

Speech & language therapy

- already part of the patient journey
- standard screening tests

- identify problems with
  - articulation, perhaps only of some sounds
  - duration
  - fundamental frequency

Screening test example: plosives

TARGET	REALISATION
1. **P**ink so**d**a tastes goo**d**	_ink so_a tastes goo_
2.**B**ig hi**pp**o wee**k**	_ig hi_o wee_
3.**G**olden ro**ck**et shi**p**	_olden ro_et shi_
4.**D**octor Mar**t**in is la**t**e	_octor Mar_in is la_e
5.**T**urn the cup**b**oard kno**b**	_urn the cup_oard kno_
6.**C**old so**gg**y do**g**	_old so_y do_

Screening test example: clusters

TARGET	REALISATION
1. The **br**ave **gr**een **fr**og **squ**eaked	The _ave _een _og _eaked
2. I **sp**y a **bl**ue **fl**y in the **sk**y	I _y a _ue _y in the _y
3. **Thr**ee **sw**eet **sm**elling **pl**ums	_ee _eet _elling _ums
4. **St**ock market **cr**ash **dr**ives up **pr**ices	_ock market _ash _ives up _ices
5. **Gl**asgow **sn**ow **sl**ows **tr**avellers	_asgow _ow _ows _avellers
6. A **qu**ick **spr**ing **cl**ean **scr**ub	A _ick _ing _ean _ub

The voicebank

A source of healthy voices

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colours. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow. Throughout the centuries people have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews it was a token that there would be no more universal floods. The Greeks used to imagine that it was a sign from the gods to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge over which the gods passed from earth to their home in the sky. Others have tried to explain the phenomenon physically. Aristotle thought that the rainbow

ed we

pares b
ar enrichm

second this month — is another attempt to break more than three years of Iranian stonewalling about allegations that Tehran is secretly working on nuclear weapons that would be armed with uranium enriched to 90 percent or more.

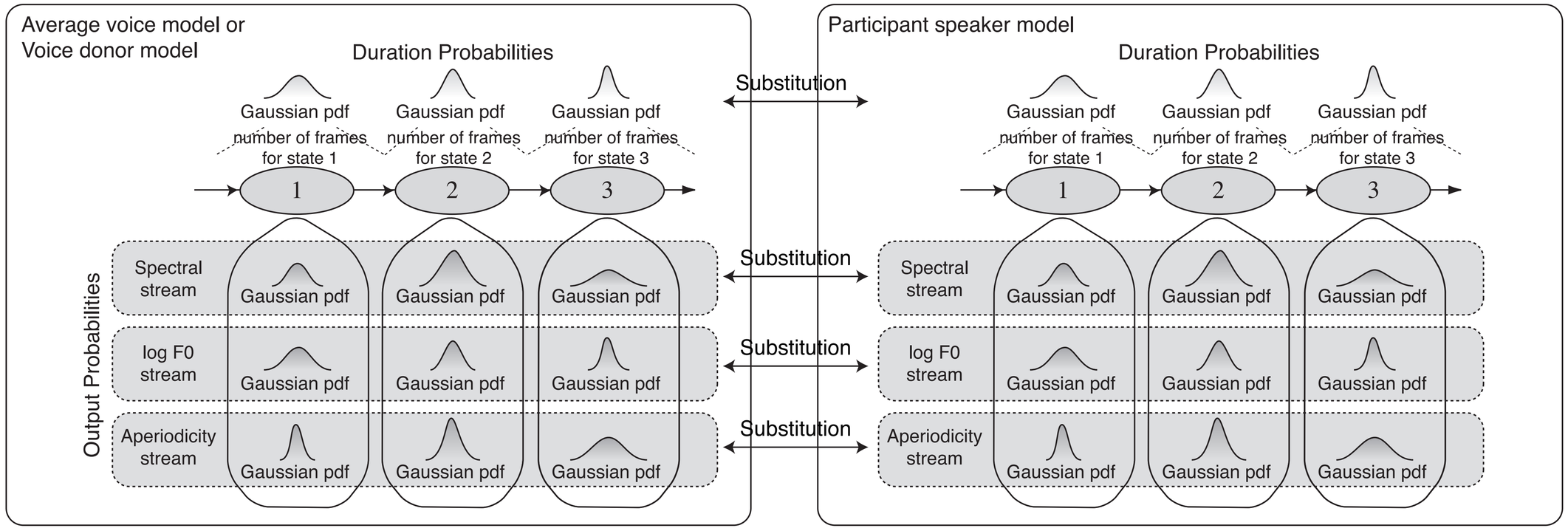
Diplomats accredited to the IAEA expected little from that visit. They said Iran was refusing to allow the agency experts to visit Parchin, the suspected site for a nuclear explosives testing and had turned down

siety escalating international nctions and the latent threat an Israeli military strike on its uclear facilities.

