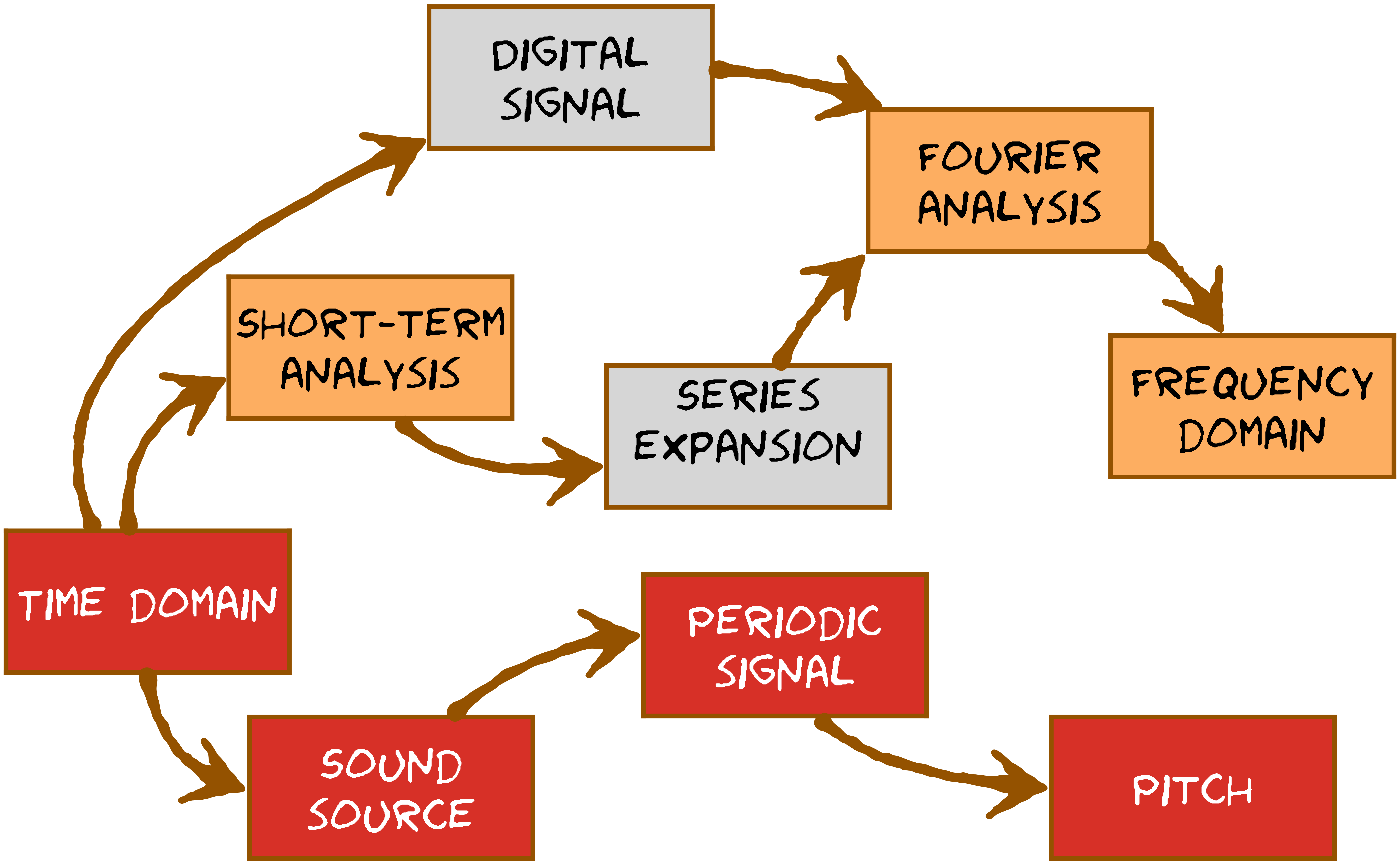
Speech Processing - modules 1 to 5

slides for topic videos on **speech.zone**

Module 1

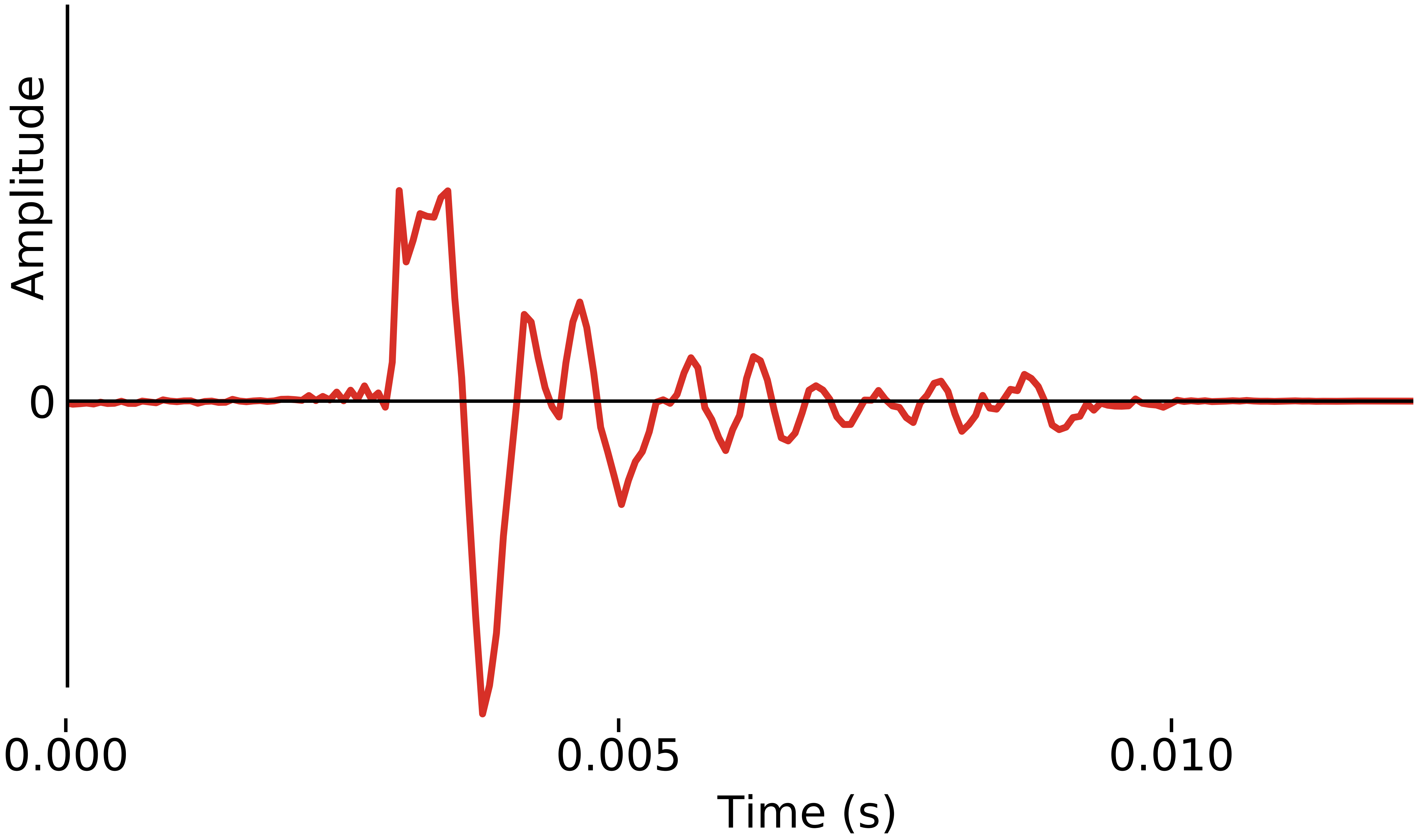
Sound


DIGITAL SIGNAL
FOURIER ANALYSIS
SHORT-TERM ANALYSIS
SERIES EXPANSION
FREQUENCY DOMAIN
TIME DOMAIN
PERIODIC SIGNAL
SOUND SOURCE
PITCH

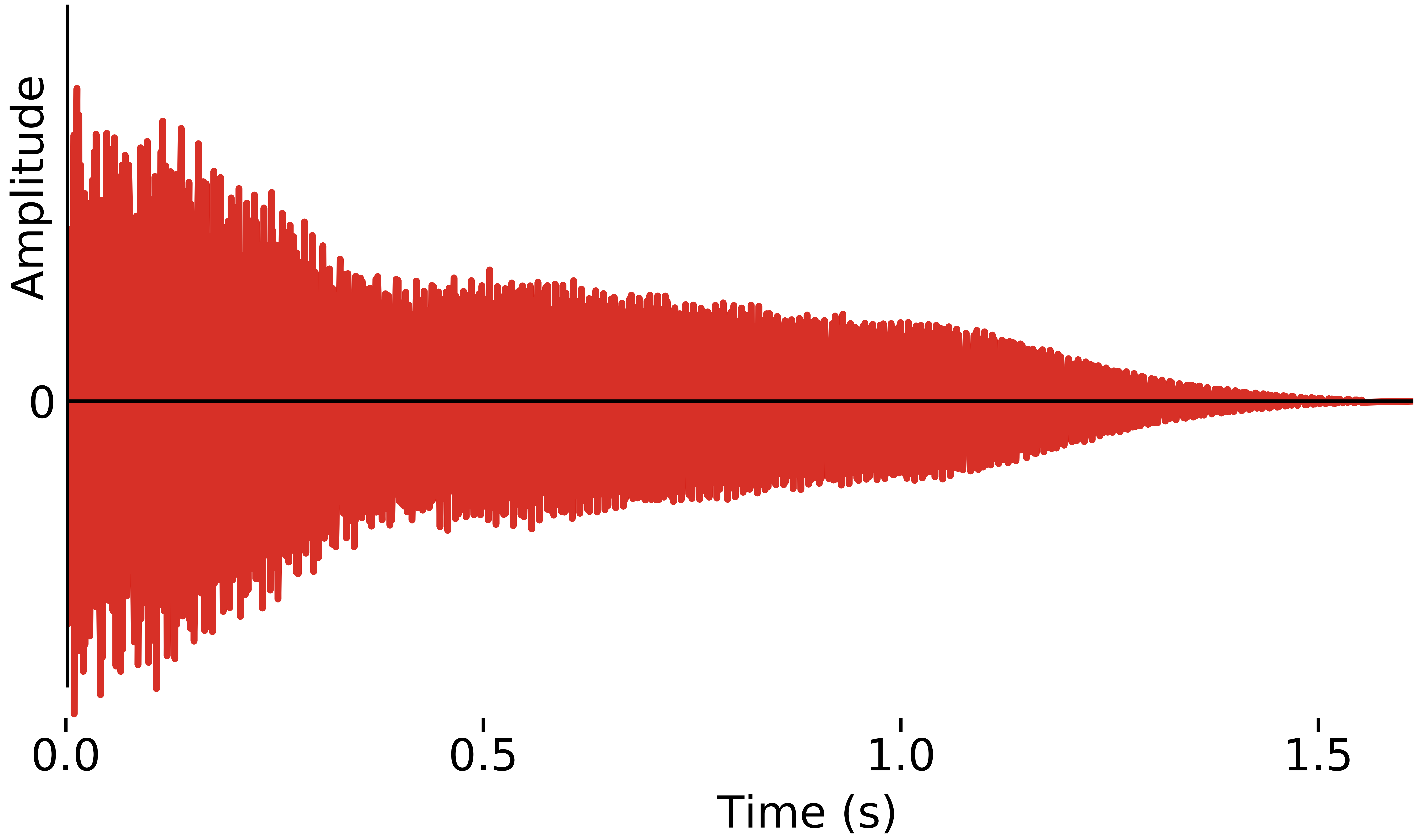
TIME DOMAIN

Sound

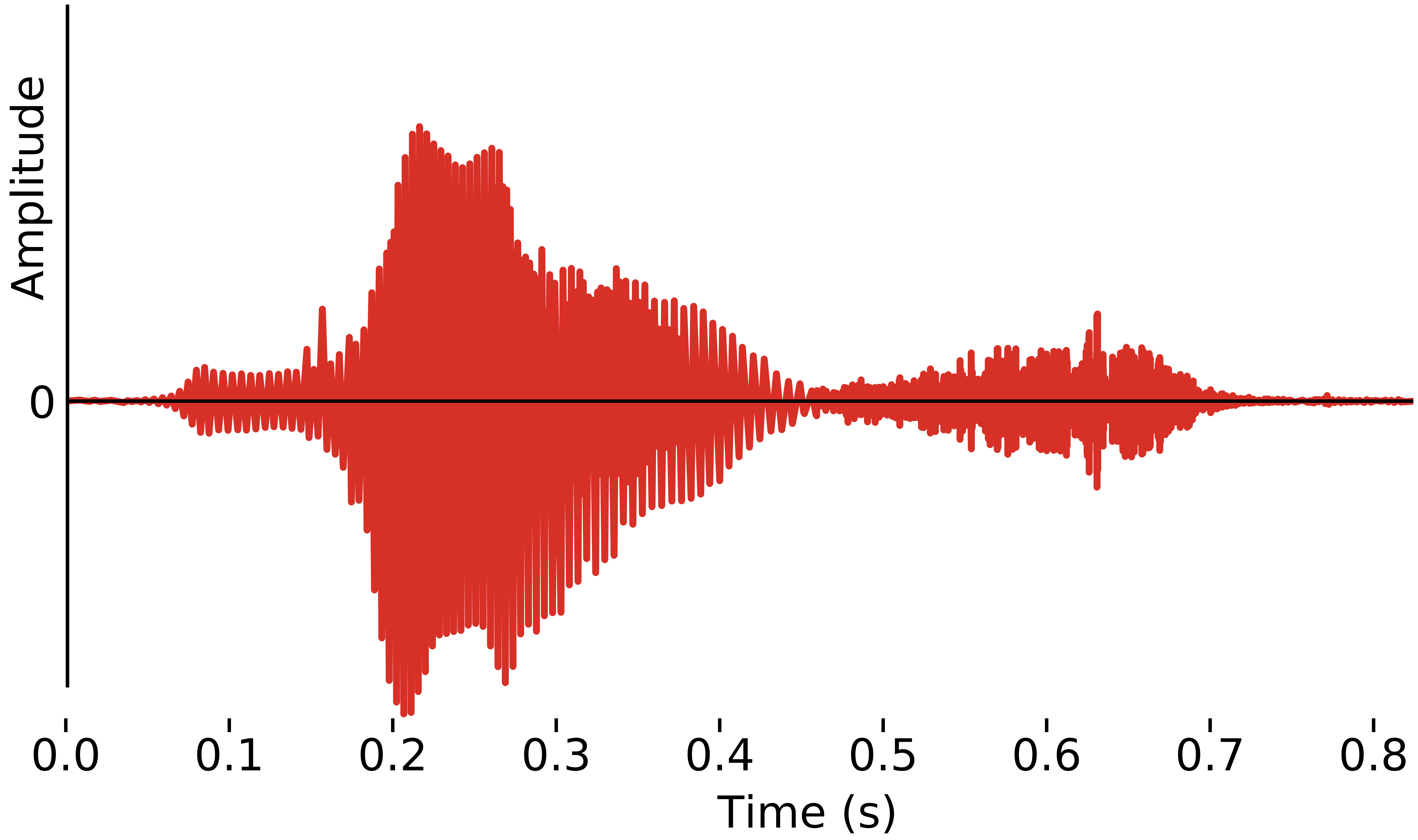
Waveform



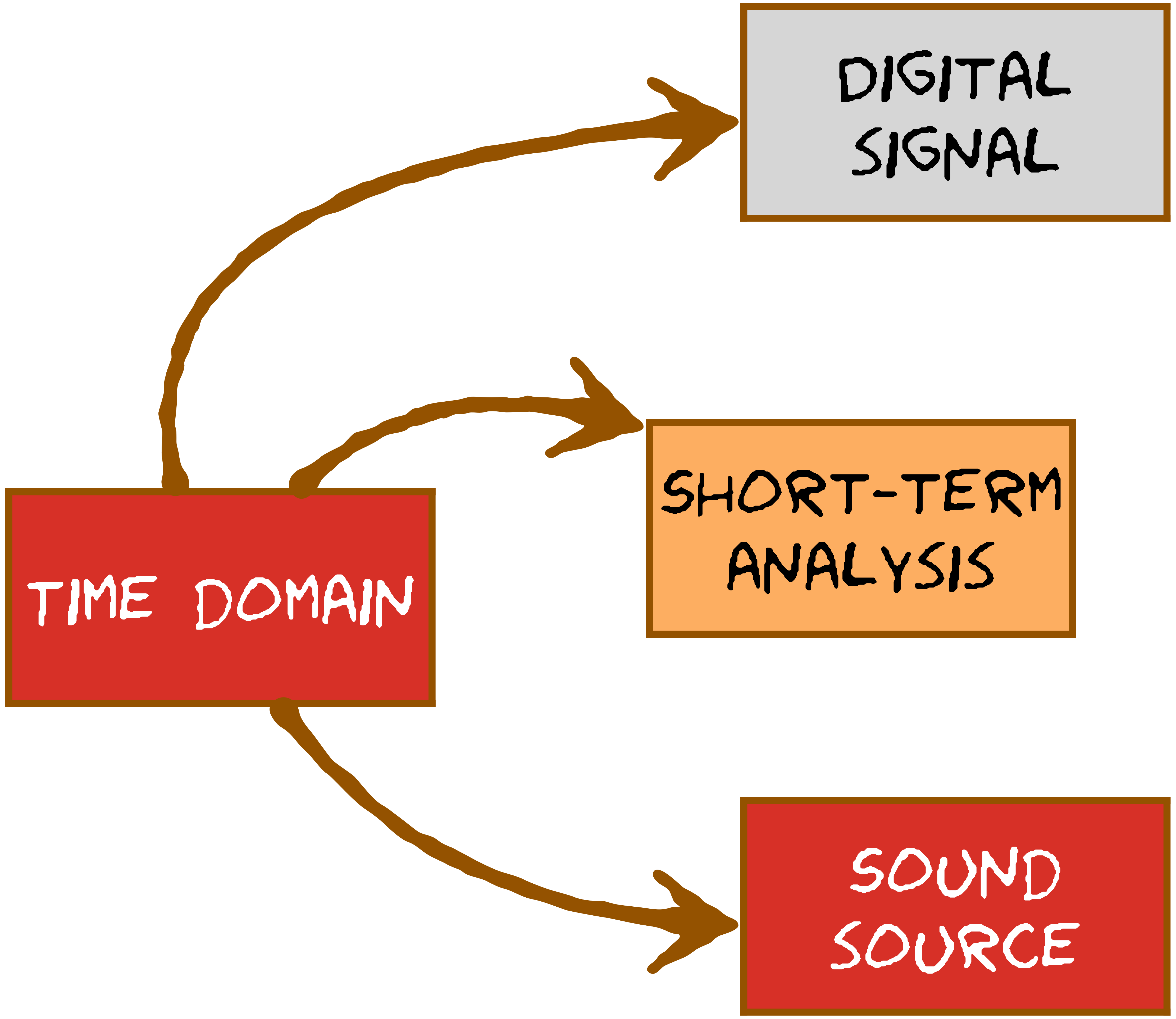
Waveform



Waveform

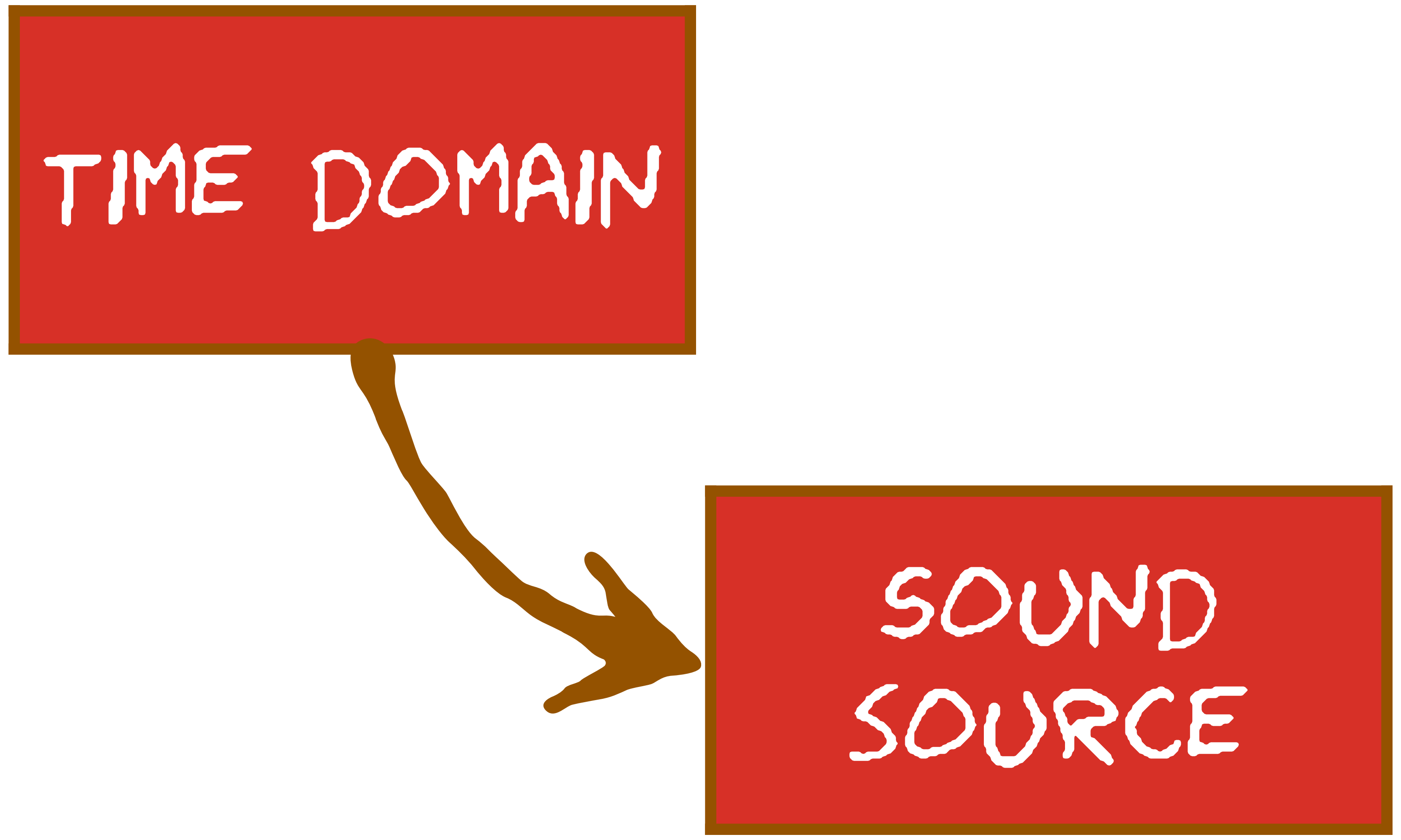


What you can learn next

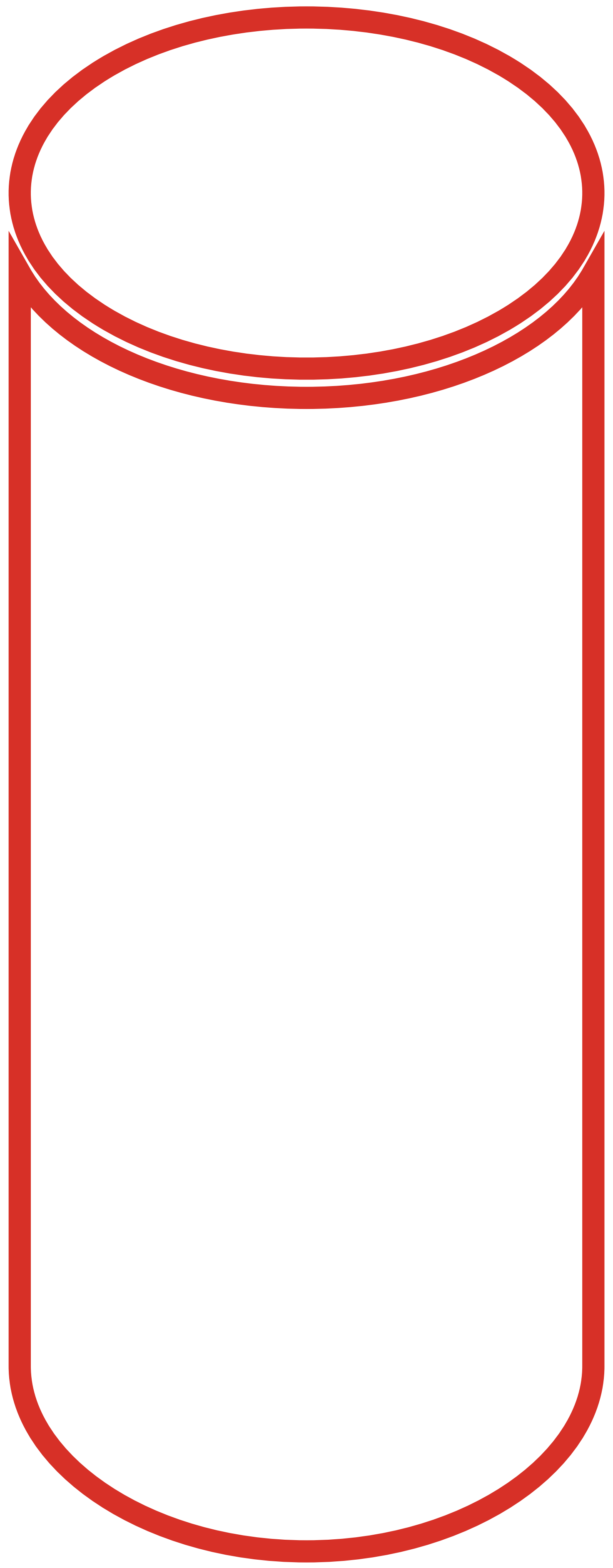


SOUND SOURCE

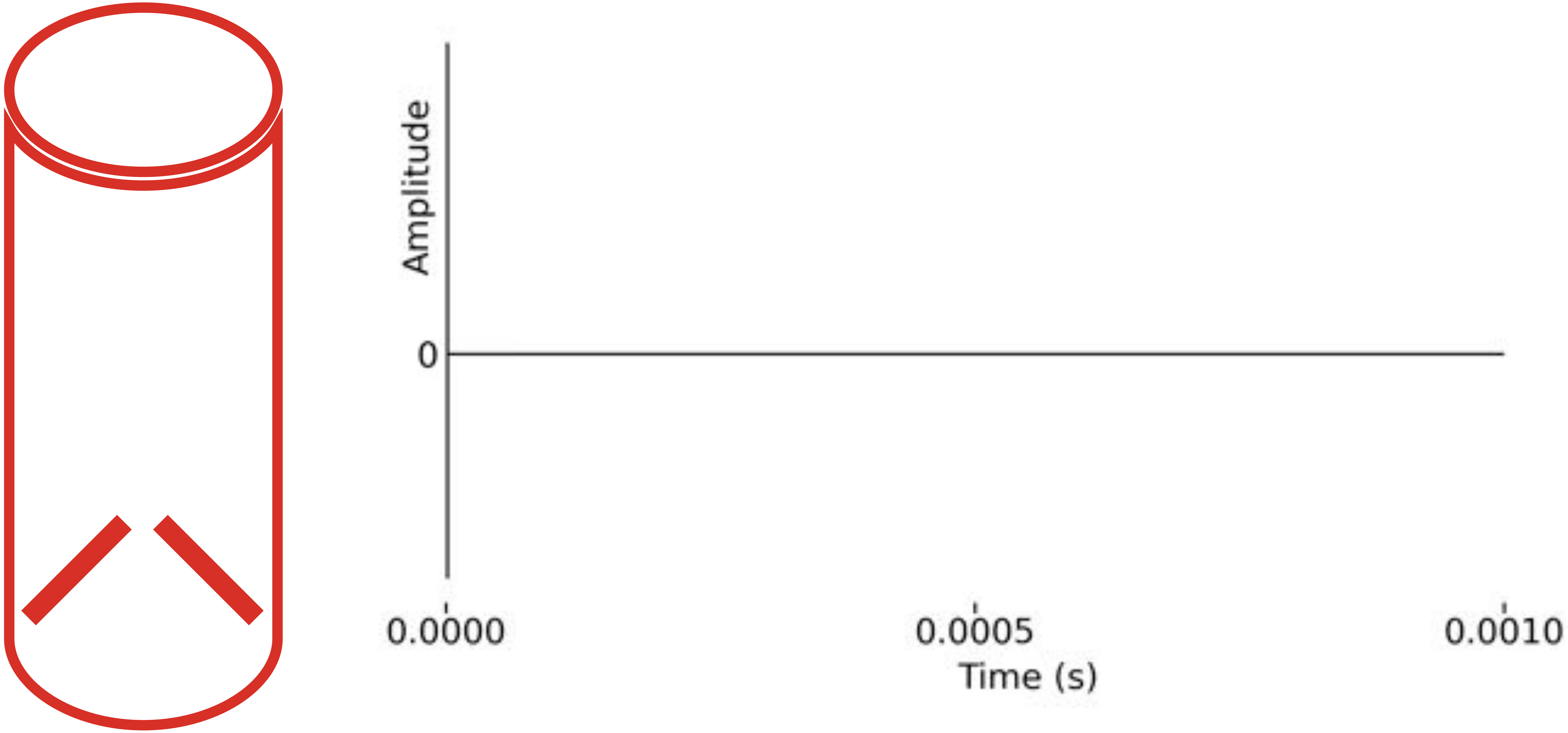
What you need to know already



The vocal tract

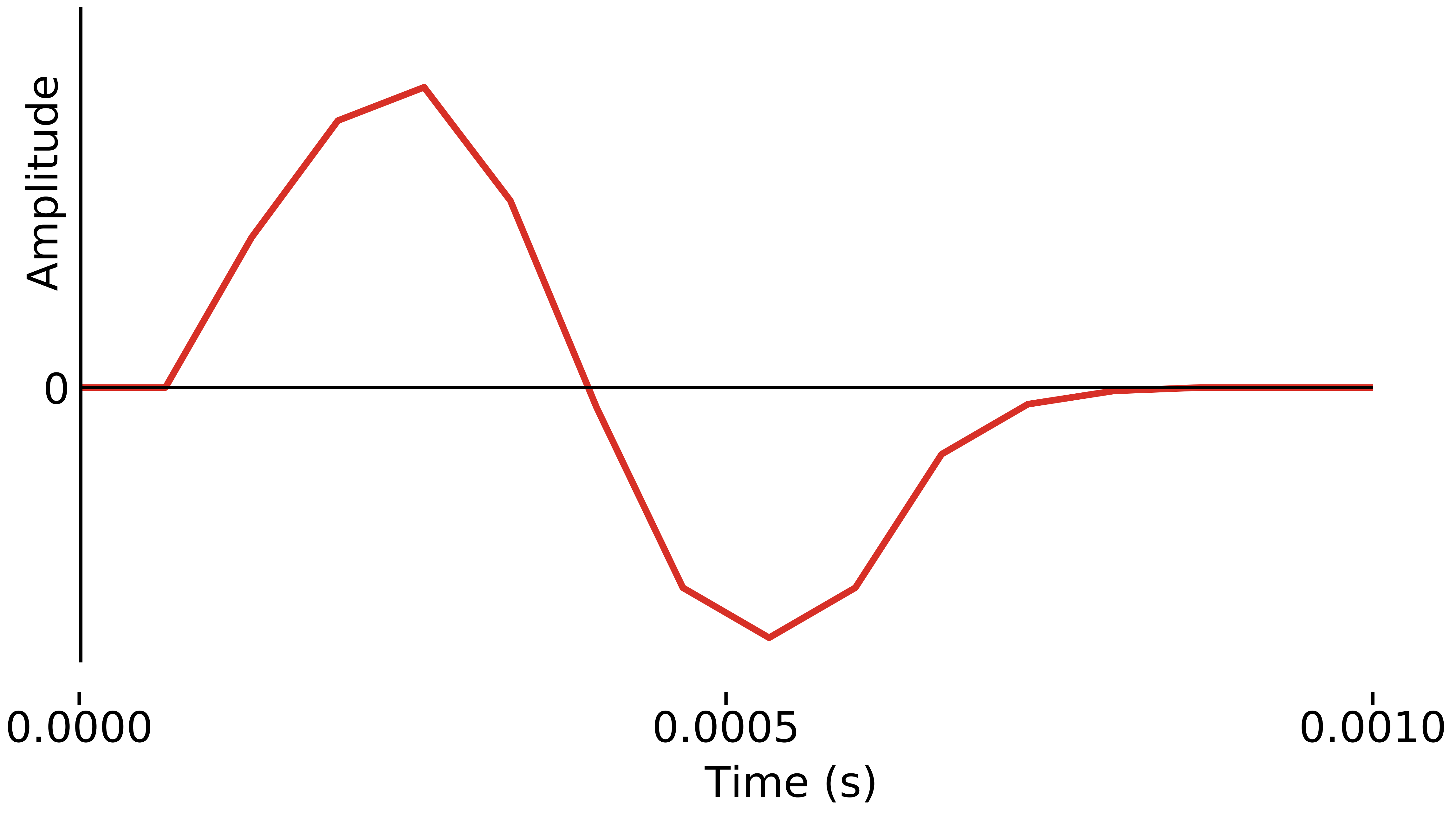


Voicing: the vocal folds

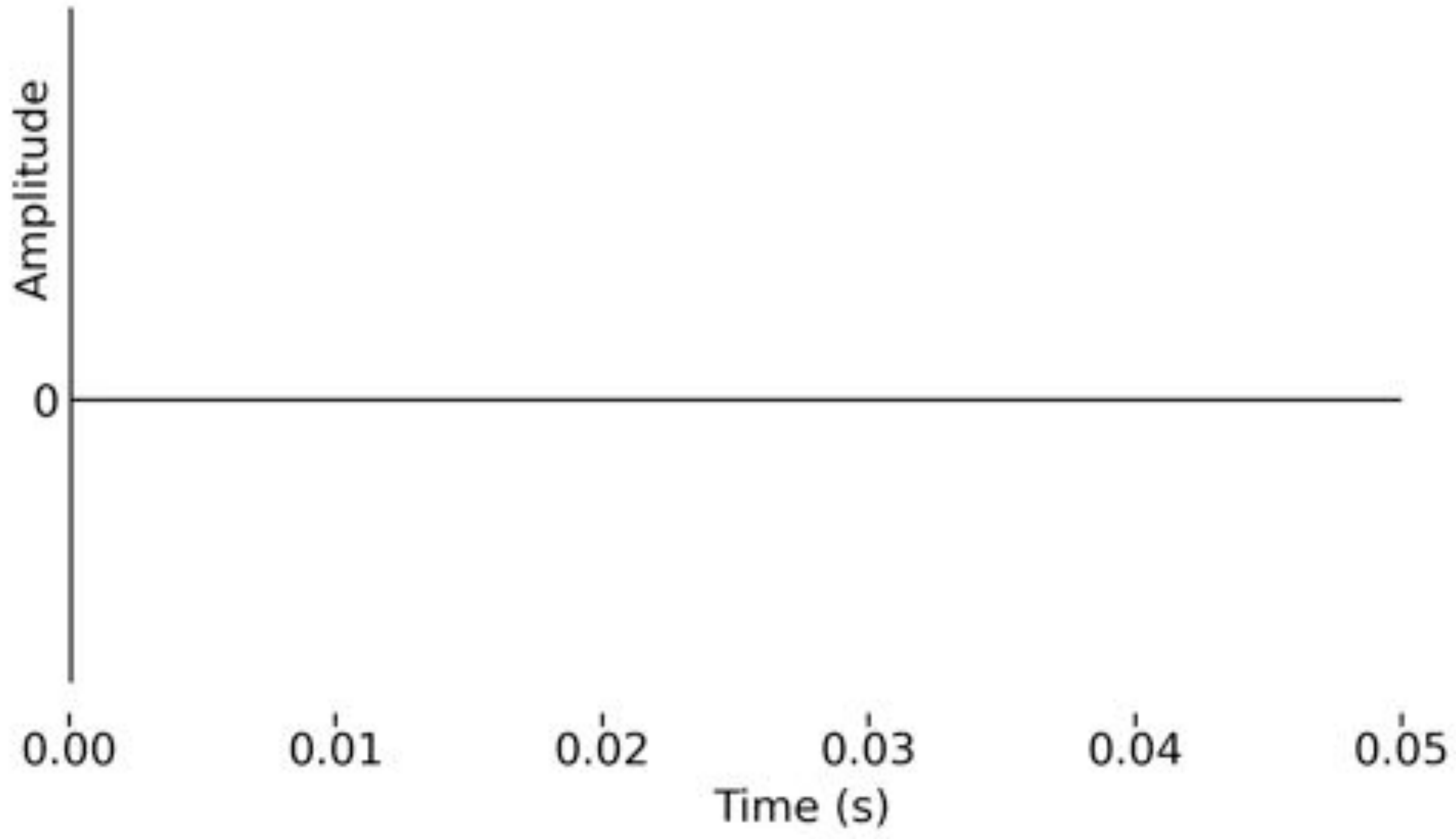


Glottal pulse

Individual pulse

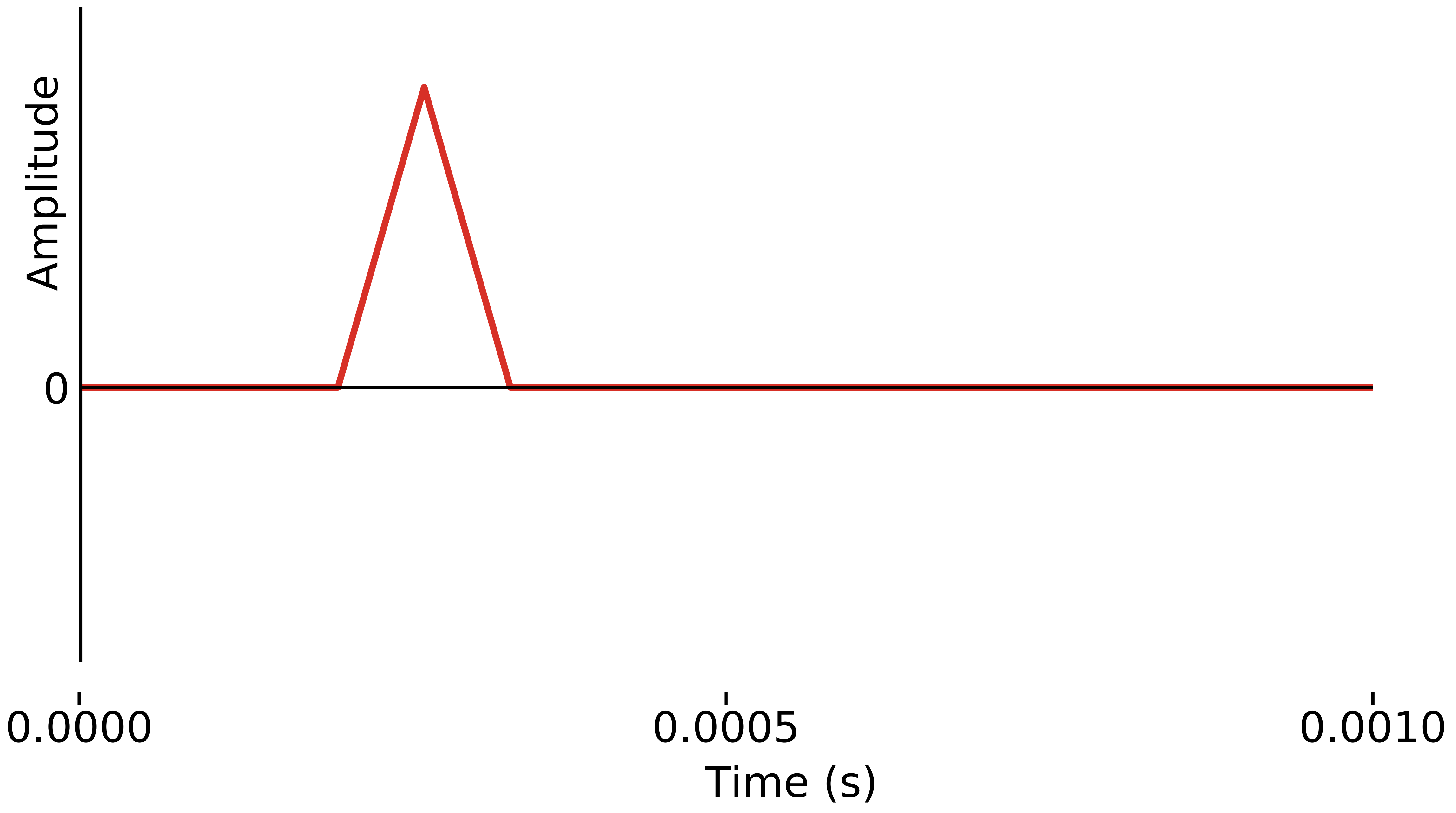


Pulse train

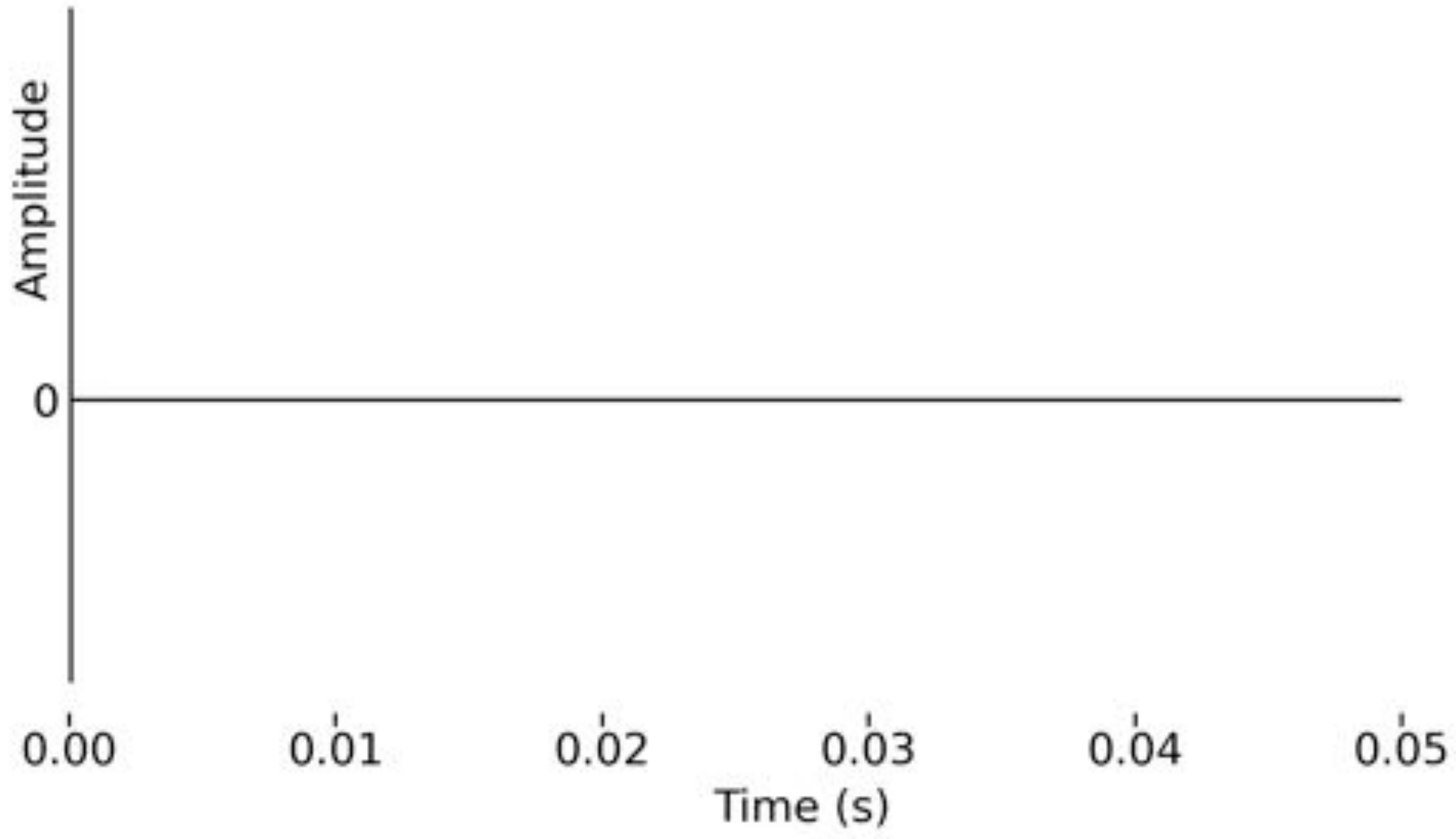


Simplified glottal pulse

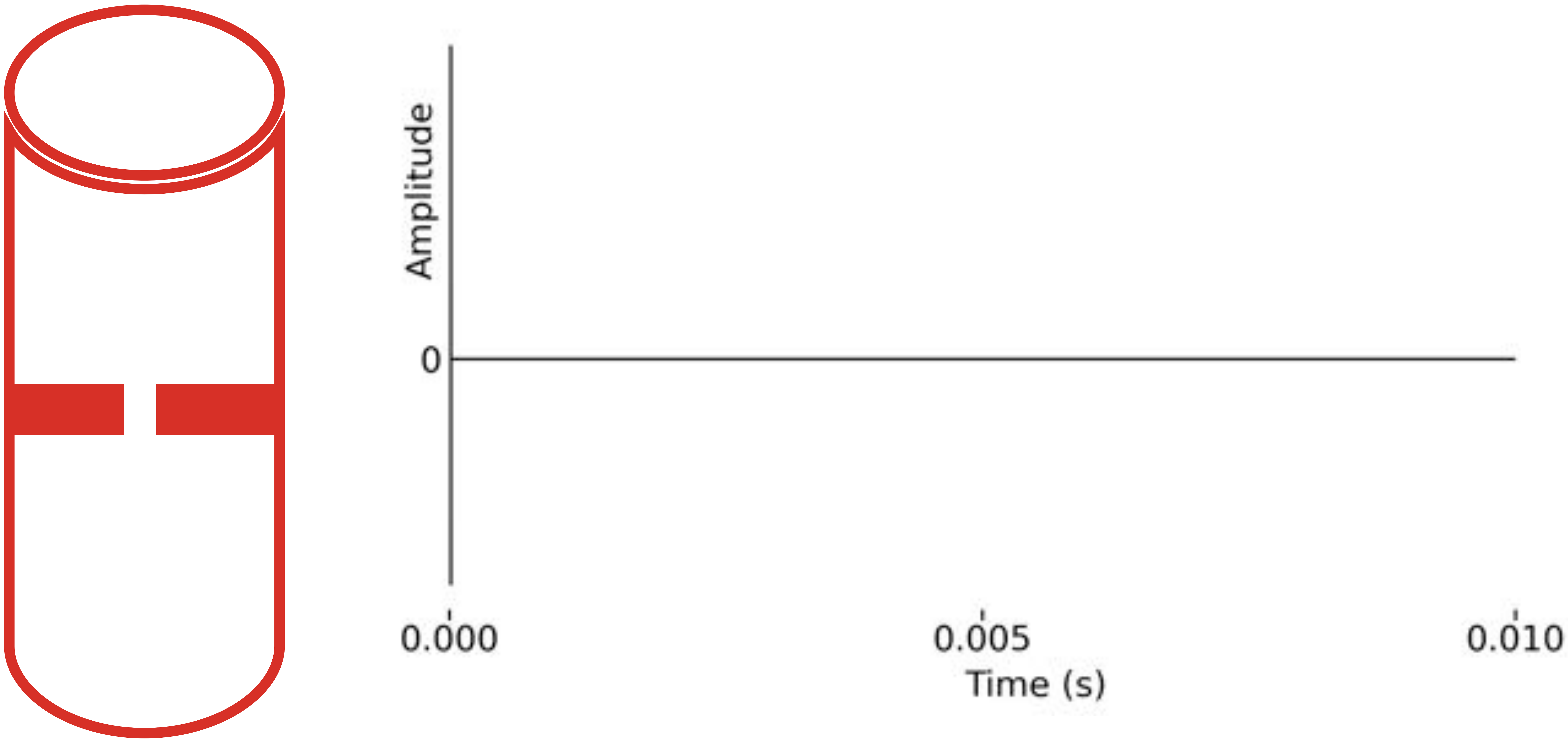
Individual pulse