Forde could be used to make ile warhead material even r such an upgrade, the ld. although old-generation new

Voice repair by using healthy voices

Using an Average Voice Model to perform voice repair



Average voice model (AVM)
*or an individual voice **donor** model*

Voice **clone** of patient
(an AVM fully adapted to their speech)

Example of parameter substitution

- **original recording**
- **speaker adapted voice (voice clone)**
- **s1:** duration + aperiodicity + GV model (aperiodicity)
- **s2:** s1 + logf0 (dynamic features, variance, V/UV weights) + GV (logf0)
- **s3:** s1 + mcep (excluding low-order static coefficients) + GV (mcep)
- **s4:** s2 + mcep (excluding low-order static coefficients) + GV (mcep)
- **accent specific average voice model**

- We can still hear problems of coarticulation
  - bad coarticulation of approximants (“reconstruction”)

Open questions & challenges

Accents

Better modelling

Personalised text processing

Accent

Short vowels		*Long vowels*		*Rising diphthongs*	
KIT	/ɪ/	FLEECE	/iː/	PRICE/PRIDE	/ai/
DRESS	/e/	FACE	/eː/	MOUTH	/au/
TRAP	/æ/	BATH	/aː/	CHOICE	/ɒi/
LOT	/ɒ/	THOUGHT	/ɒː/	GOAT	/oʊ/
STRUT	/ʌ/	SOFT	/ɒ(ː)/		
FOOT	/ʊ/	GOOSE	/ʉː/		

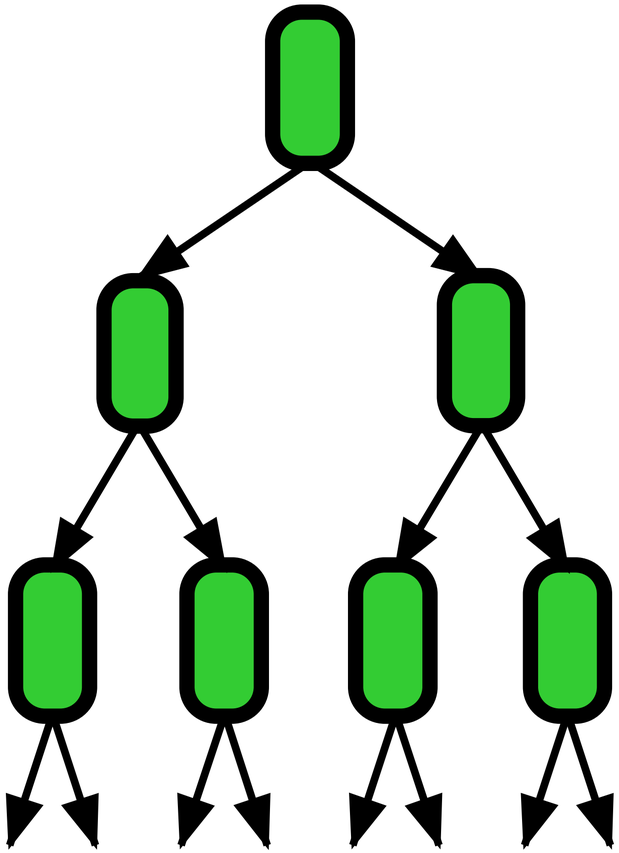
Centring diphthongs / rhotacised vowels; unstressed vowels					
NEAR	/iər/	SQUARE	/eər/	CURE	/ʉər/
START	/ɑɪ/	NORTH	/ɒɪ/	FORCE	/oɪ/
NURSE	/ər/	TERM	/ər/	LETTER	/-ər/
COMMA	/-ə/	HAPPY	/-i/		

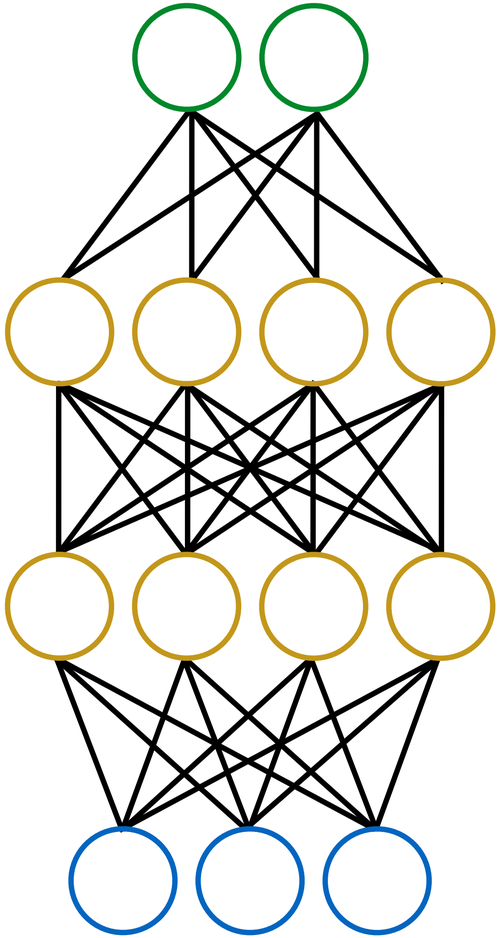
Low vowels before (i) nasal + obstruent, (ii) voiceless fricatives			
DANCE	/æ/	PATH	/aː/

Image credit: Sima Brankov, on sblanguagemaps.wordpress.com

Image credit: Manchester Metropolitan University

Modelling





Personalised text processing

www.cstr.ed.ac.uk

www.speech.zone

www.speakunique.org

THE UNIVERSITY *of* EDINBURGH