Pulse train

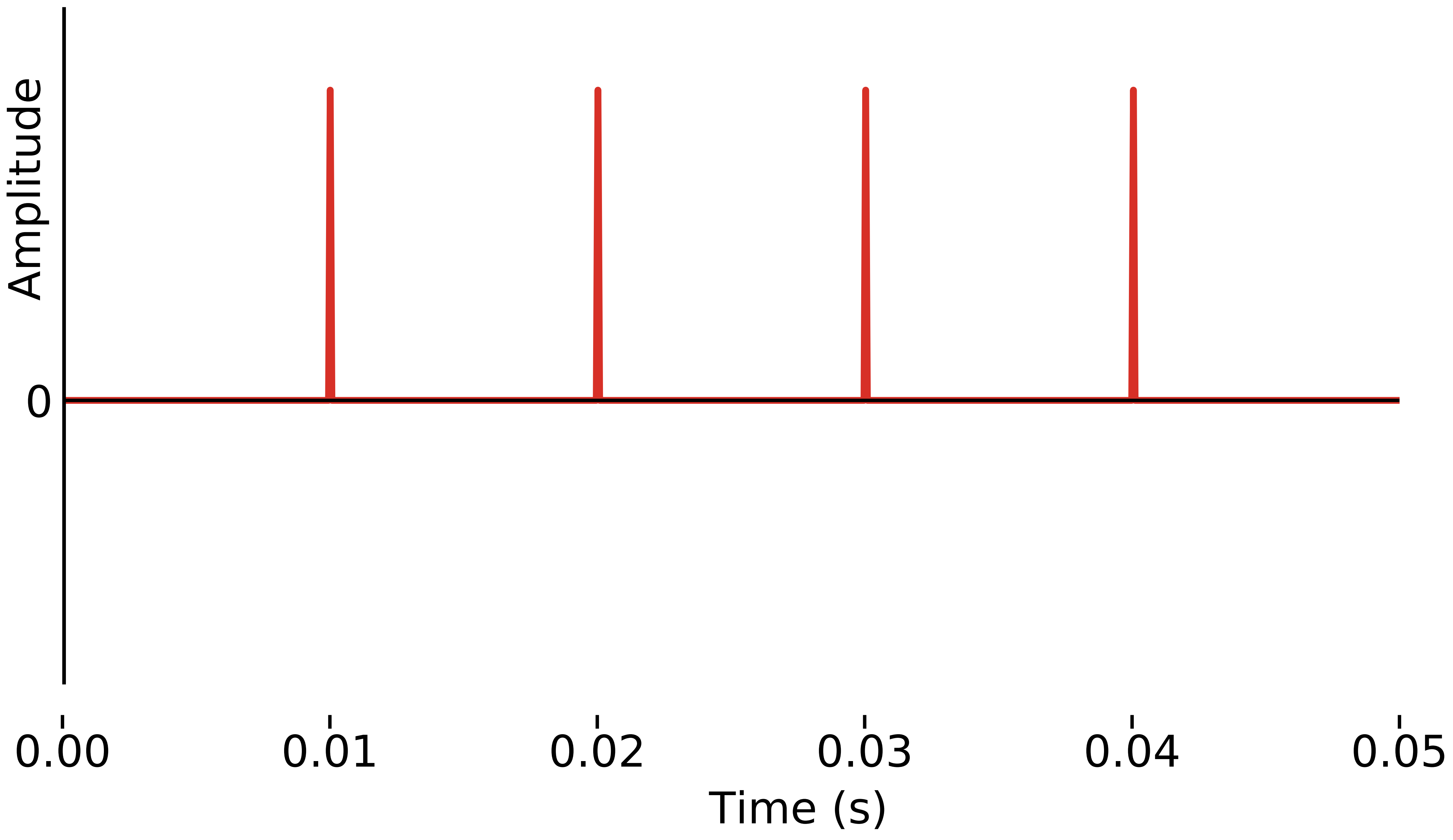


Frication: turbulent airflow caused by a constriction

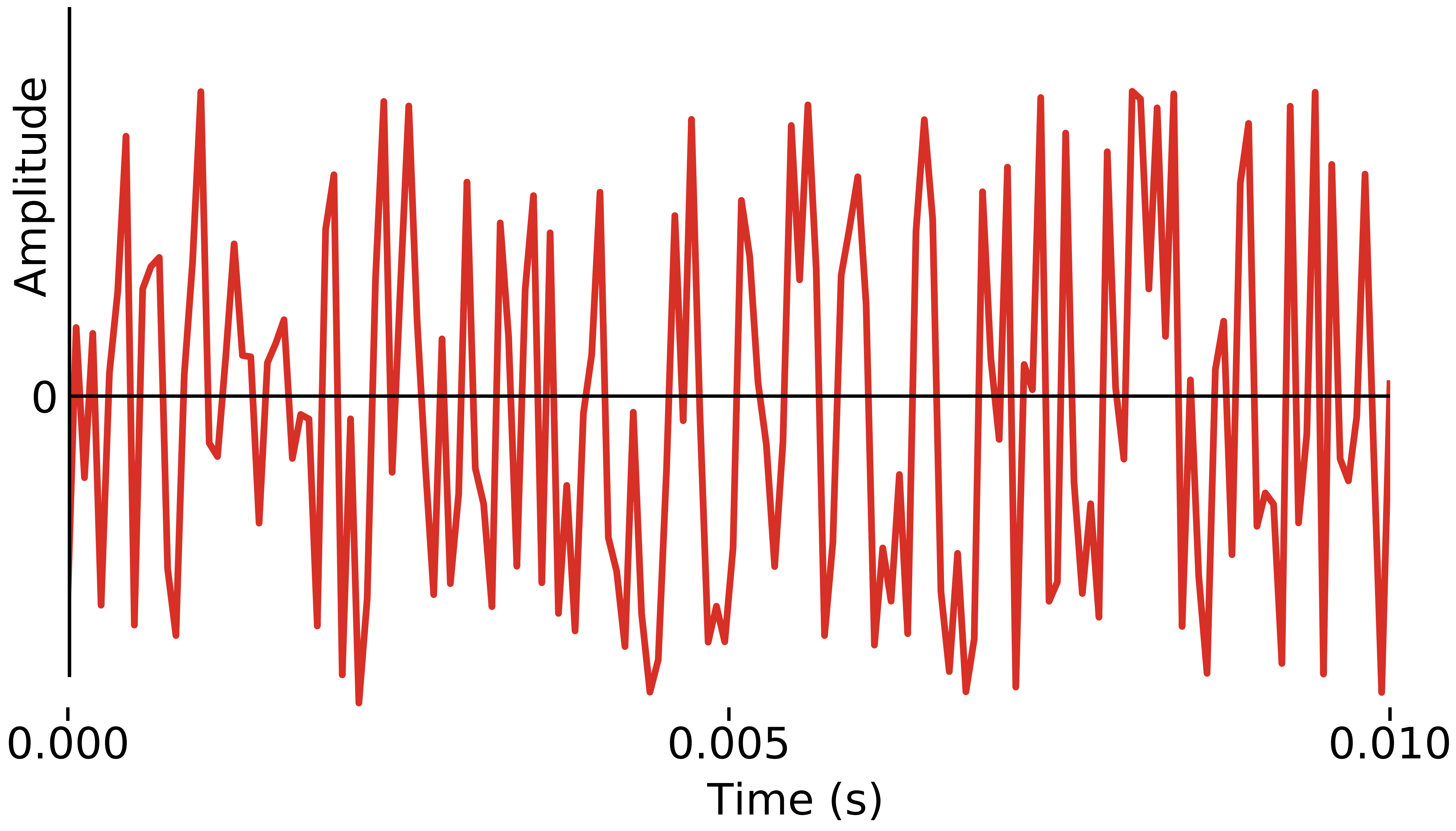


The two main sound sources in speech

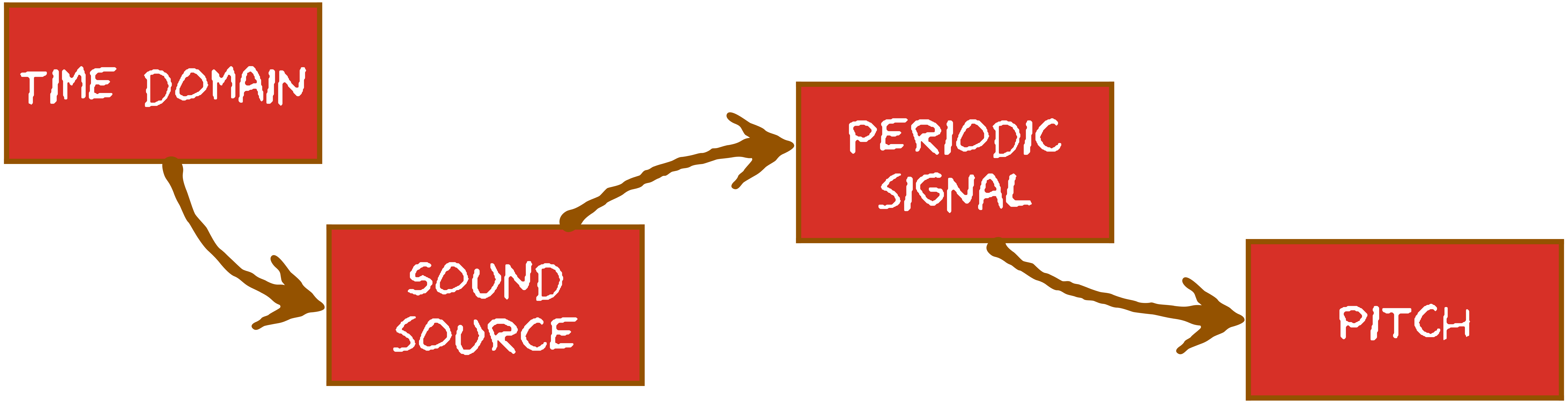
voicing



frication

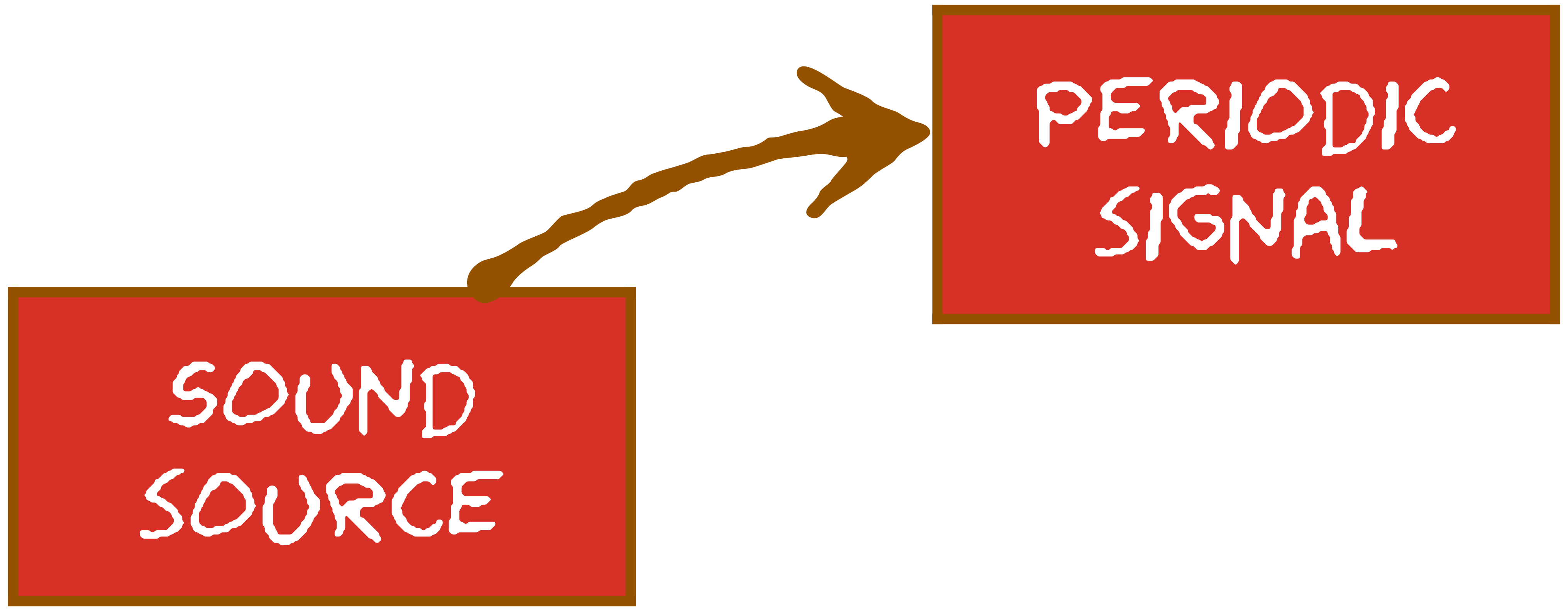


What you can learn next



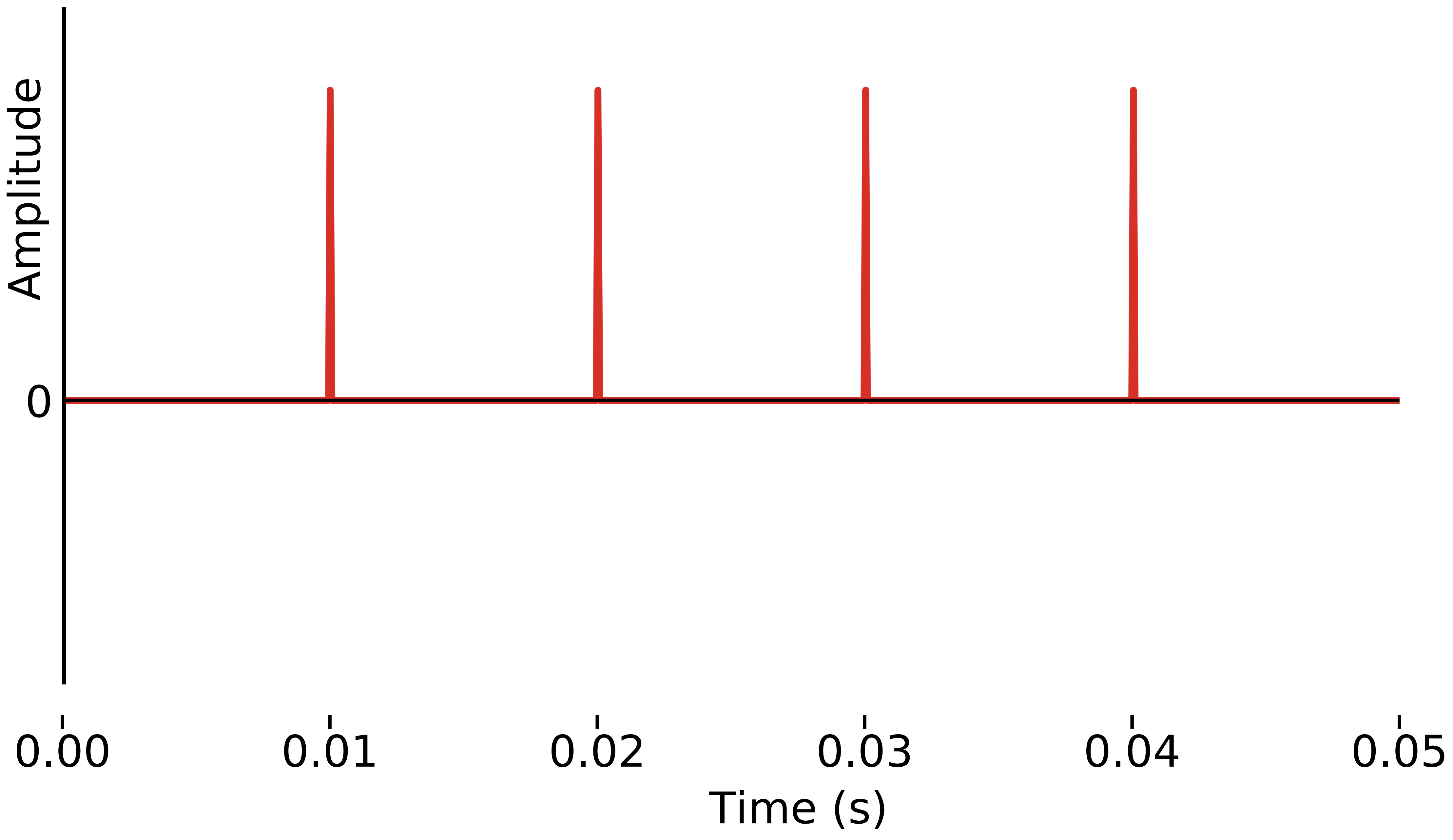
PERIODIC SIGNAL

What you need to know already

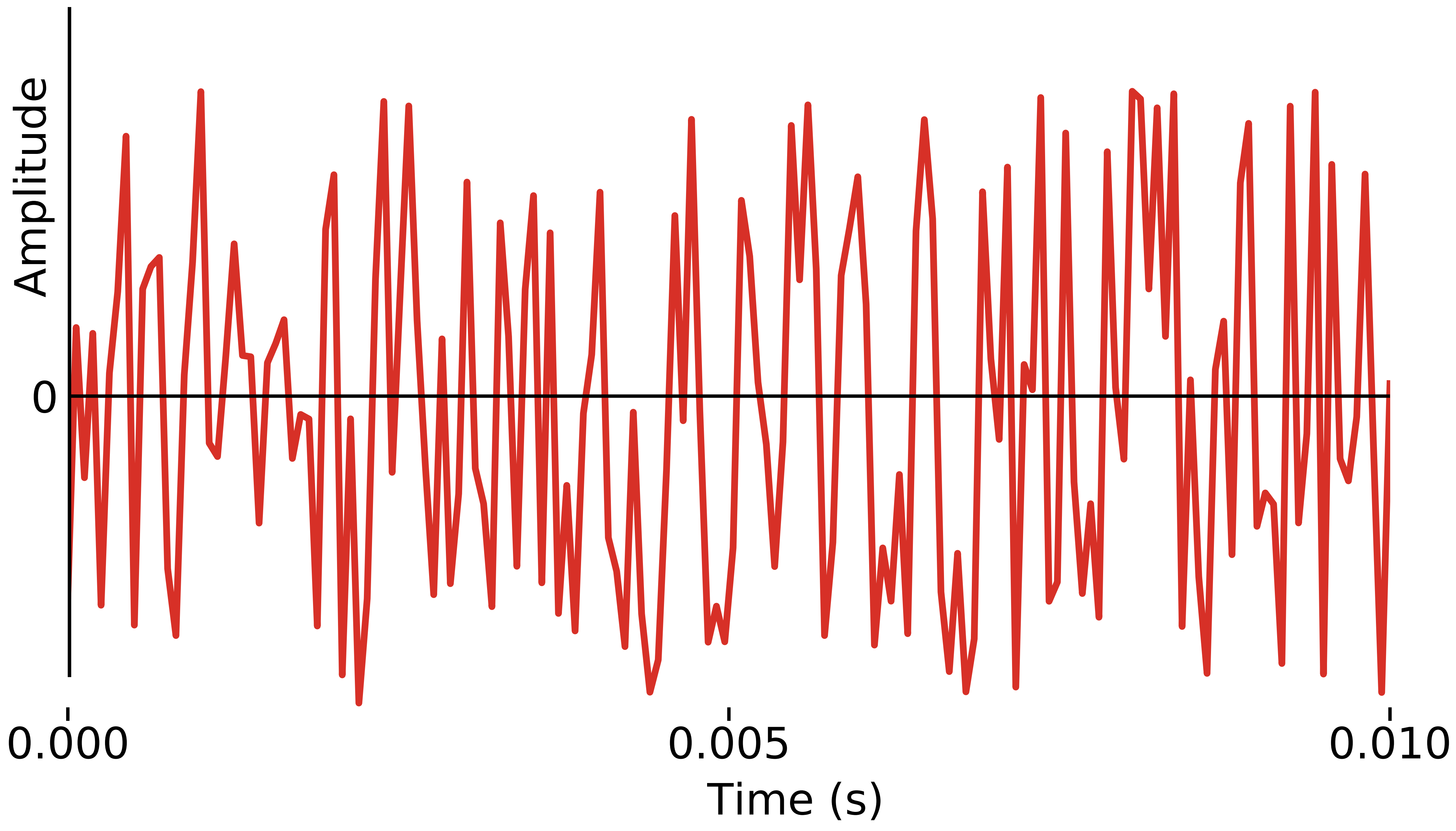


The two main sound sources in speech

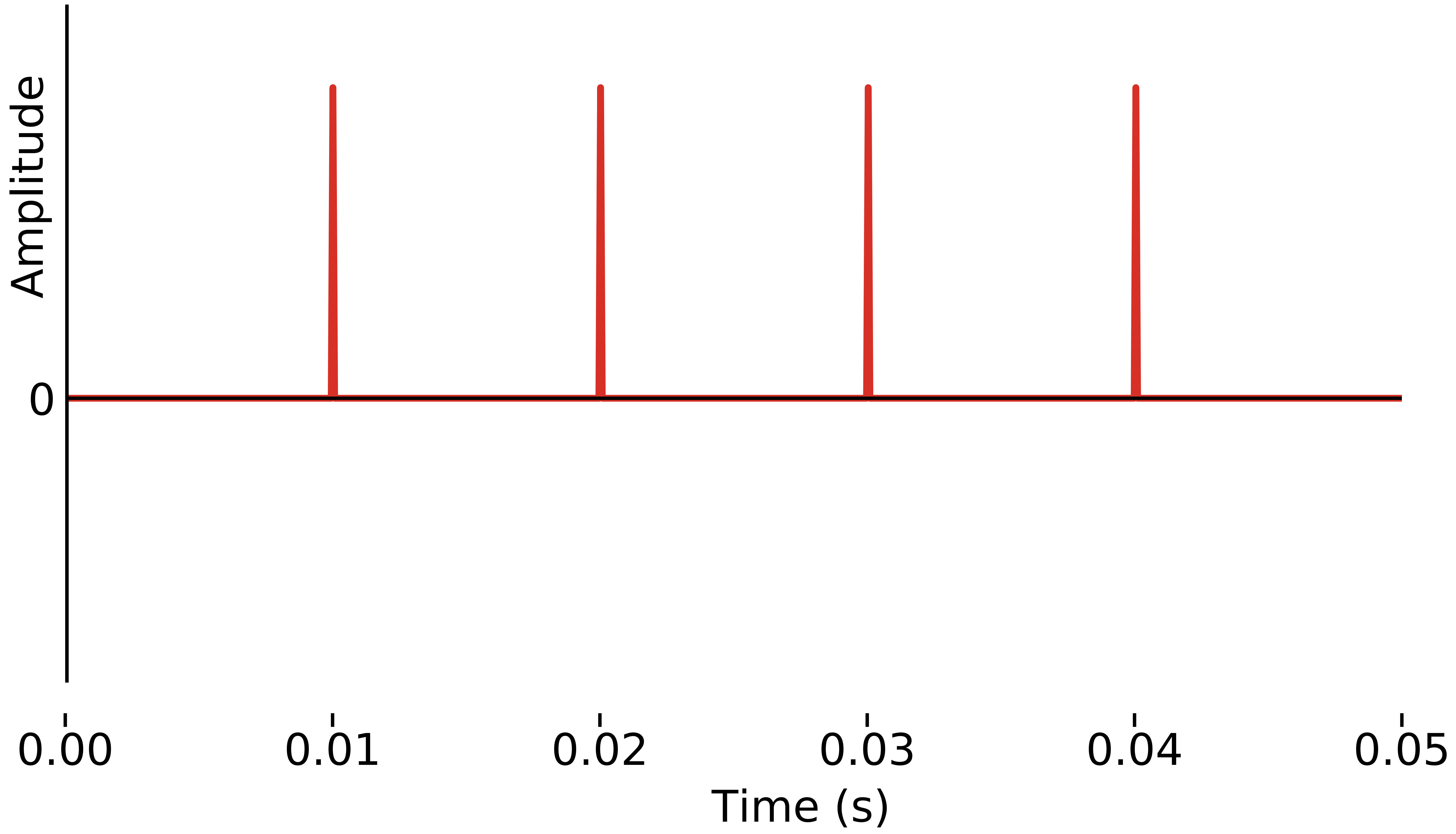
voicing

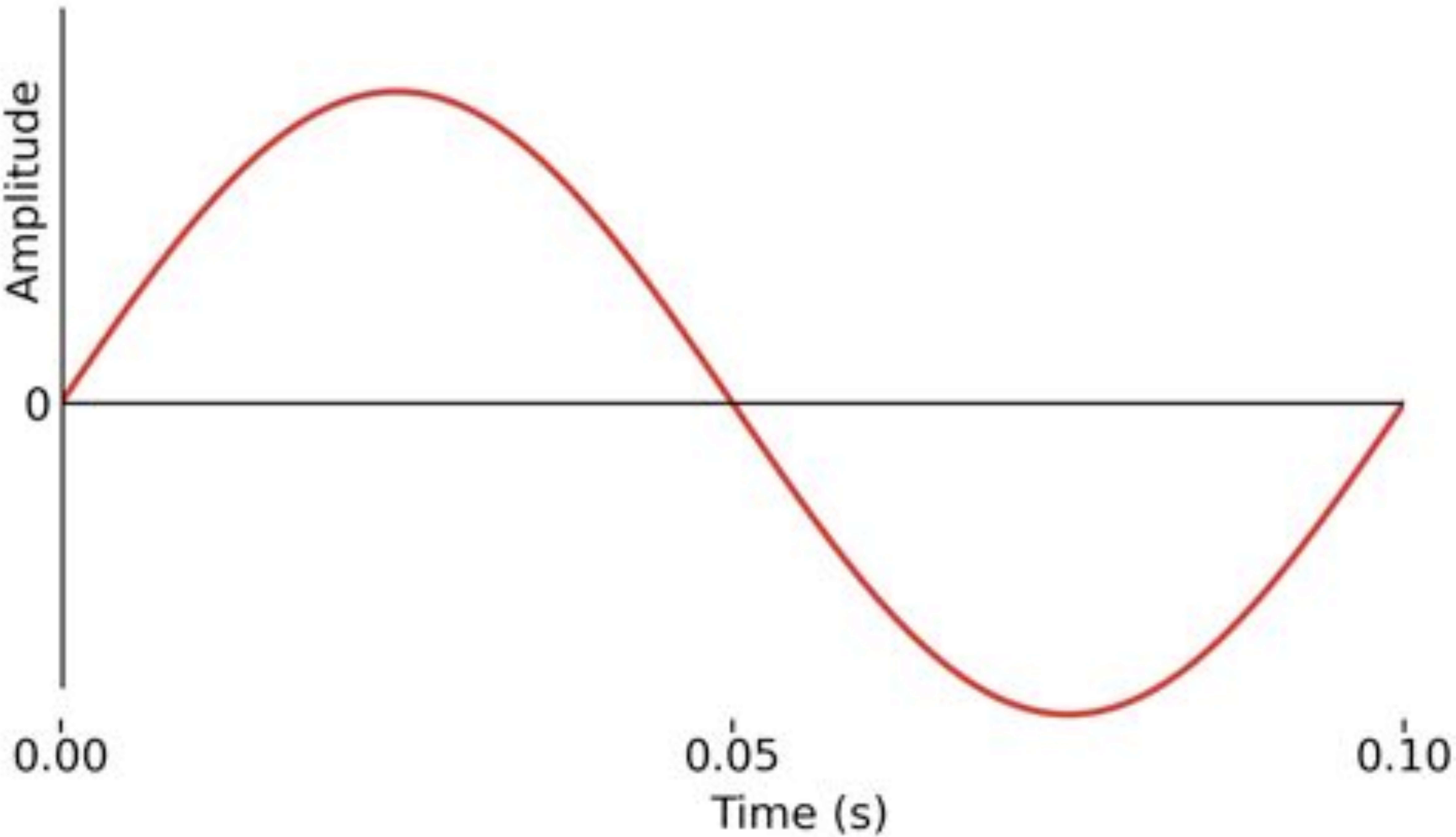


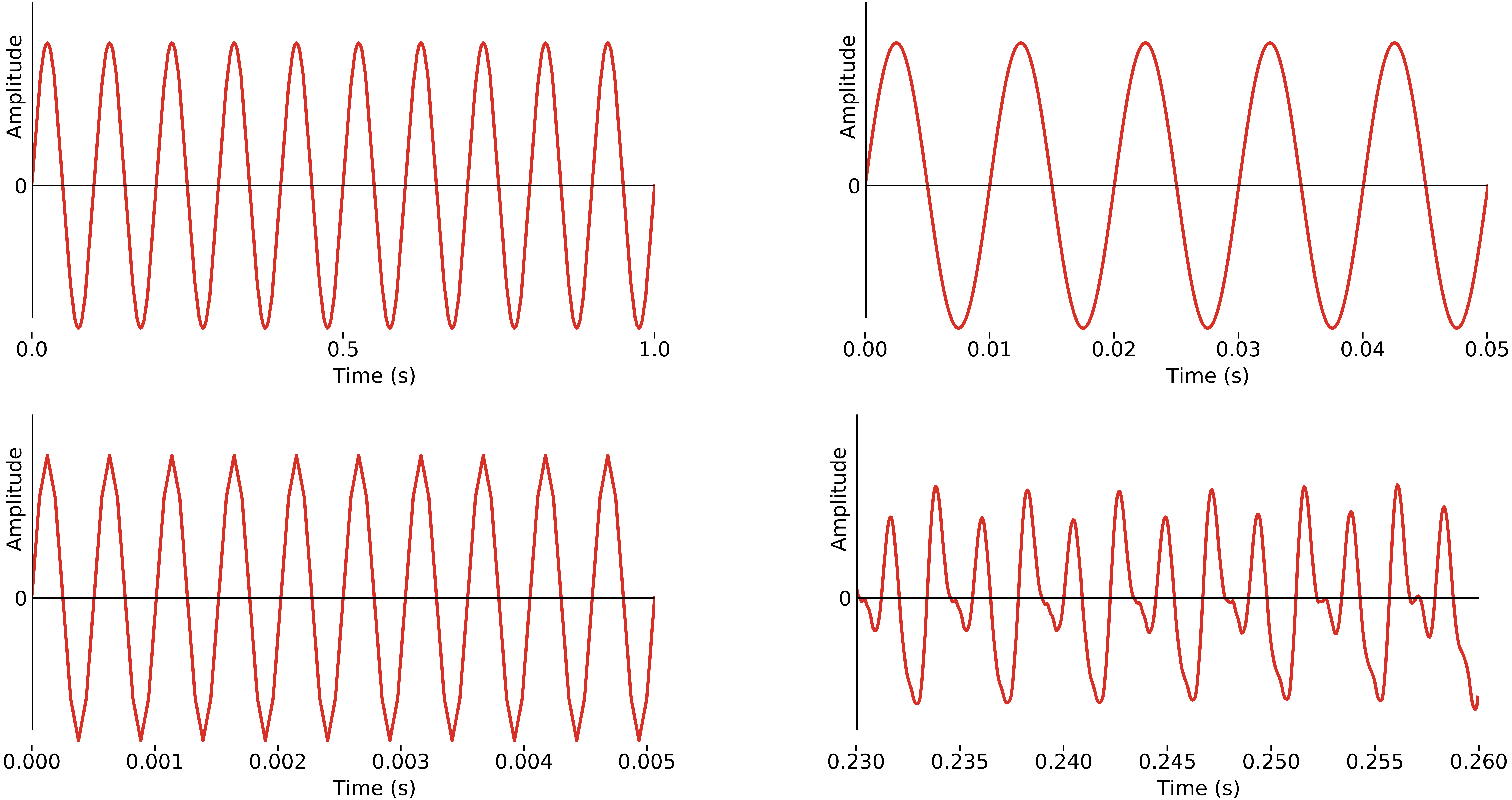
frication



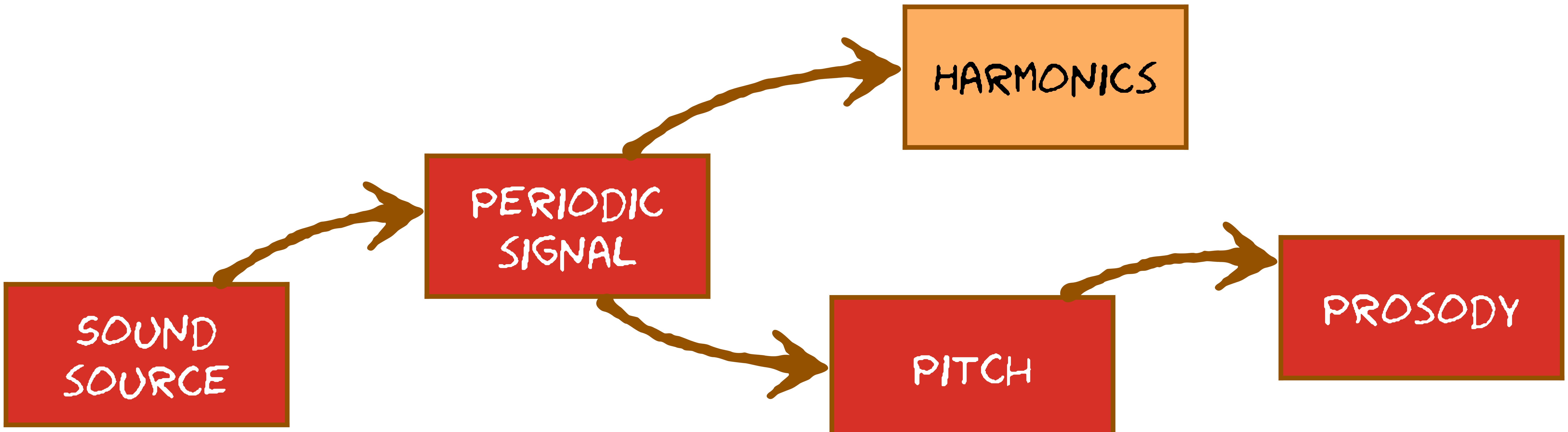
A periodic signal has a repeating pattern




Amplitude
0
0.00
0.05
0.10
Time (s)

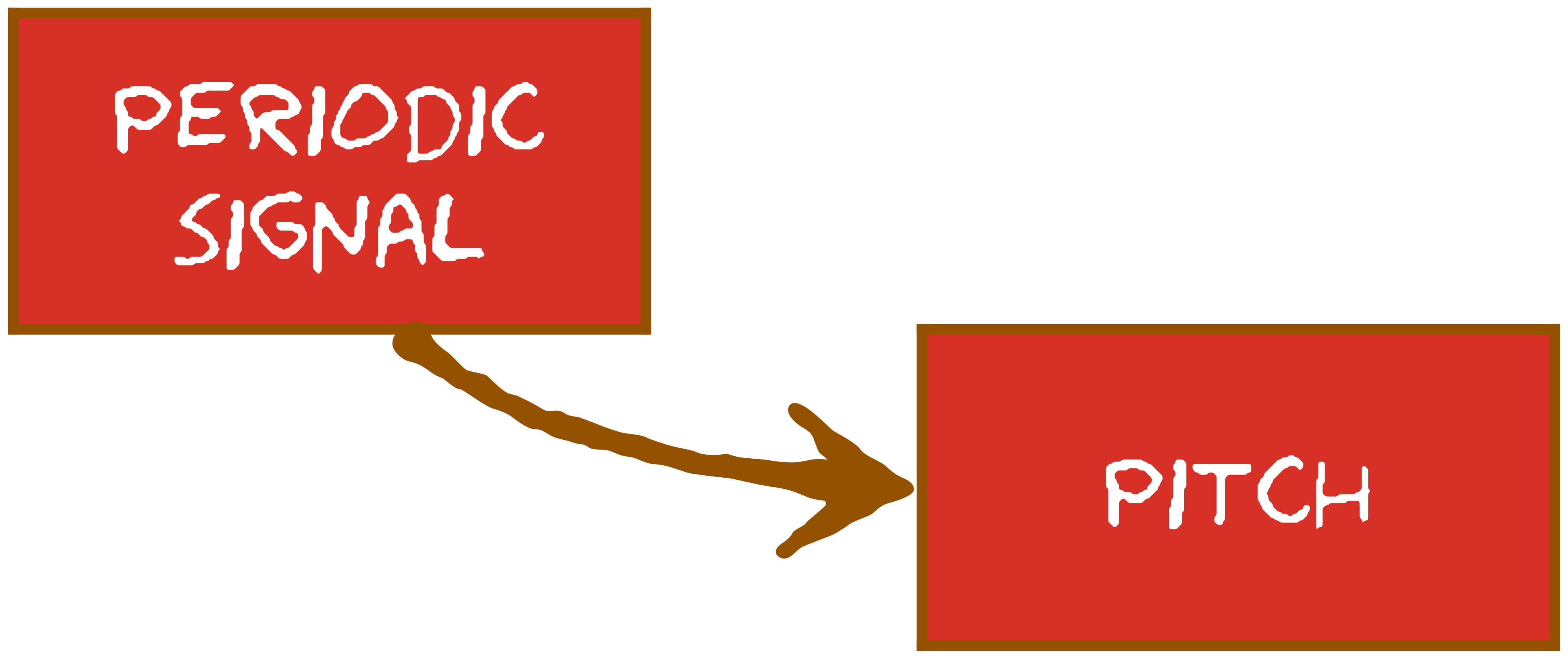

Amplitude
0
0.0
0.5
1.0
Time (s)
Amplitude
0
0.00
0.01
0.02
0.03
0.04
0.05
Time (s)
Amplitude
0
0.000
0.001
0.002
0.003
0.004
0.005
Time (s)
Amplitude
0
0.230
0.235
0.240
0.245
0.250
0.255
0.260
Time (s)

What you can learn next

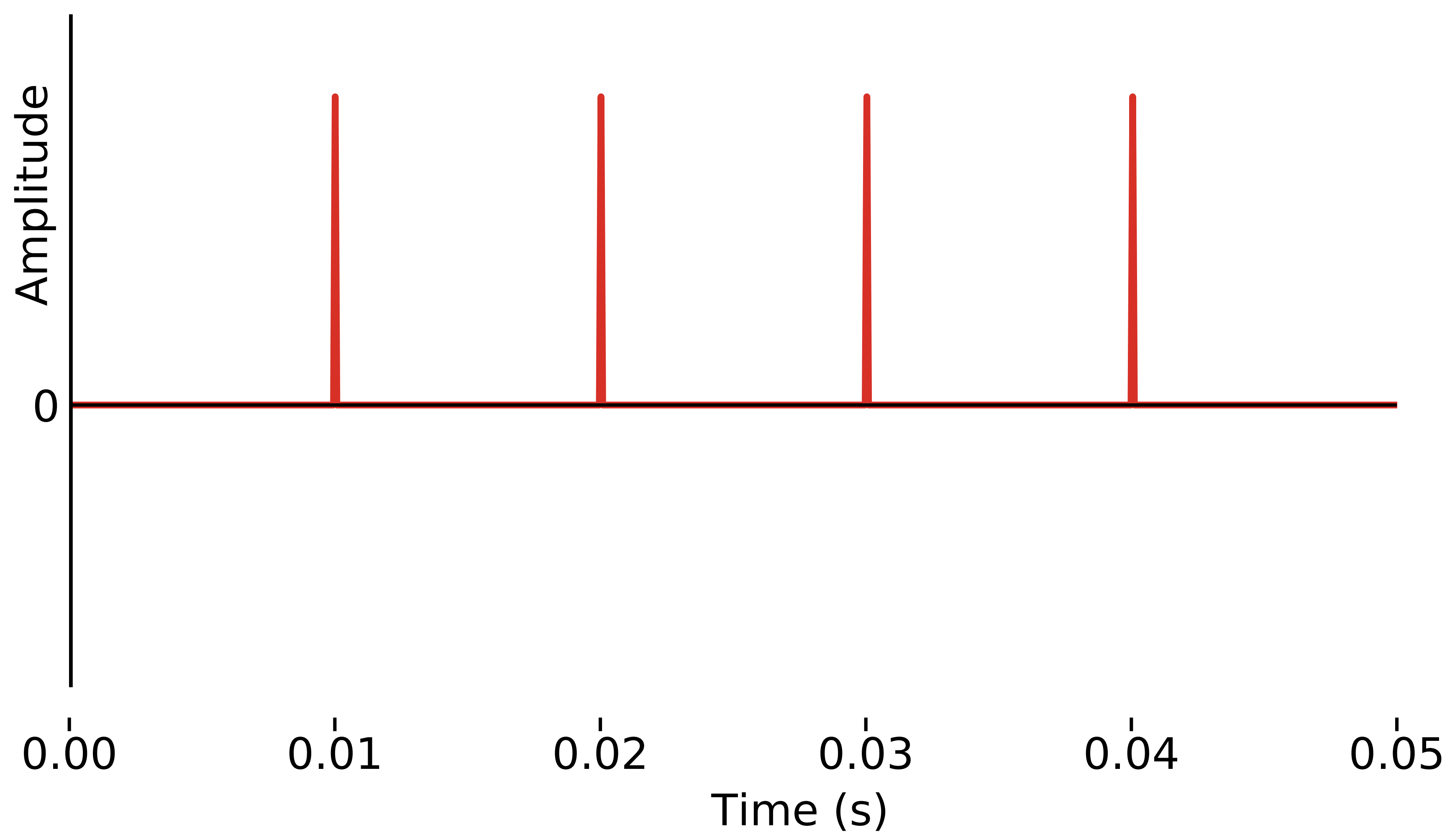


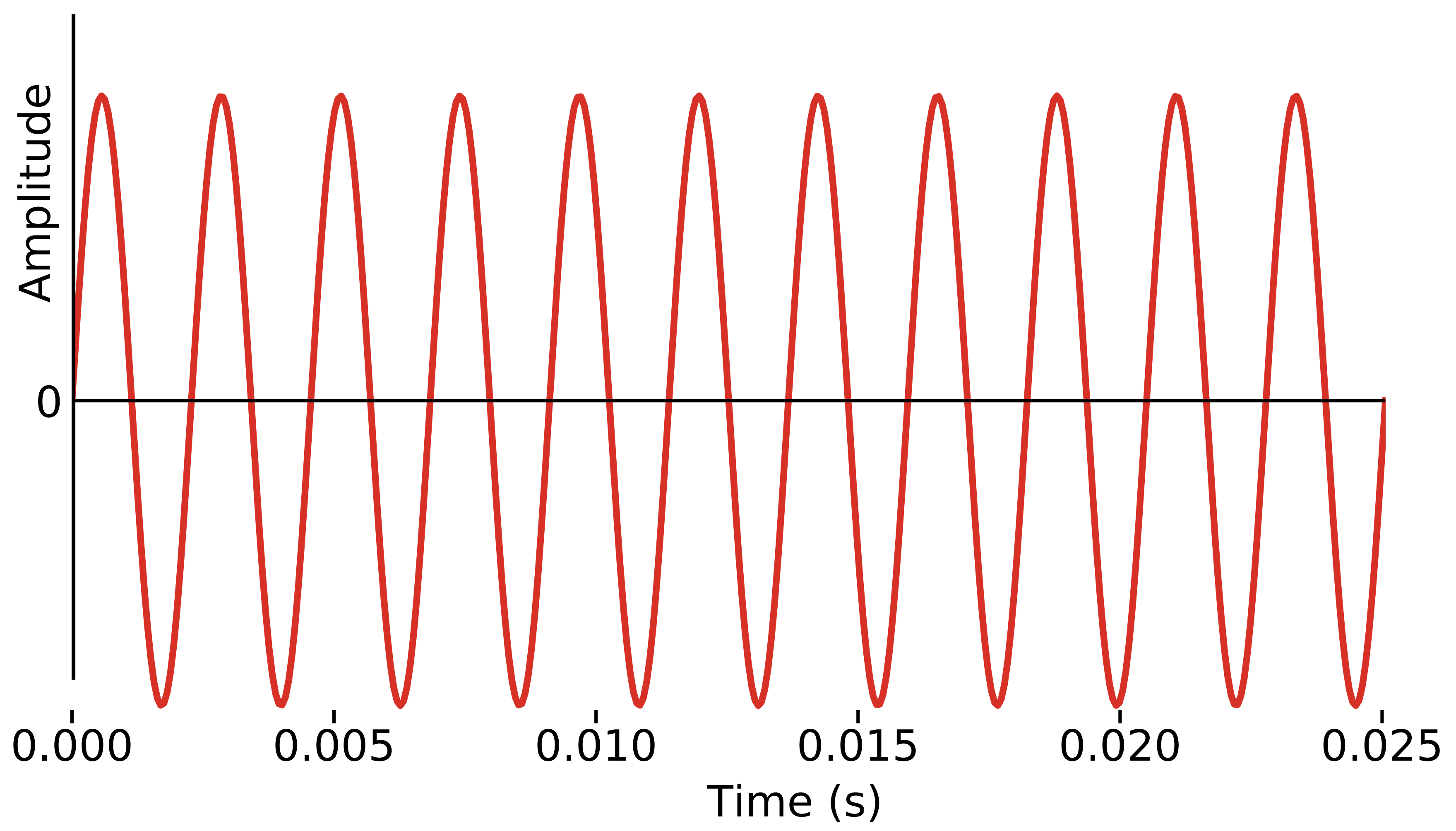
PITCH

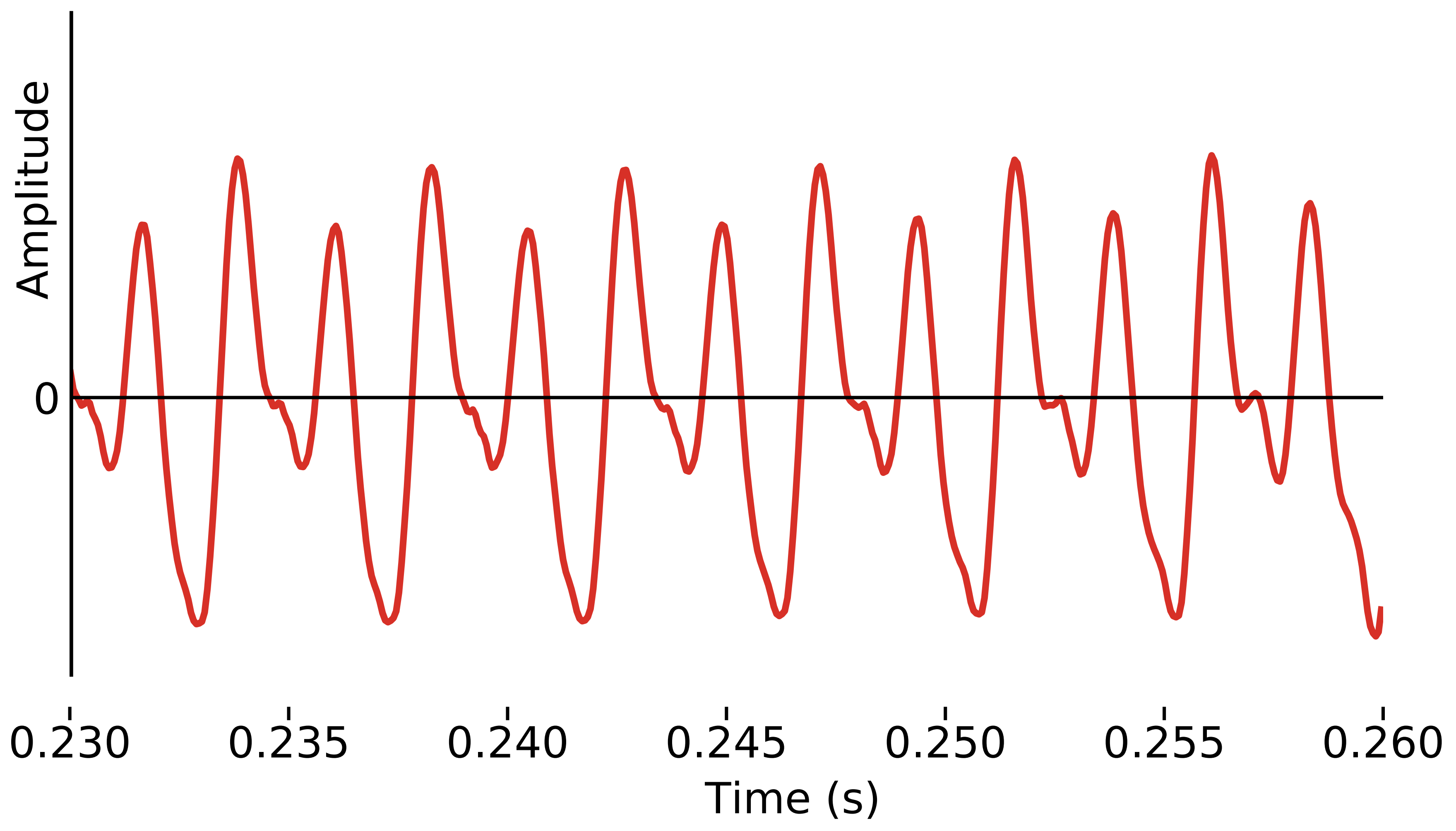
What you need to know already



Periodic signals are perceived as having pitch: a musical note



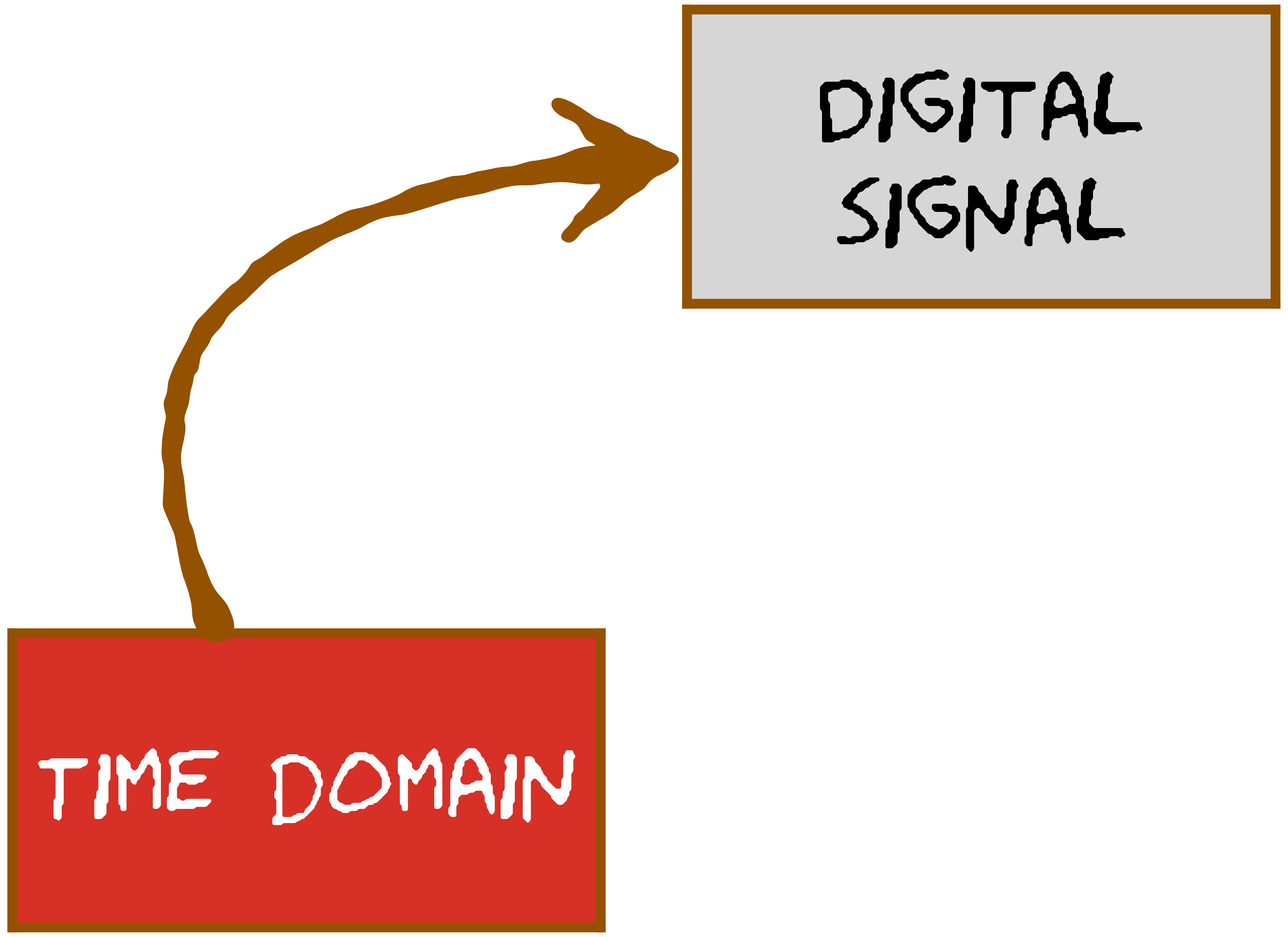




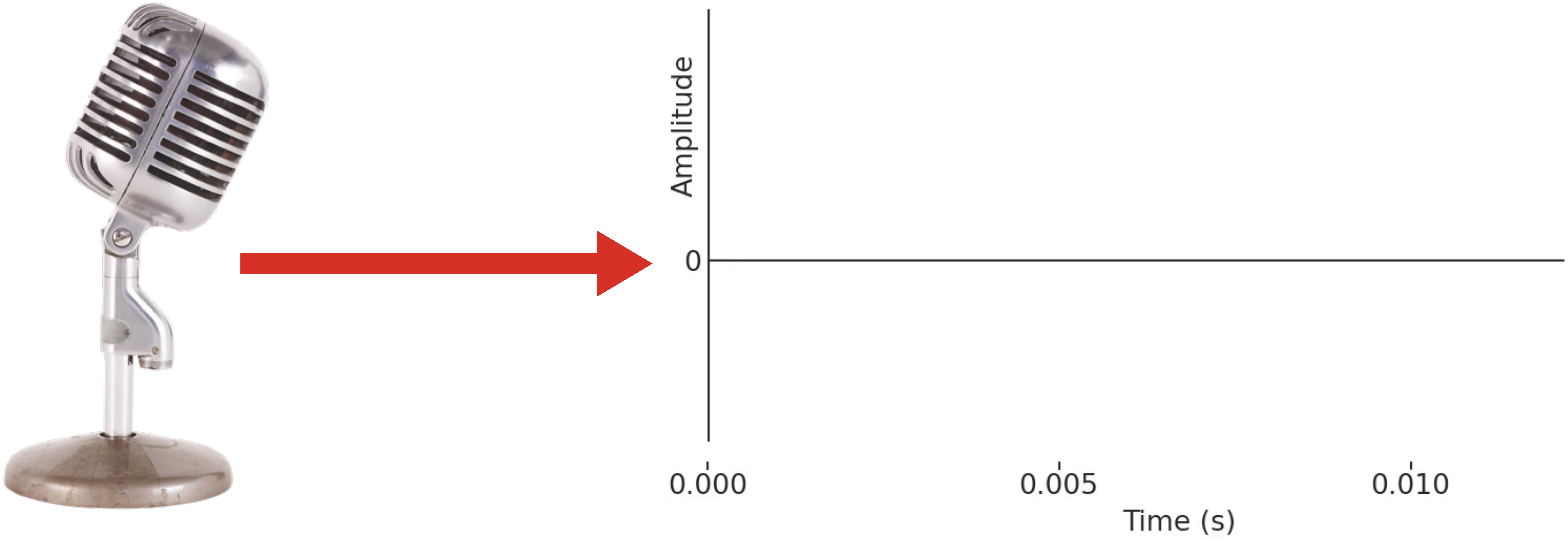
What you can learn next

DIGITAL SIGNAL

What you need to know already



Analogue-to-digital conversion = sampling and quantisation

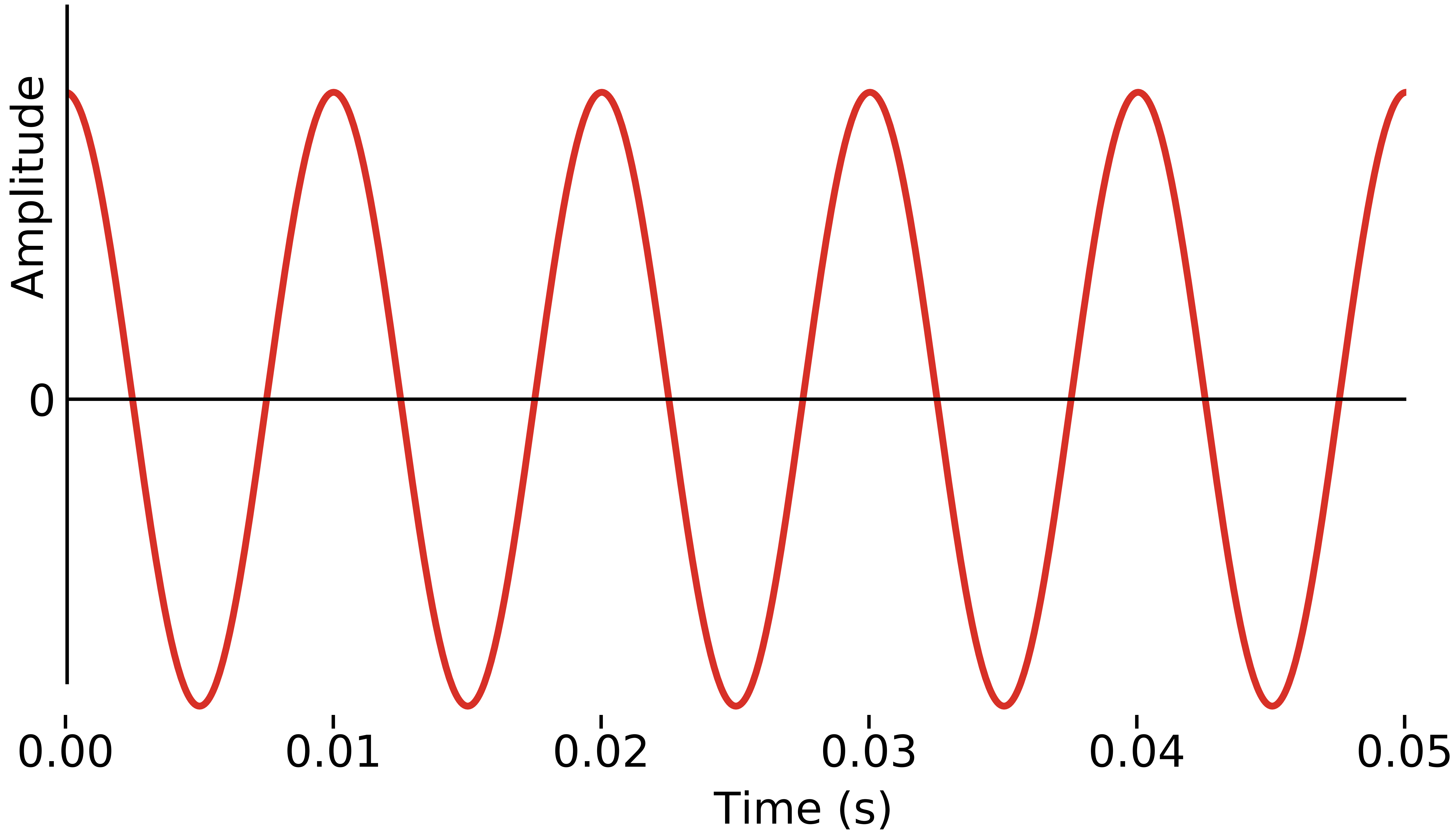


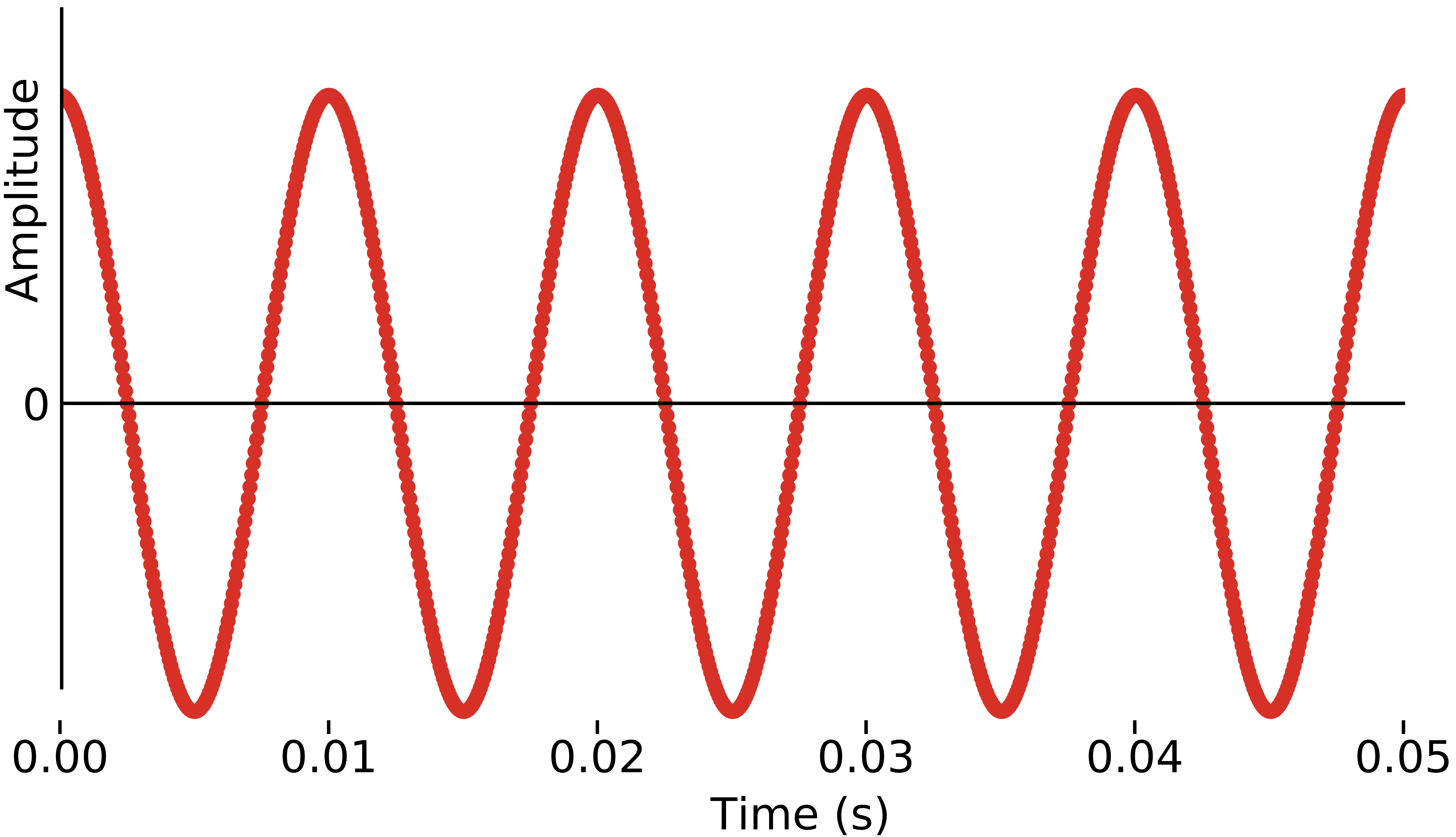
Why does “digital” mean making everything **discrete**?

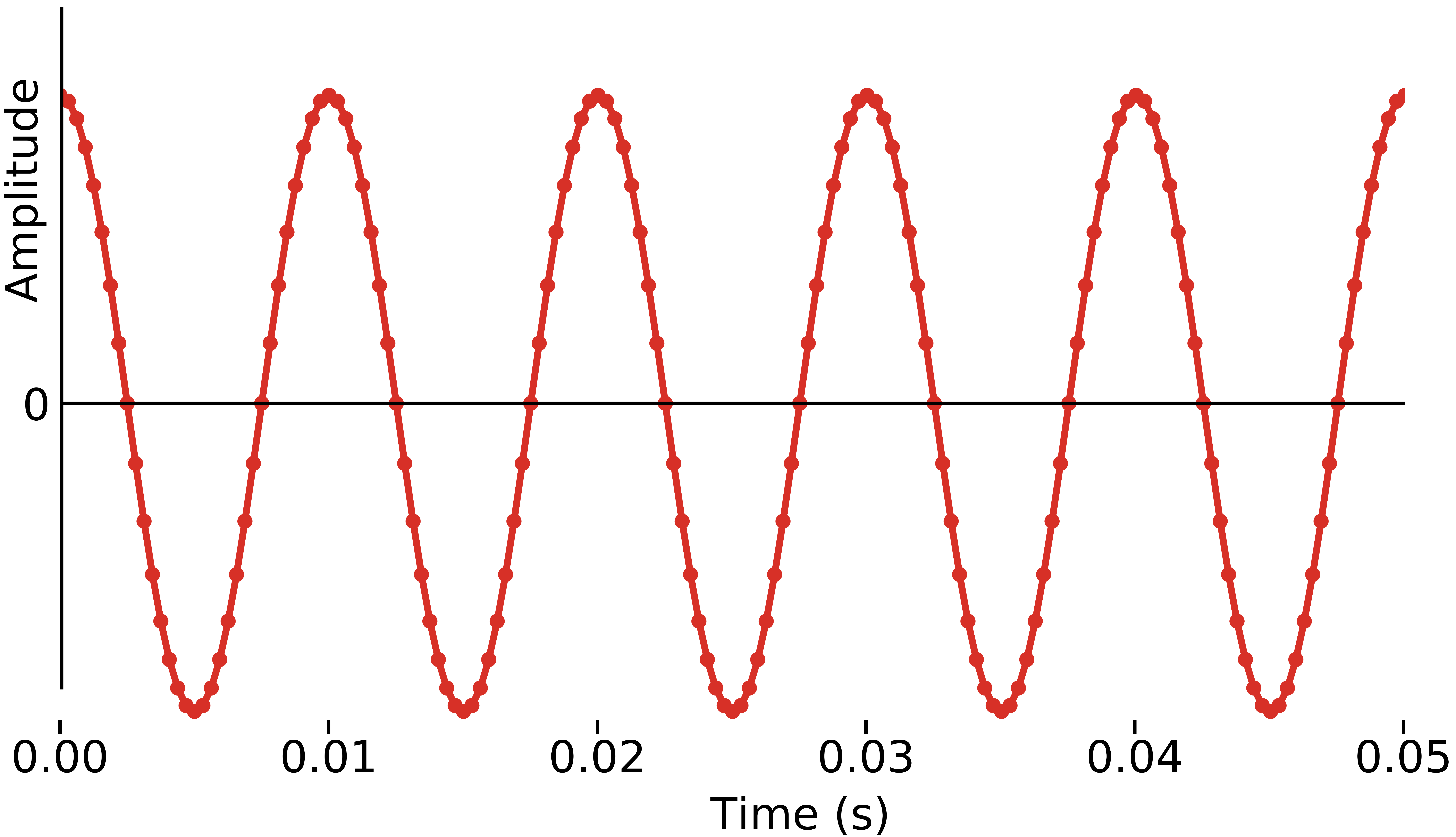
1-bit binary numbers	2-bit binary numbers	3-bit binary numbers
1	1 1	1 1 1
0	1 0	1 1 0
	0 1	1 0 1
	0 0	1 0 0
		0 1 1
		0 1 0
		0 0 1
		0 0 0

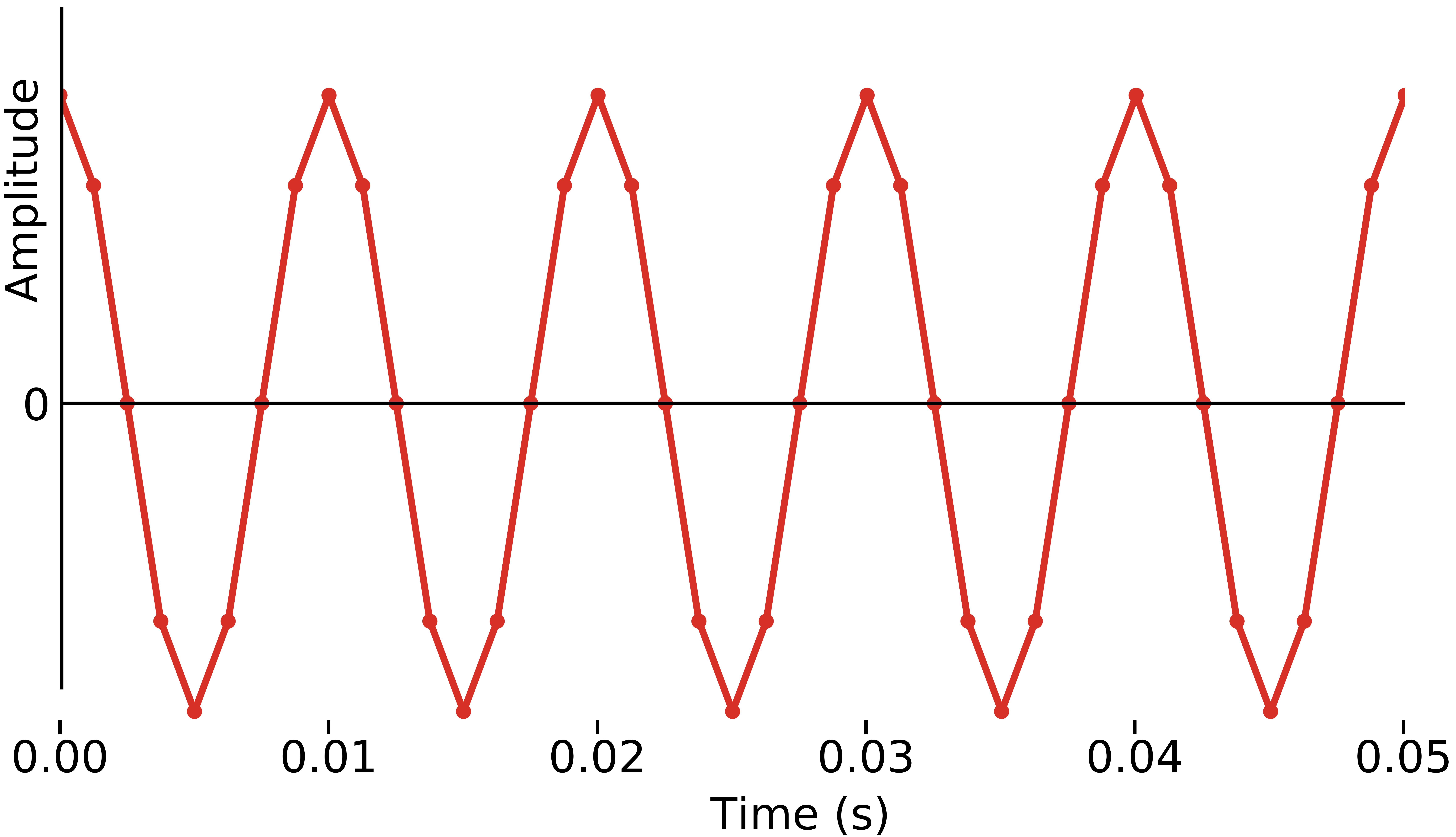
Sampling = making time digital

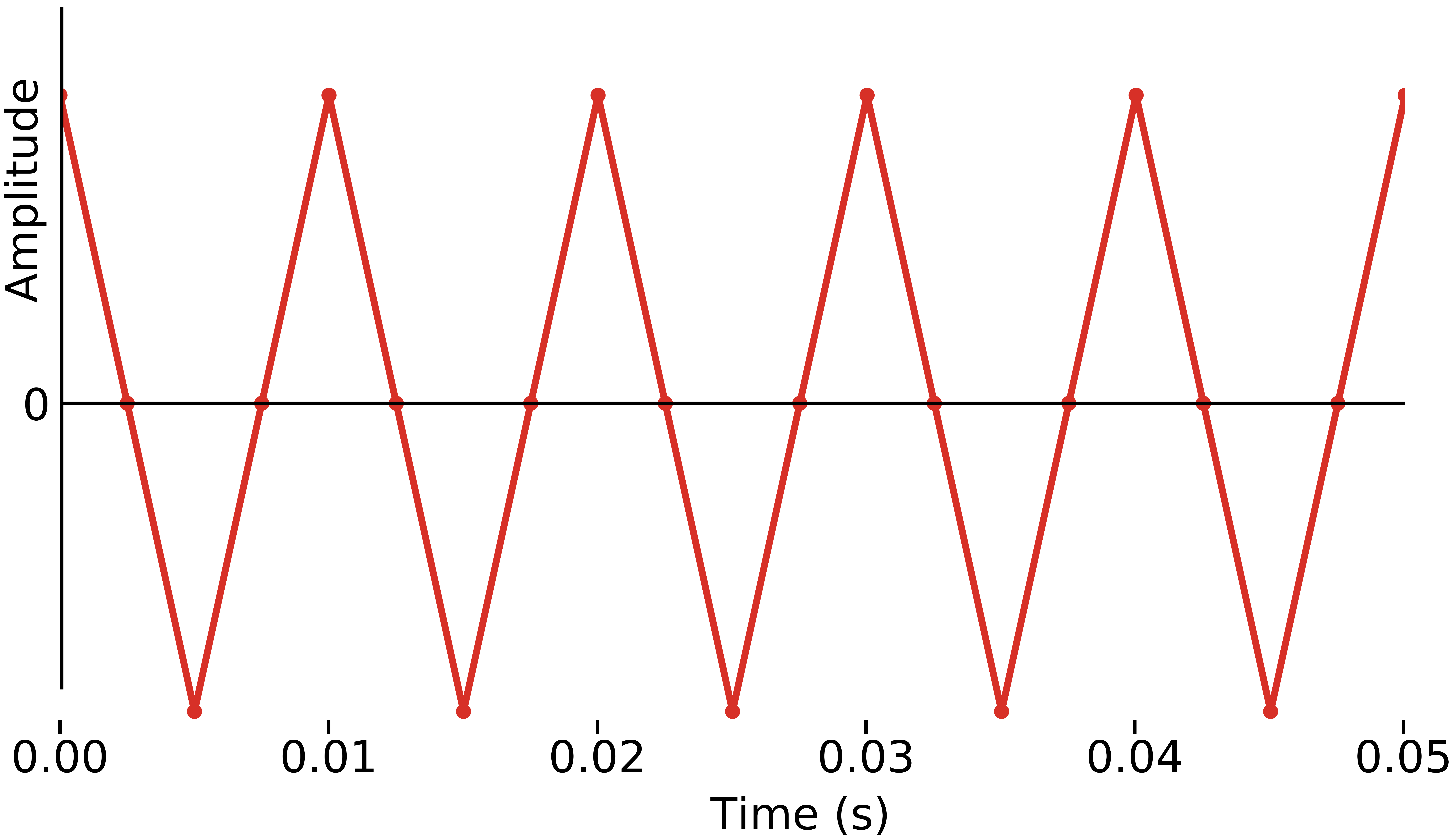
How frequently should we sample the waveform?

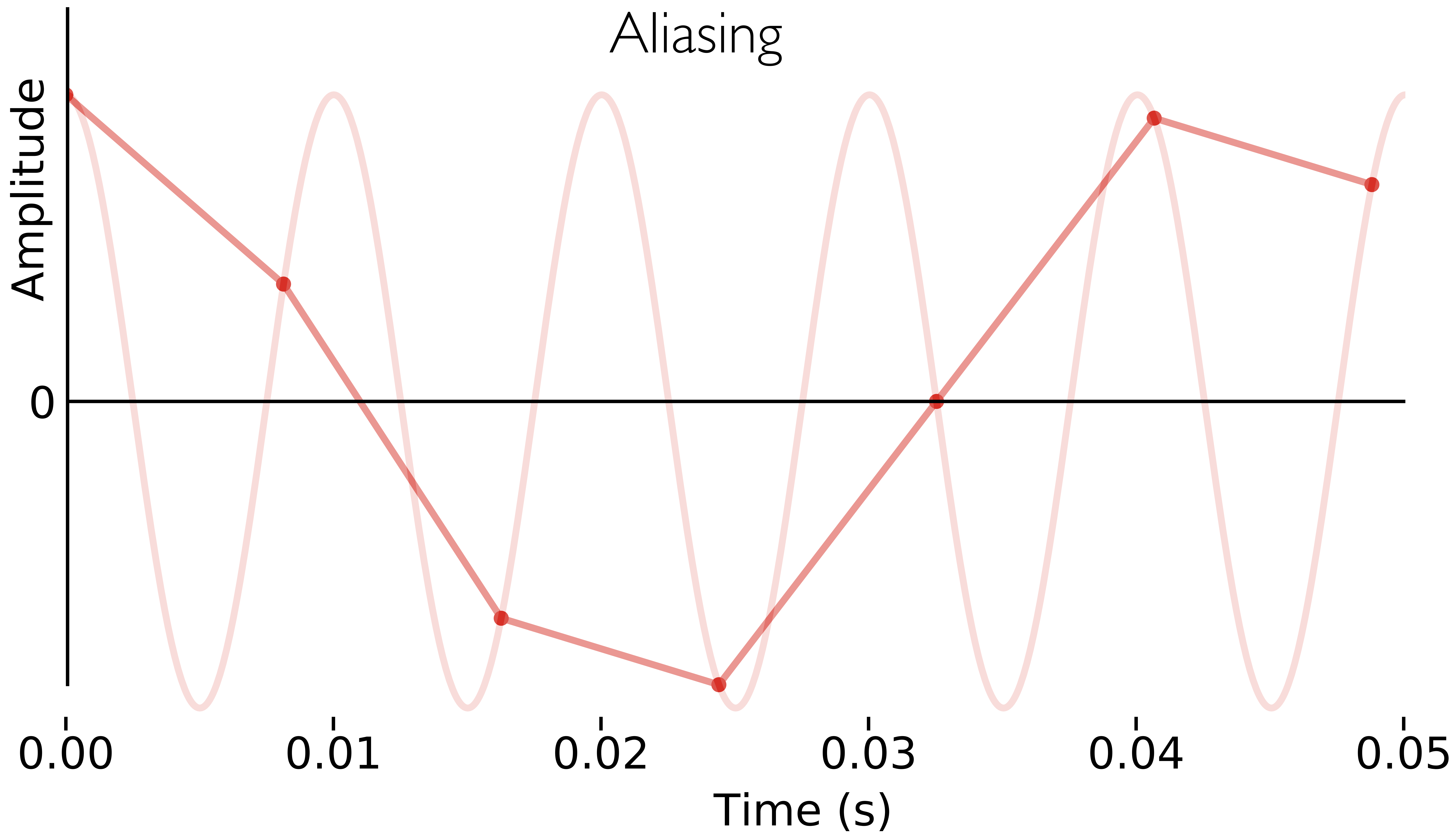



Amplitude
0
0.00
0.01
0.02
0.03
0.04
0.05
Time (s)

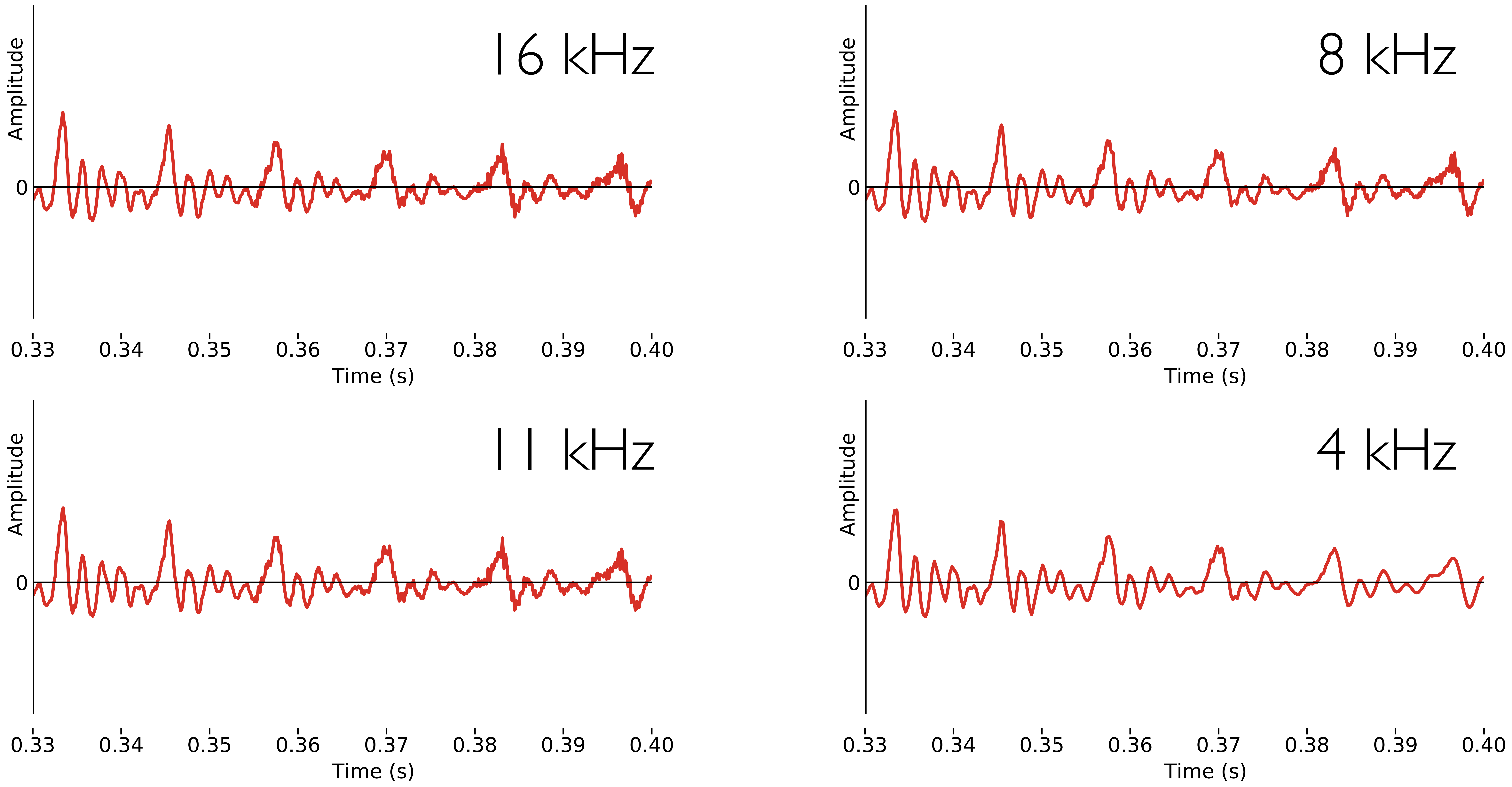

Amplitude
0
0.00
0.01
0.02
0.03
0.04
0.05
Time (s)


Amplitude
0
0.00
0.01
0.02
0.03
0.04
0.05
Time (s)


Amplitude
0
0.00
0.01
0.02
0.03
0.04
0.05
Time (s)

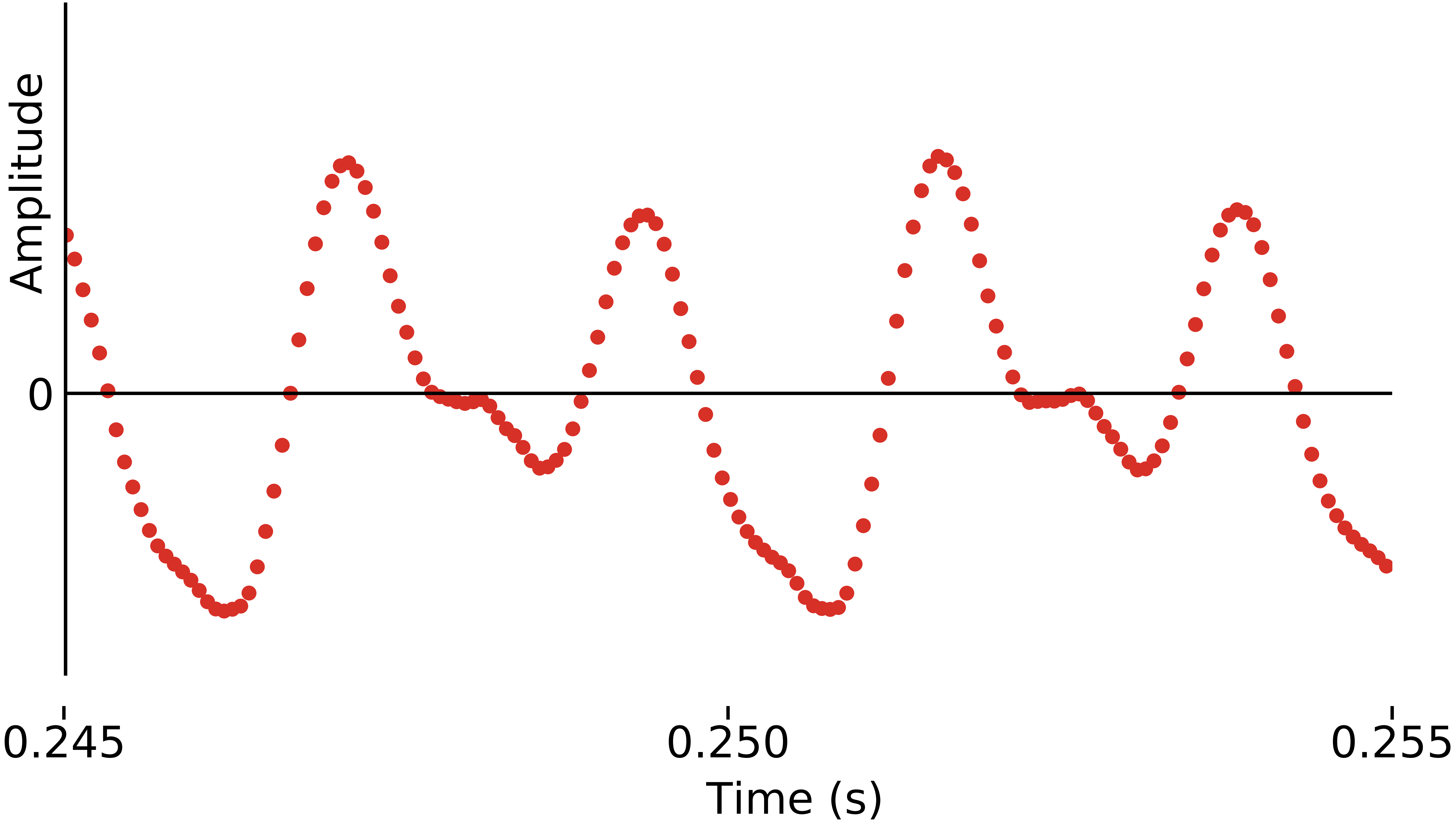

Aliasing
Amplitude
0
0.00
0.01
0.02
0.03
0.04
0.05
Time (s)

The audible effect of reducing the sampling frequency

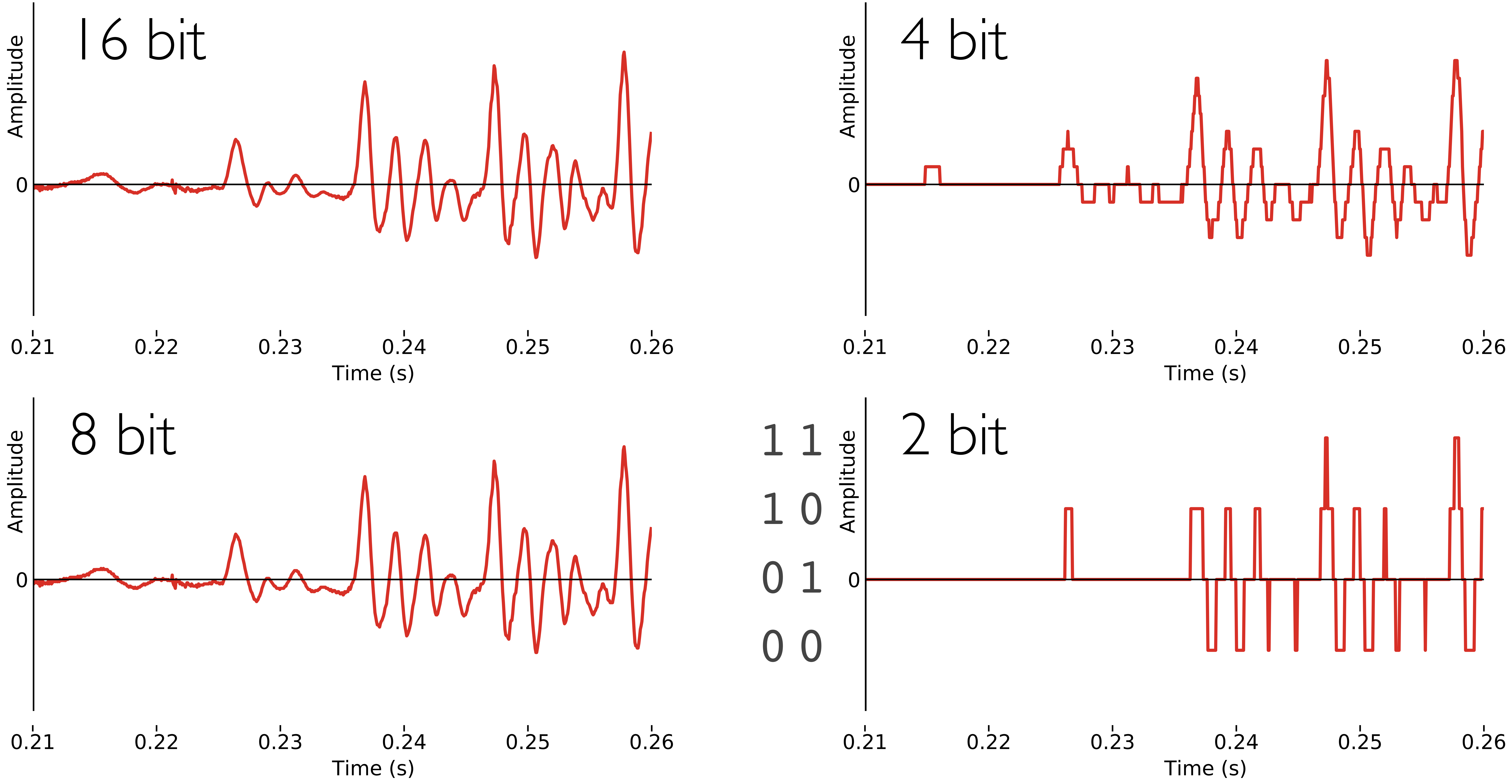


Quantisation = making amplitude digital

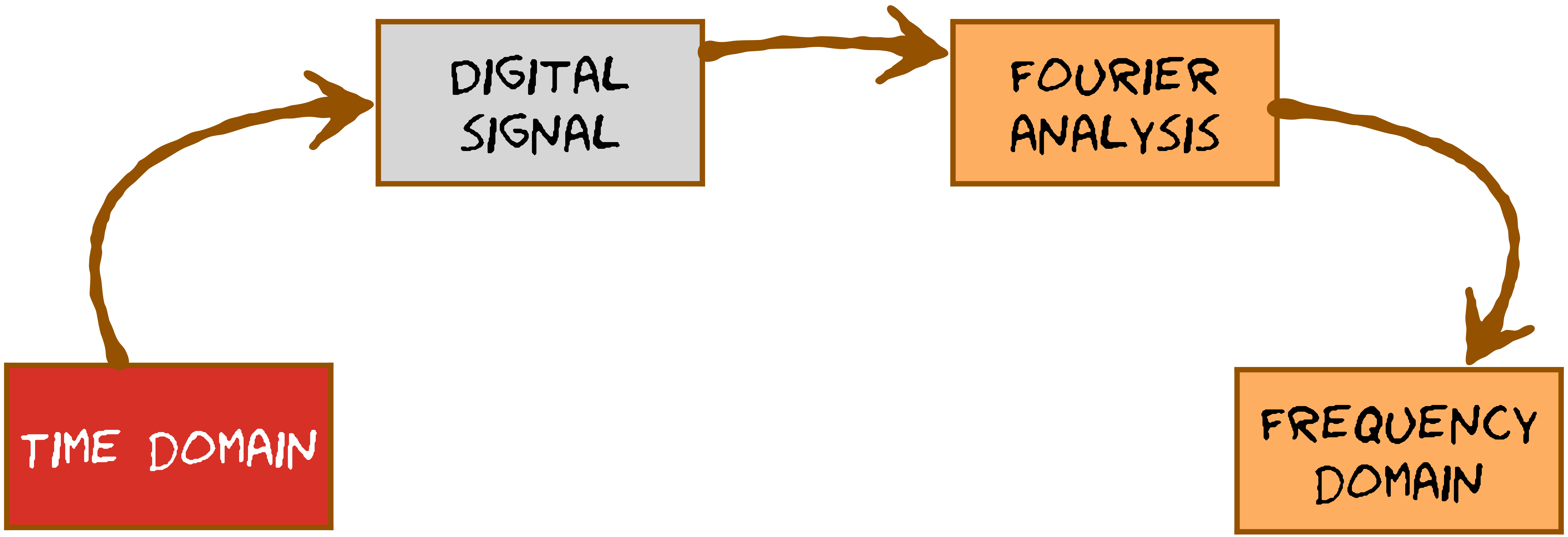
bit depth



The audible effect of reducing the bit depth



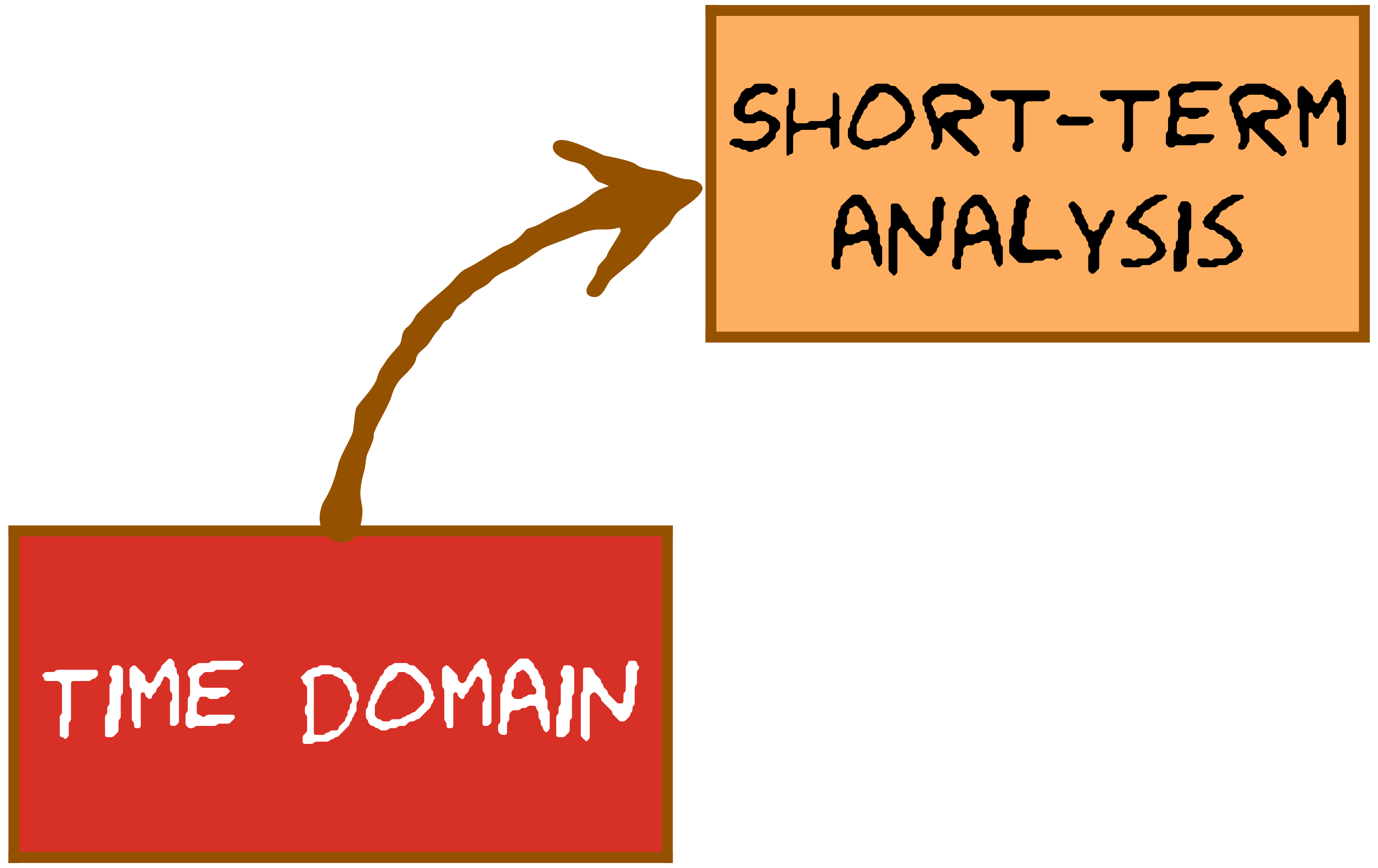
What you can learn next



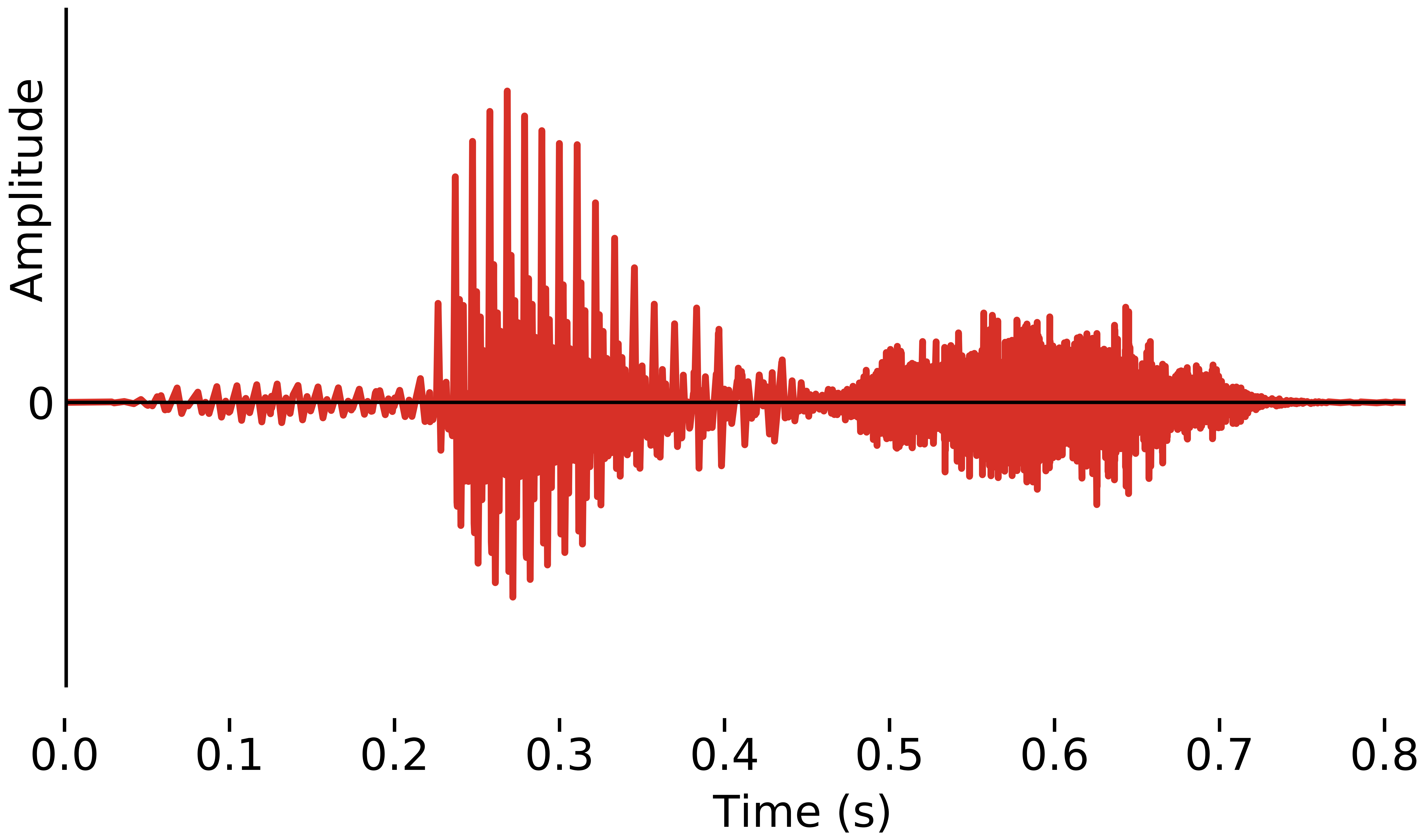
SHORT-TERM ANALYSIS

FREQUENCY DOMAIN AND BEYOND

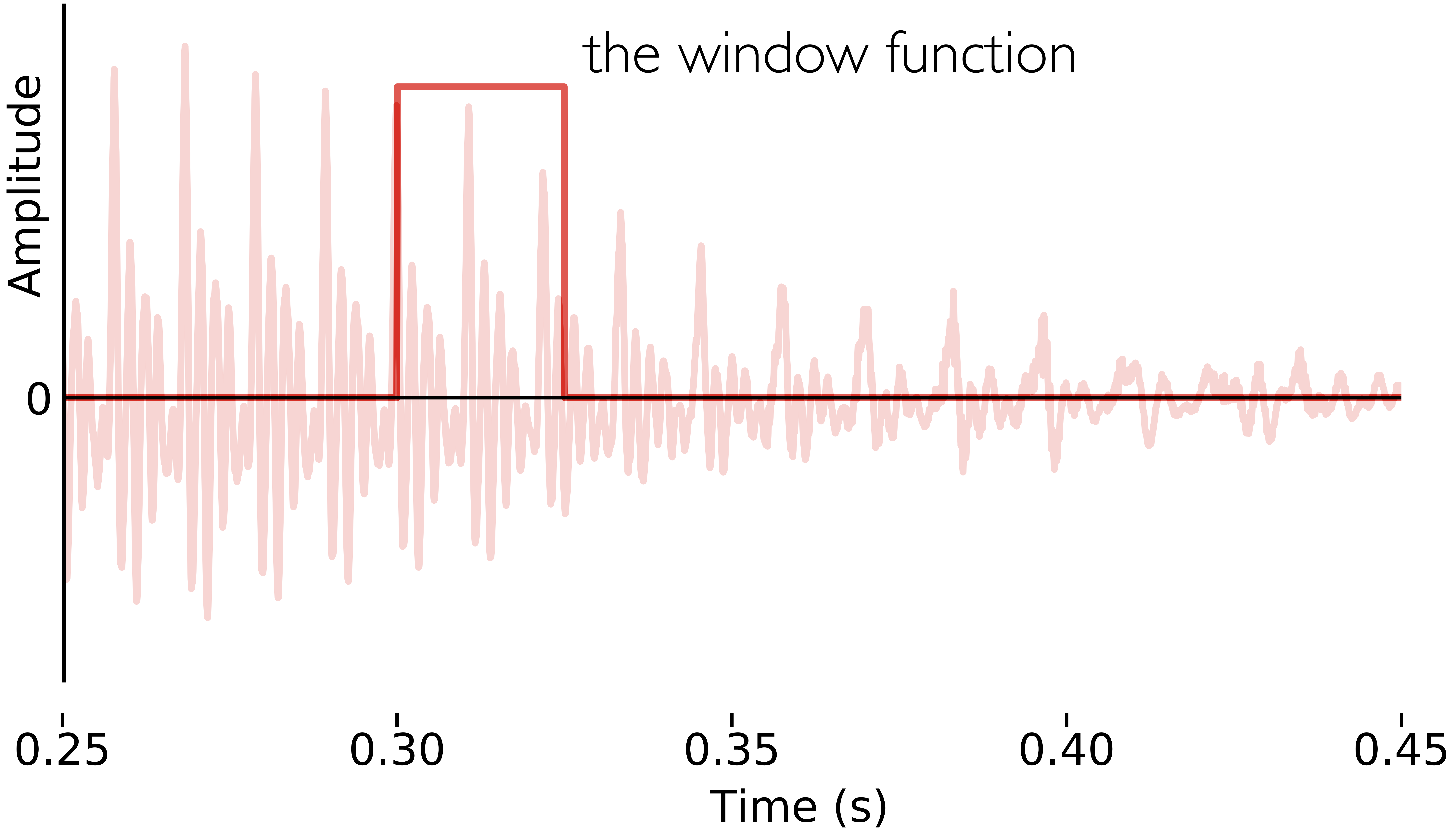
What you need to know already



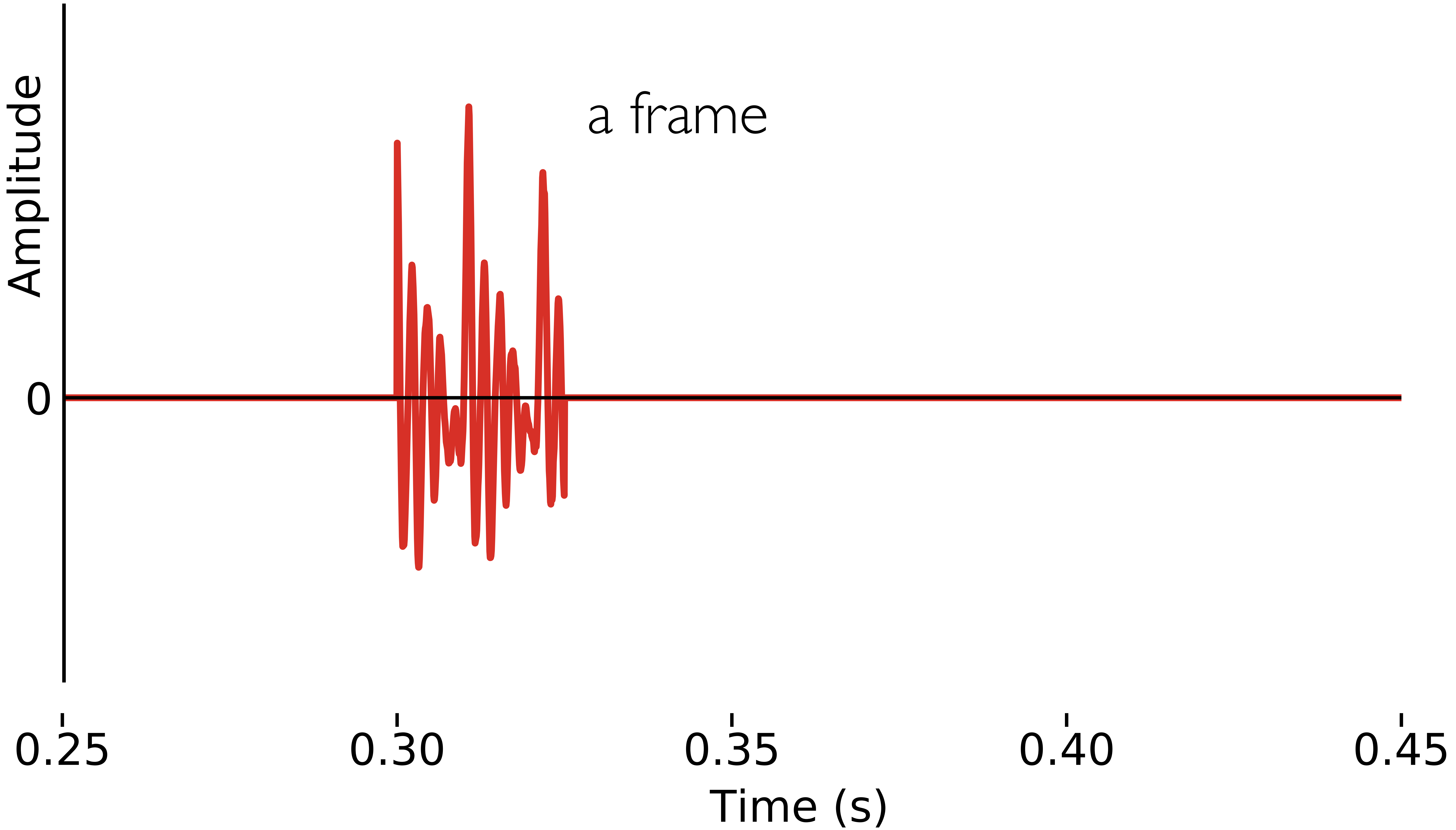
Speech waveforms vary over time



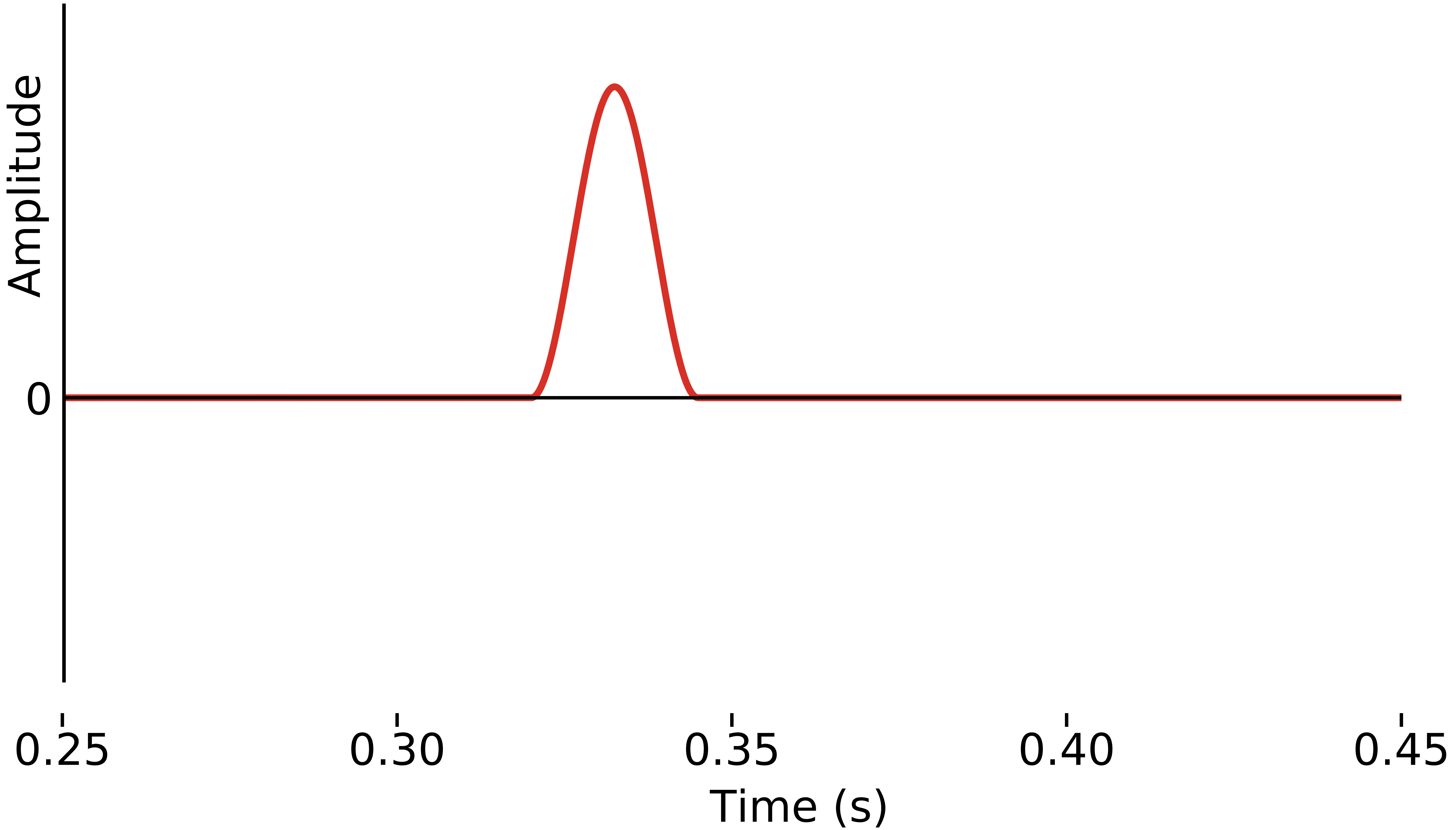
Short-term analysis - defining a frame of the waveform



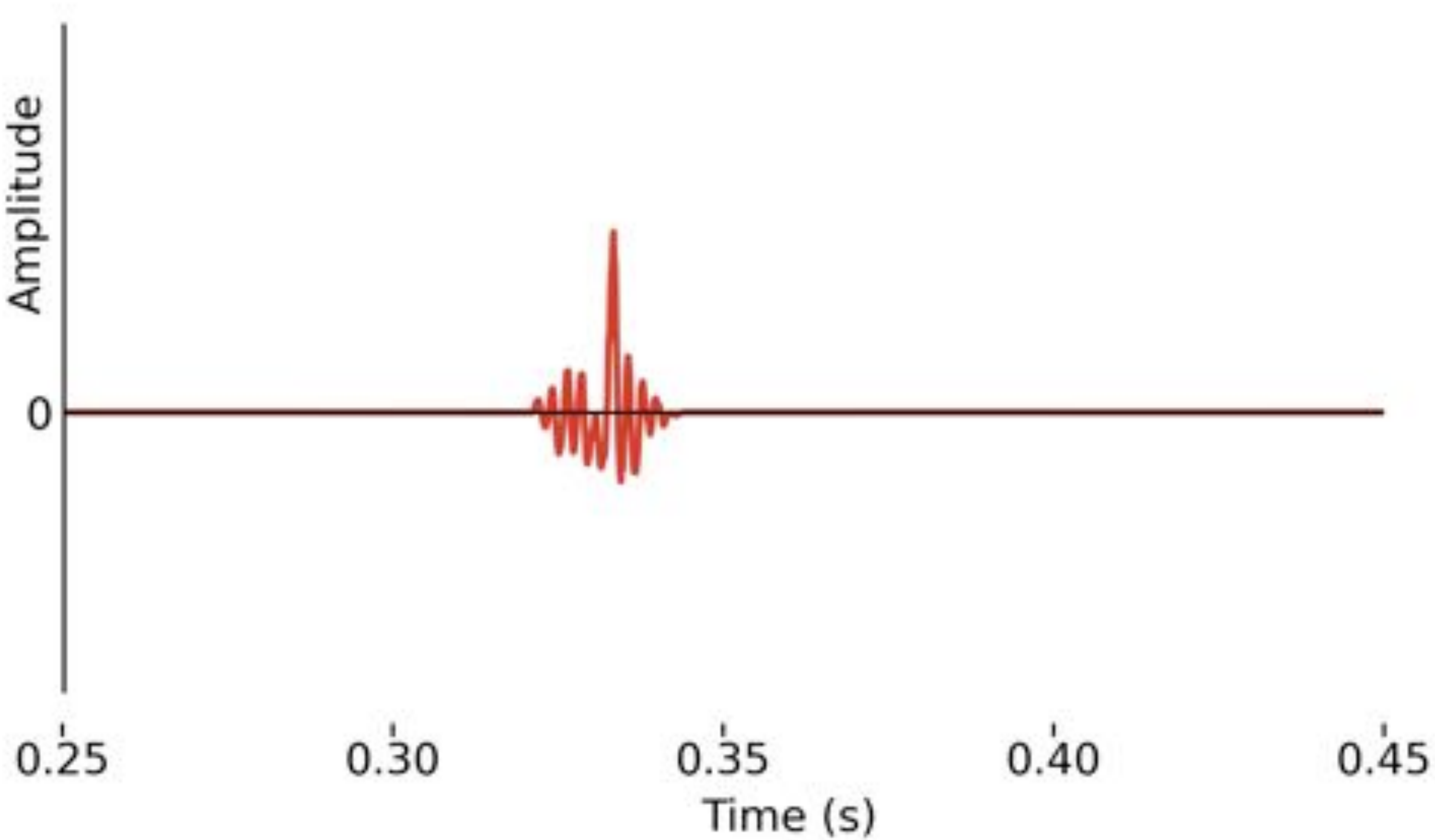
Short-term analysis - defining a frame of the waveform



Short-term analysis - applying a tapered window to a frame



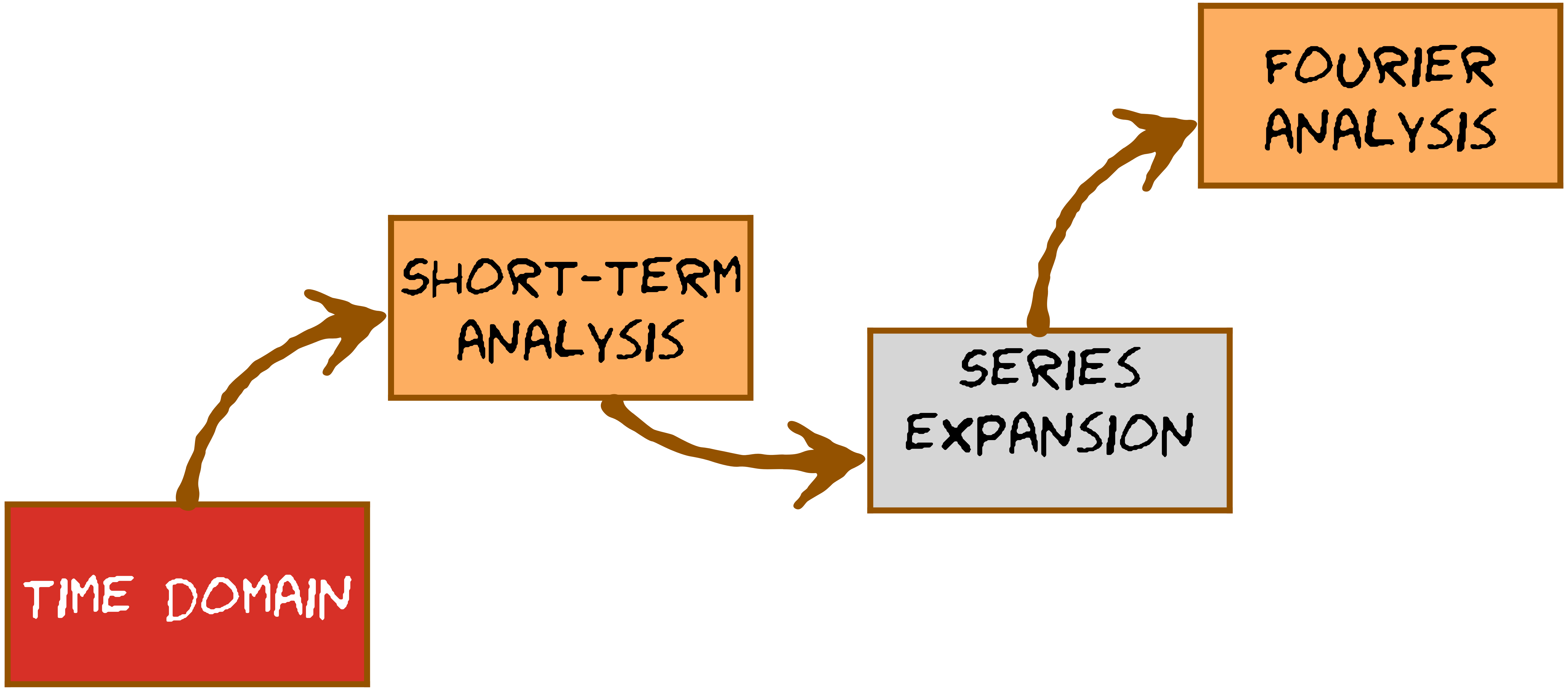
Short-term analysis - applying a tapered window to a frame



Typical values:

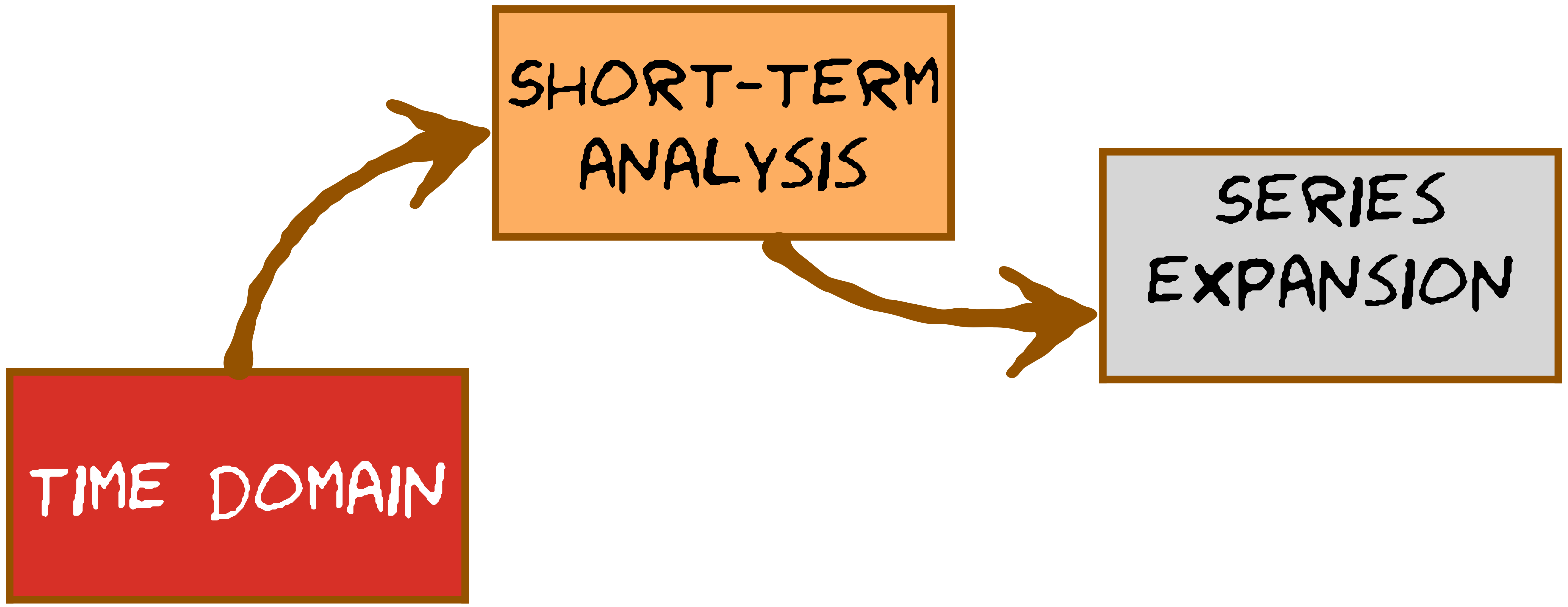
frame duration 25 ms
frame shift 10 ms

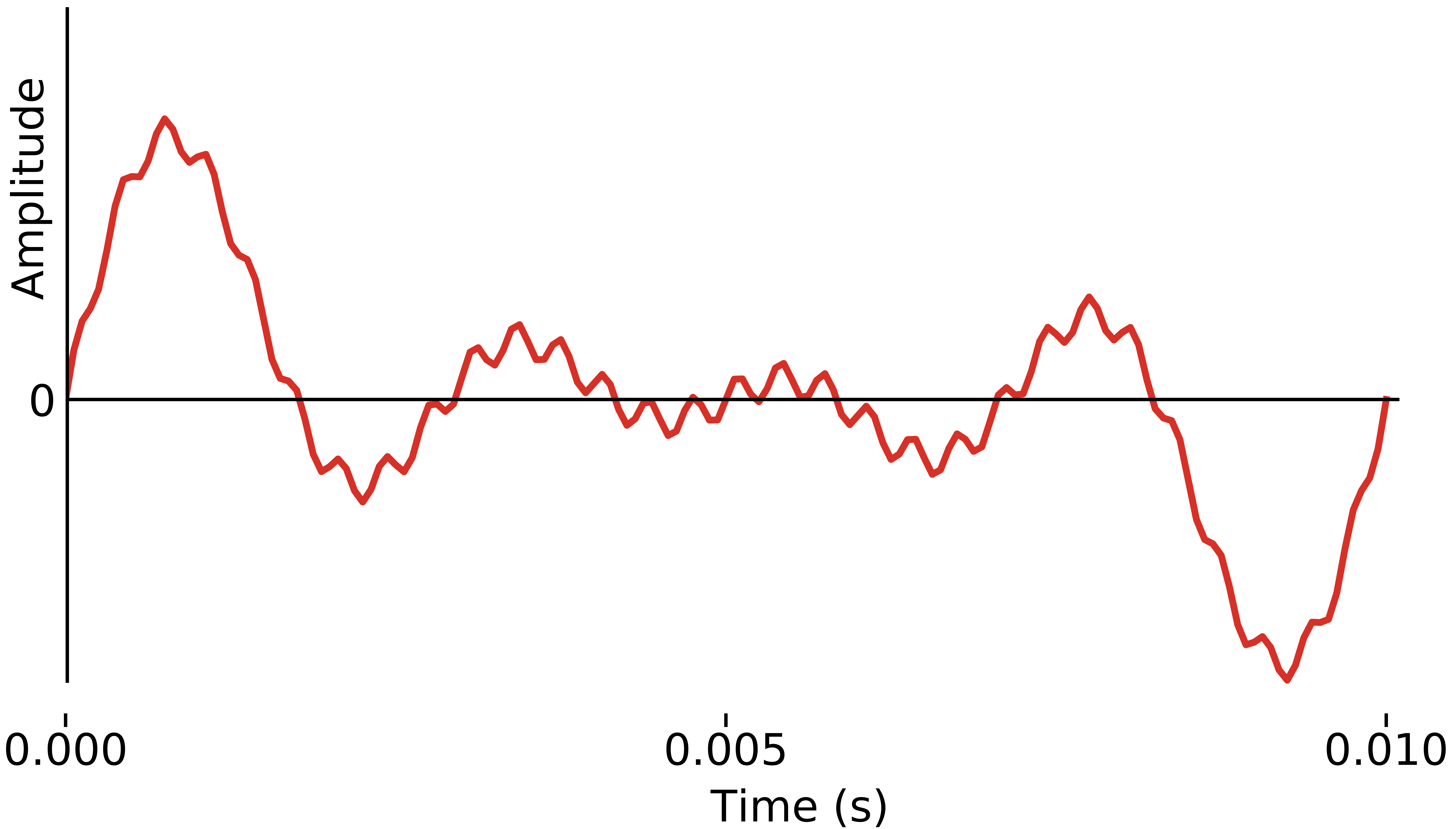
What you can learn next

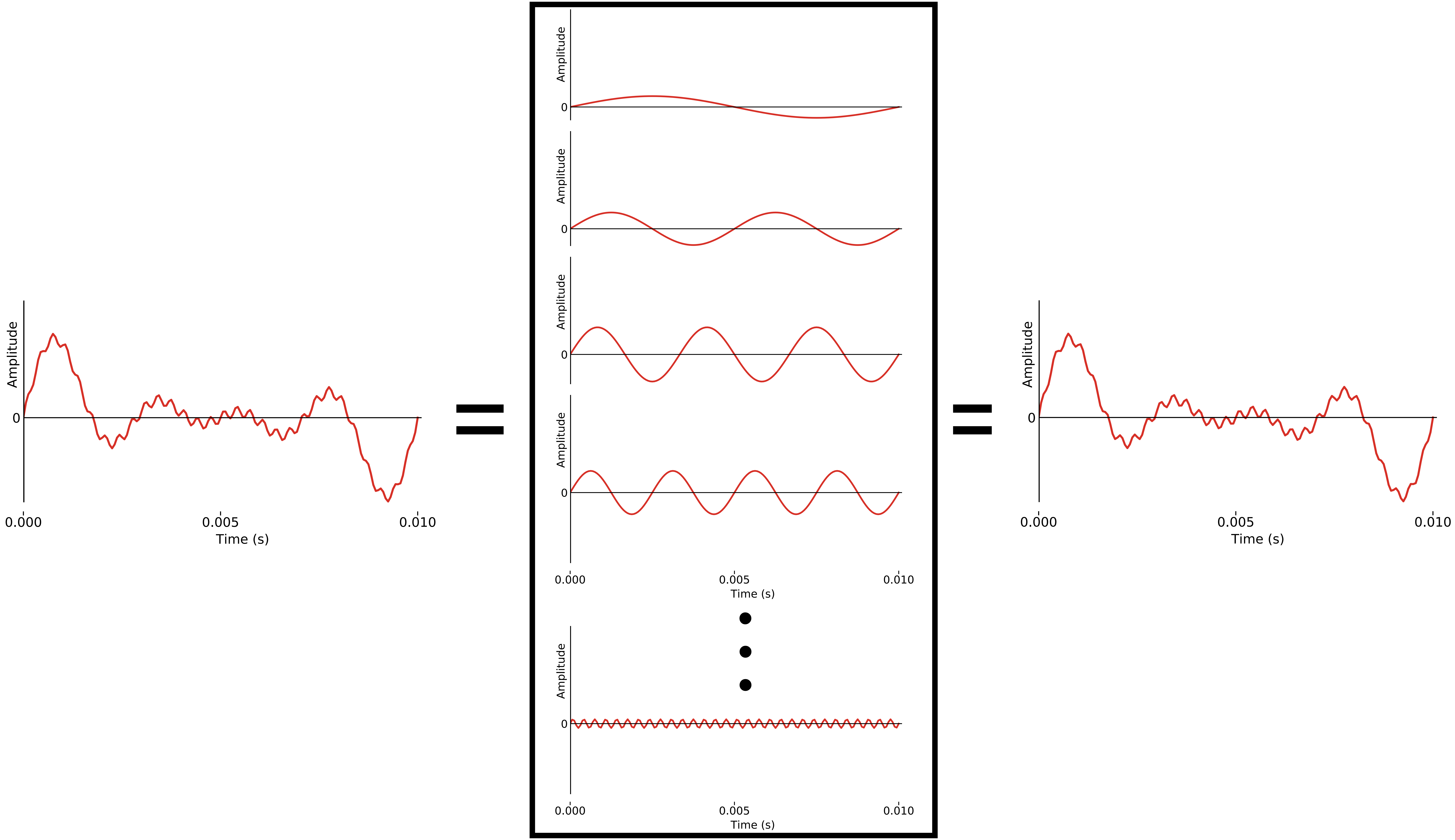


SERIES EXPANSION

What you need to know already

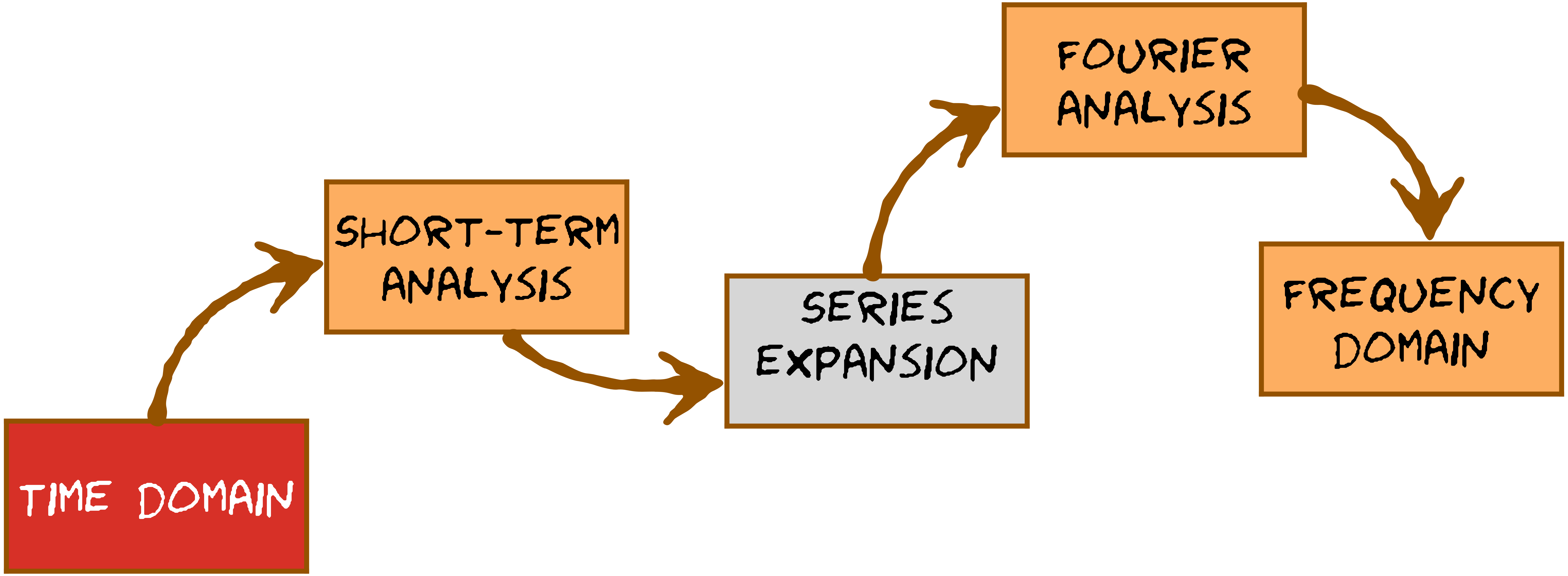



Amplitude
0
0.000
0.005
0.010
Time (s)



Amplitude
0
0.000
0.005
0.010
Time (s)

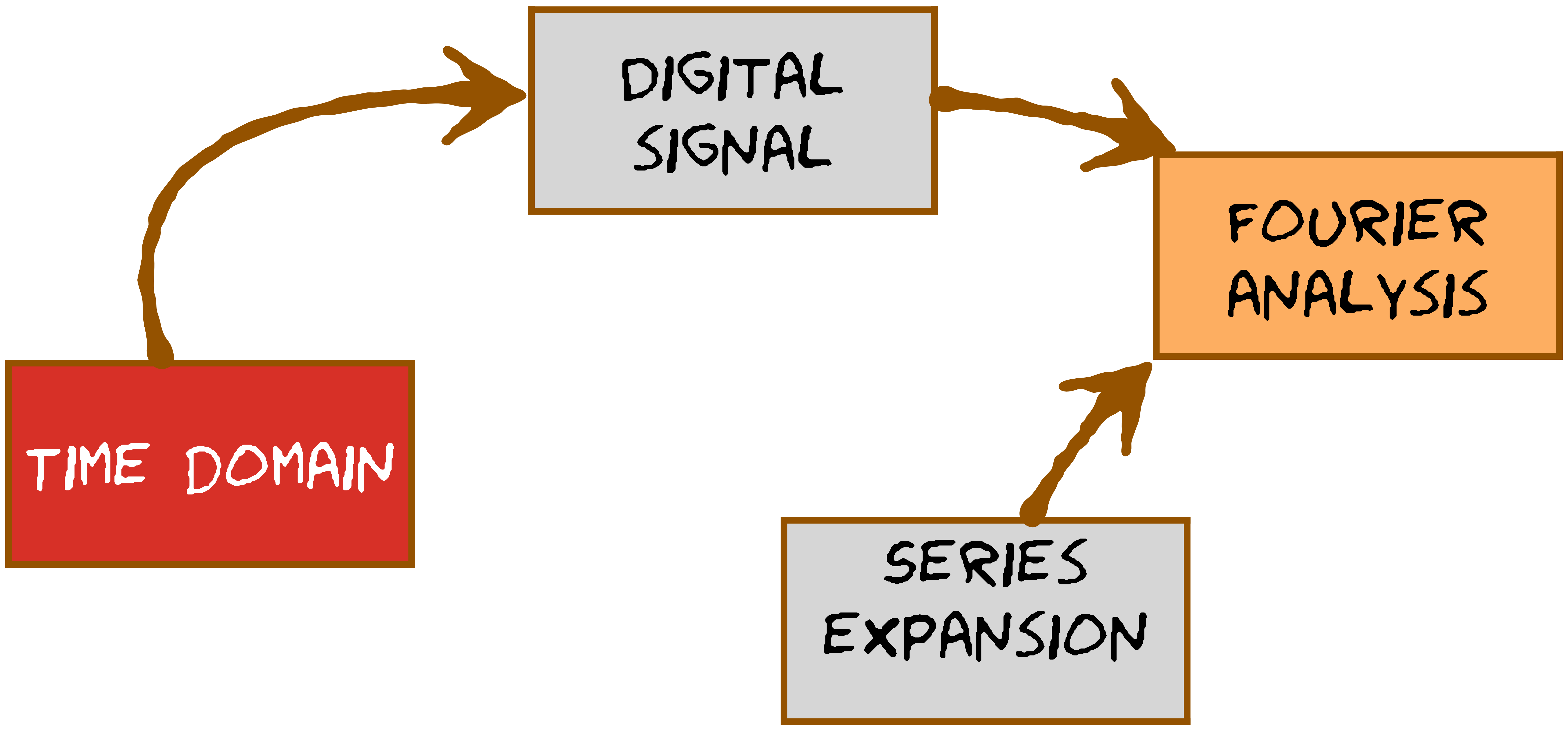
What you can learn next

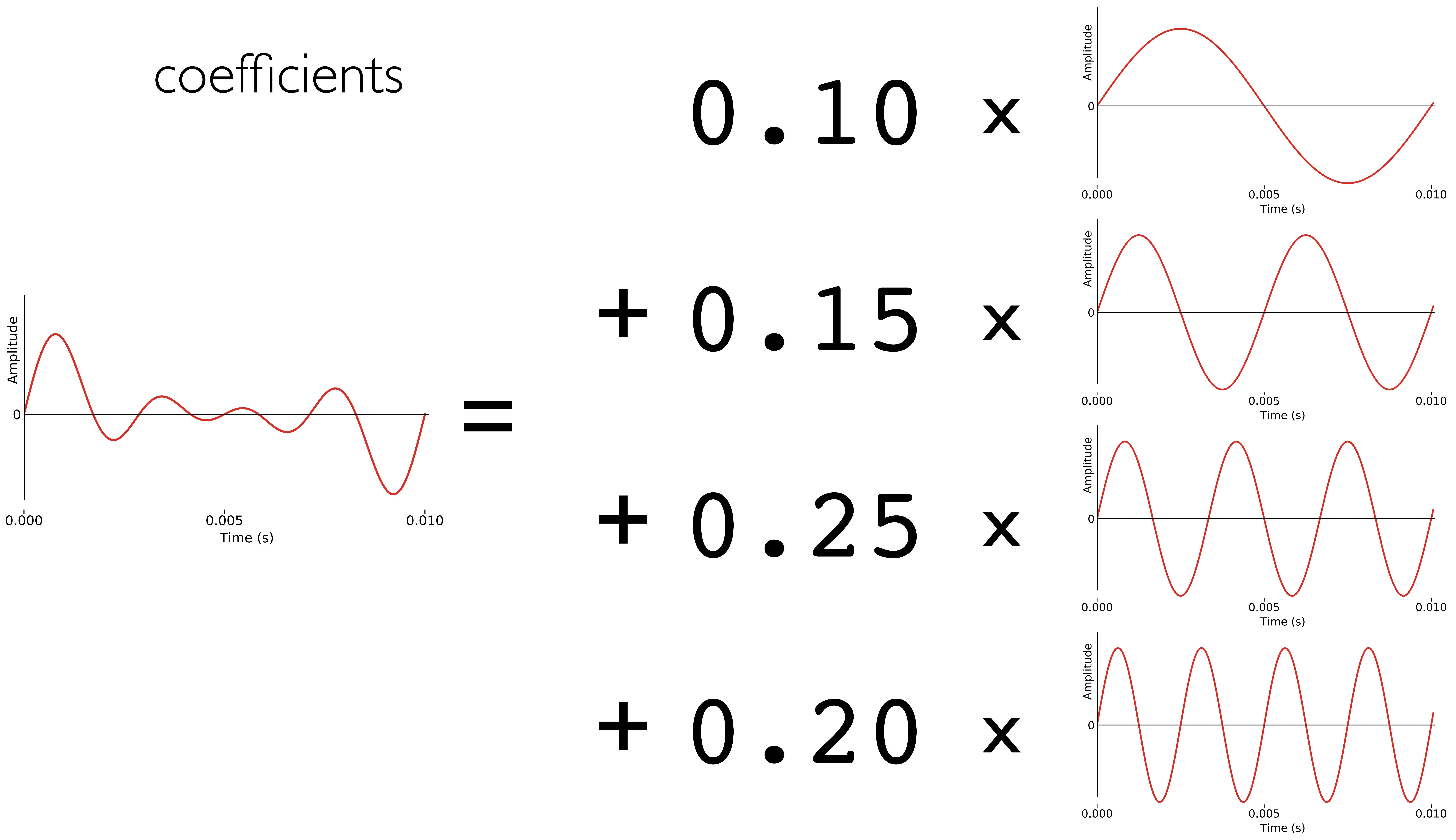


FOURIER ANALYSIS

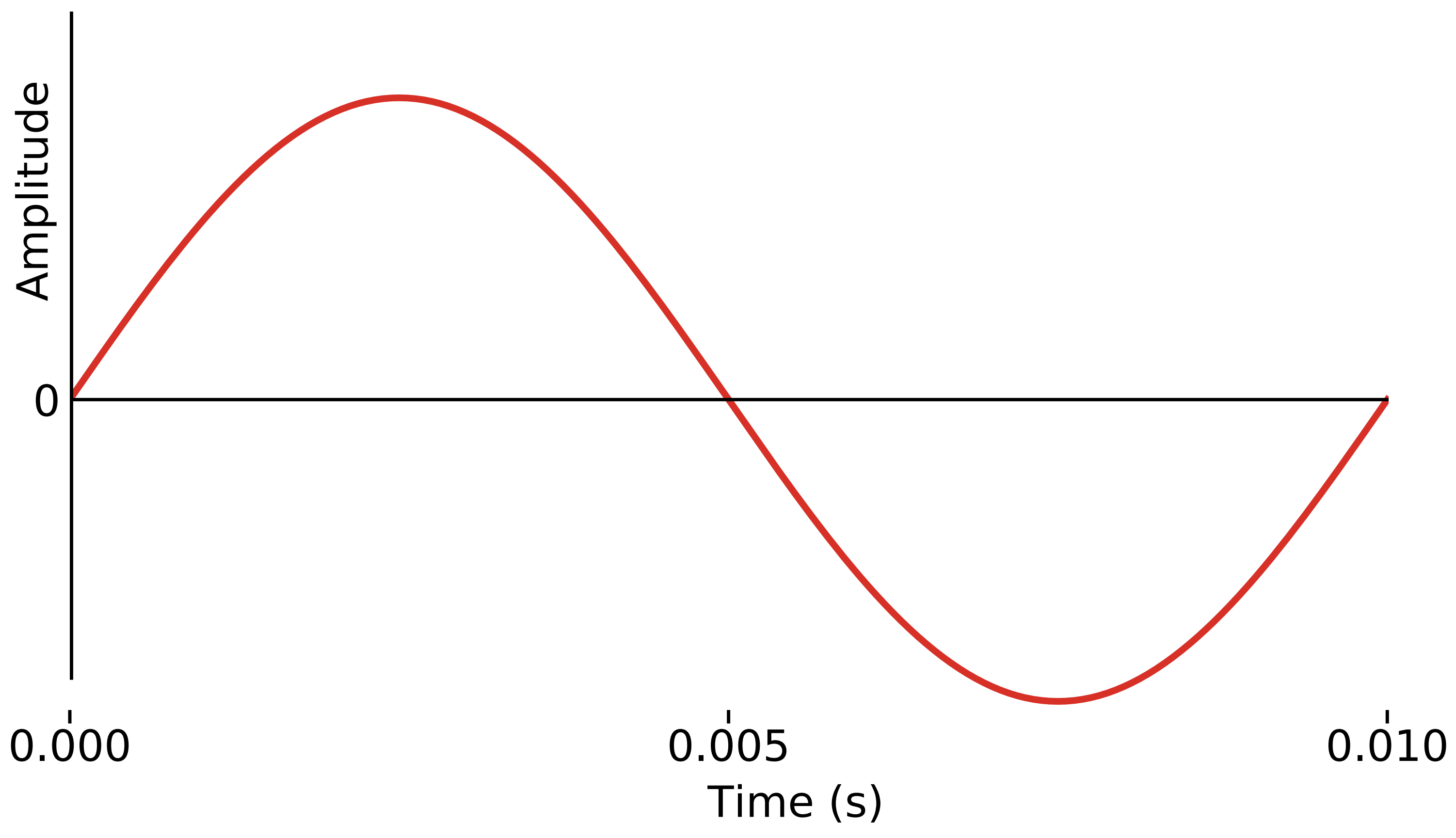
FREQUENCY DOMAIN AND BEYOND

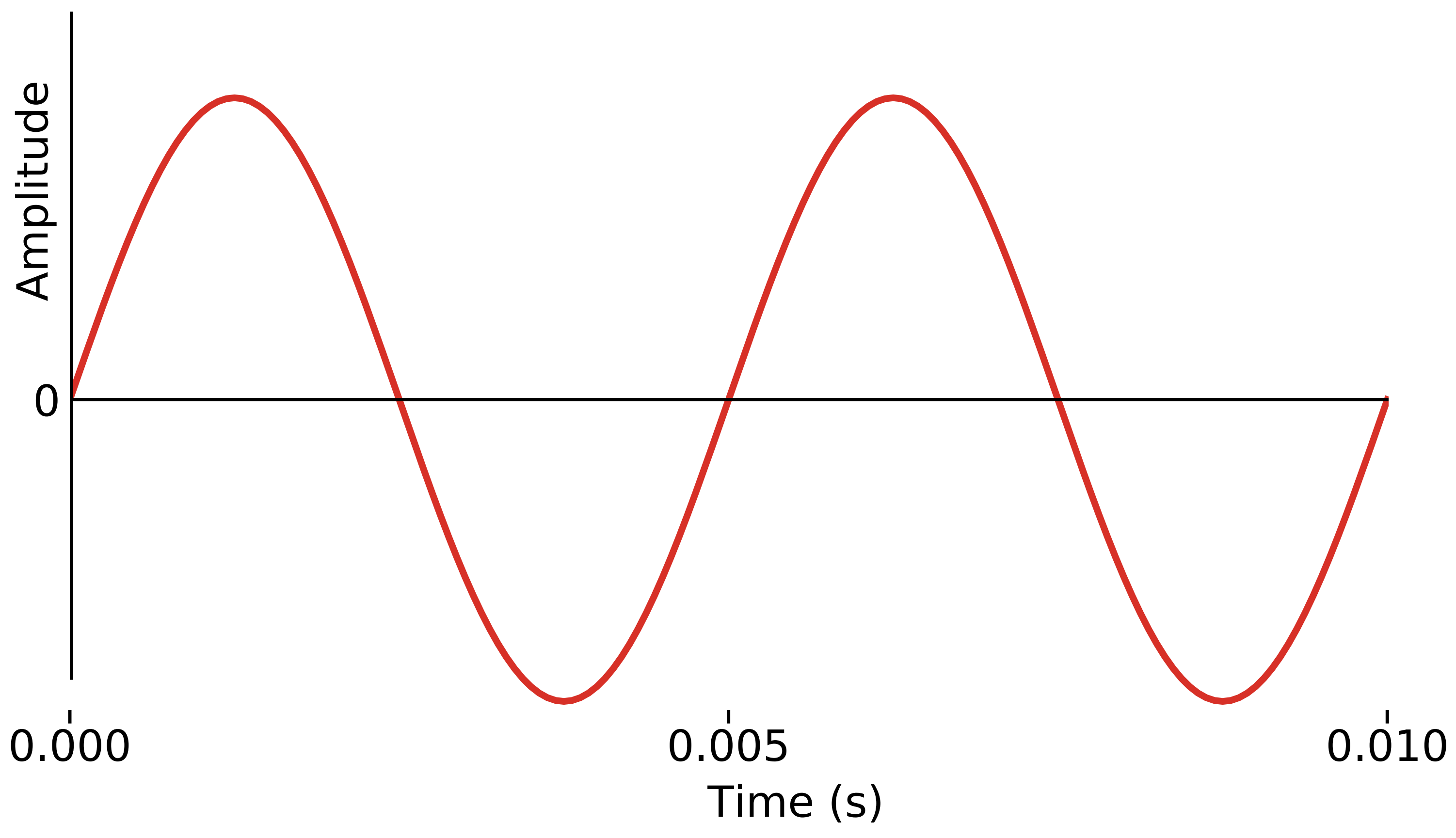
What you need to know already

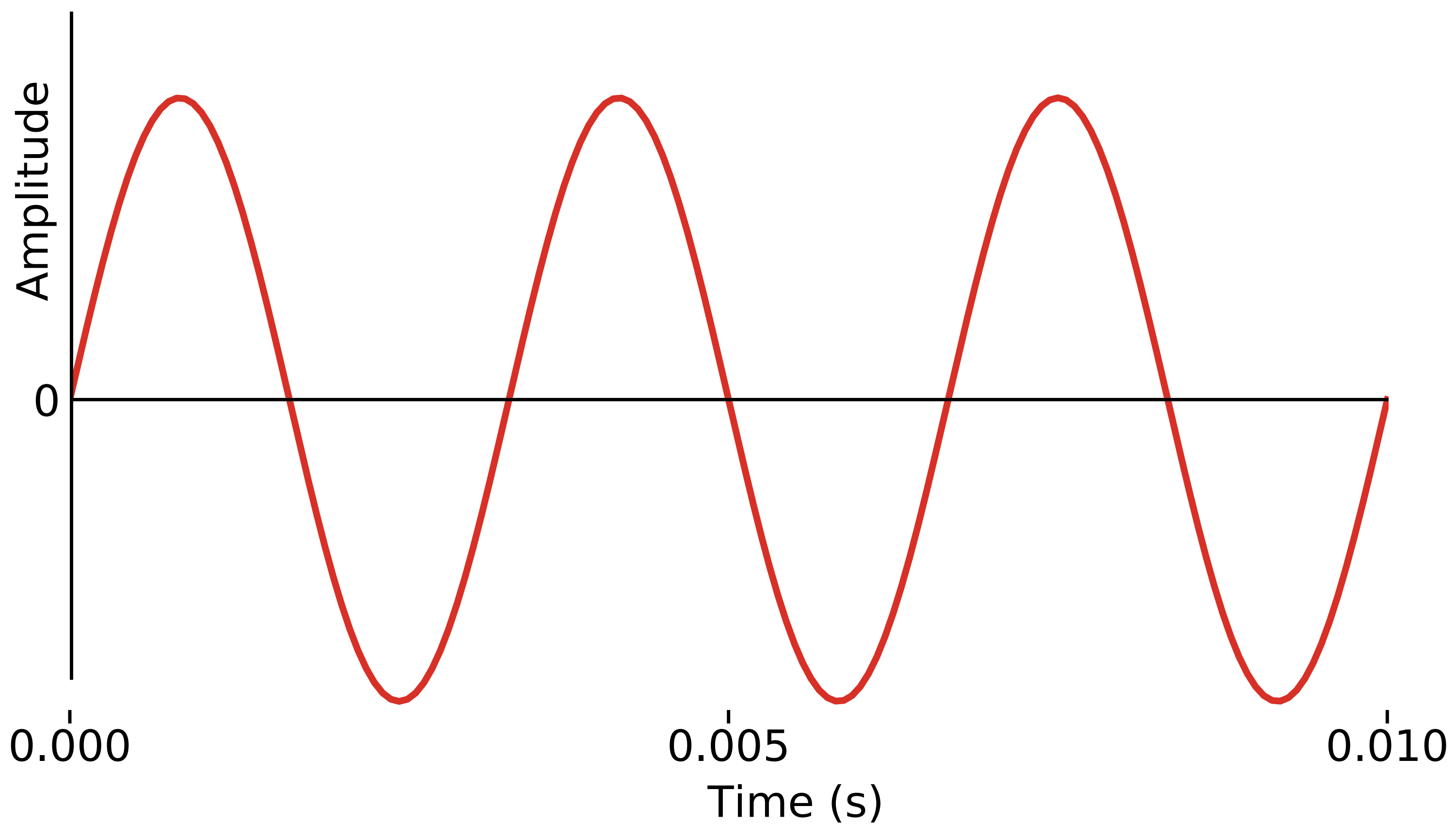


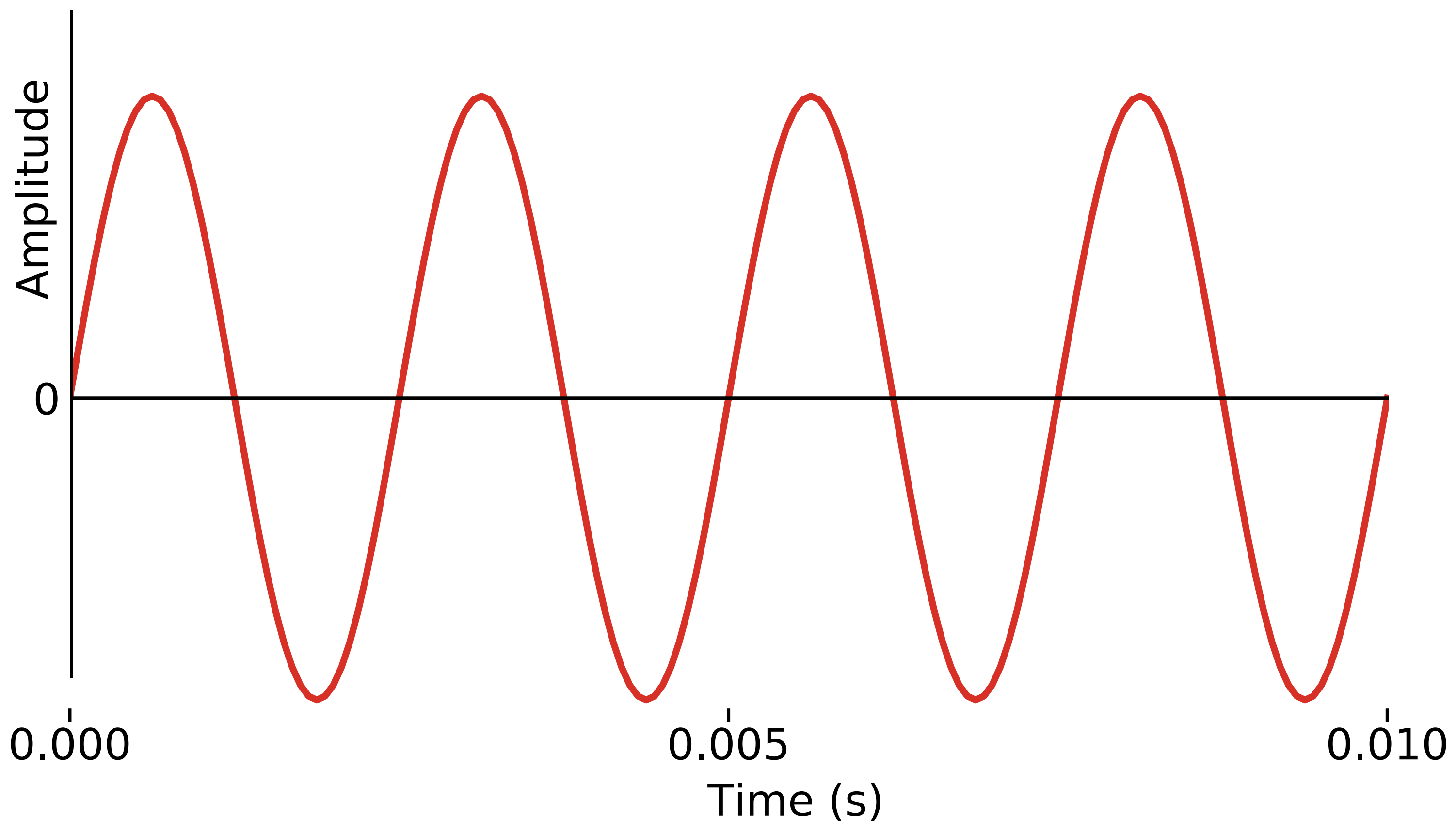

coefficients
Amplitude
0
0.000
0.005
0.010
Time (s)
=
0.10 x
+ 0.15 x
+ 0.25 x
+ 0.20 x
Amplitude
0
0.000
0.005
0.010
Time (s)

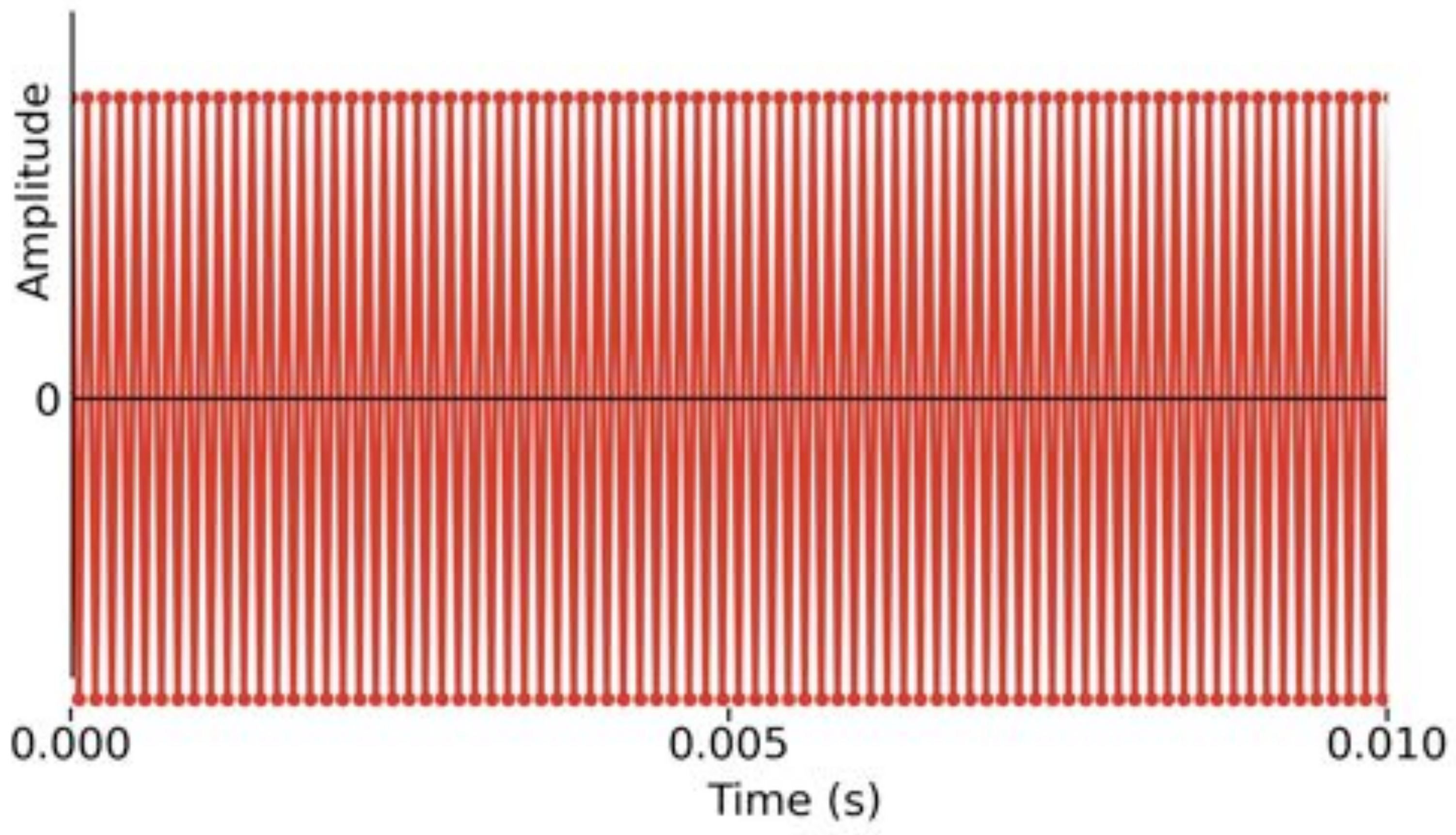
The basis functions



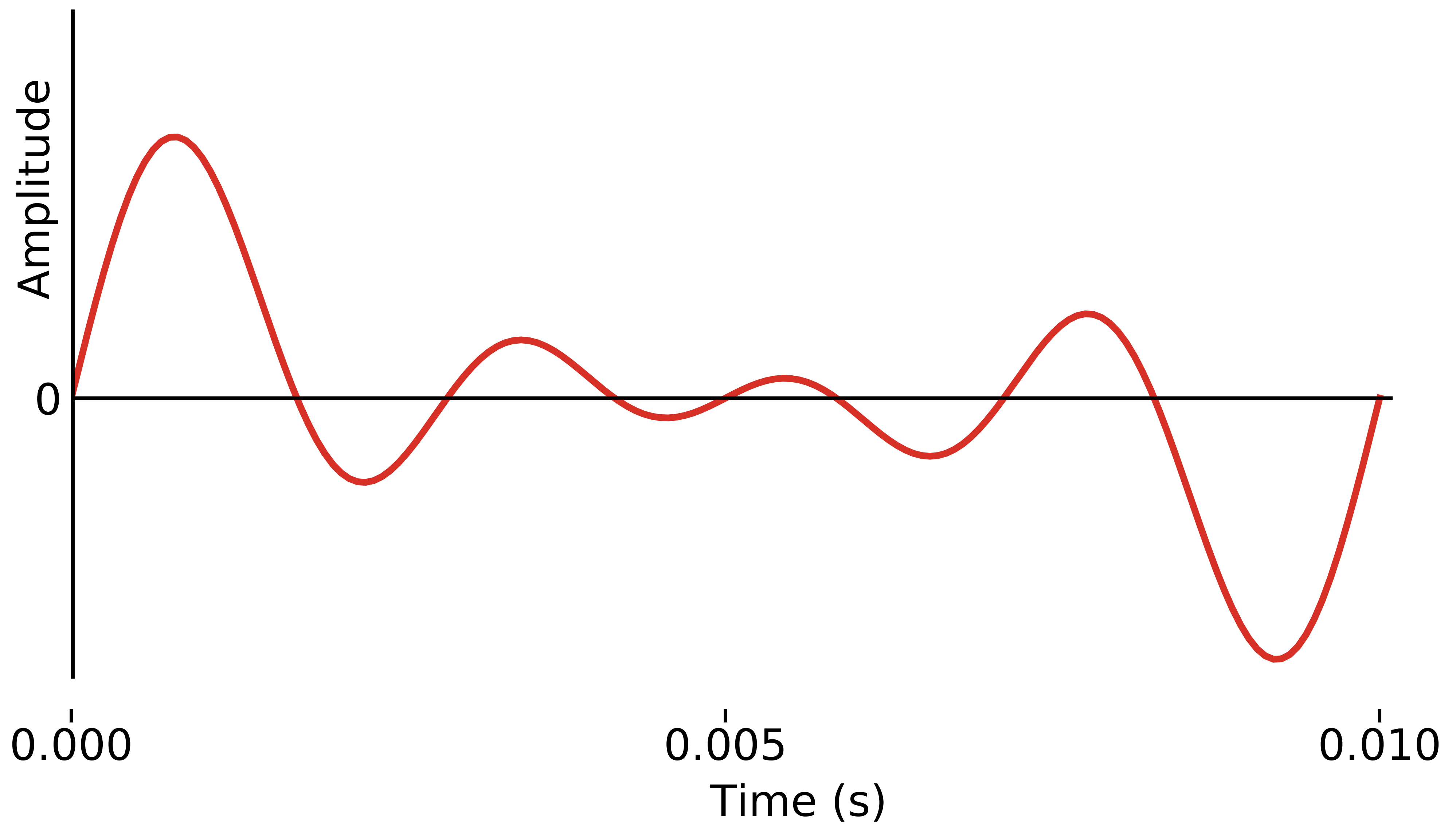


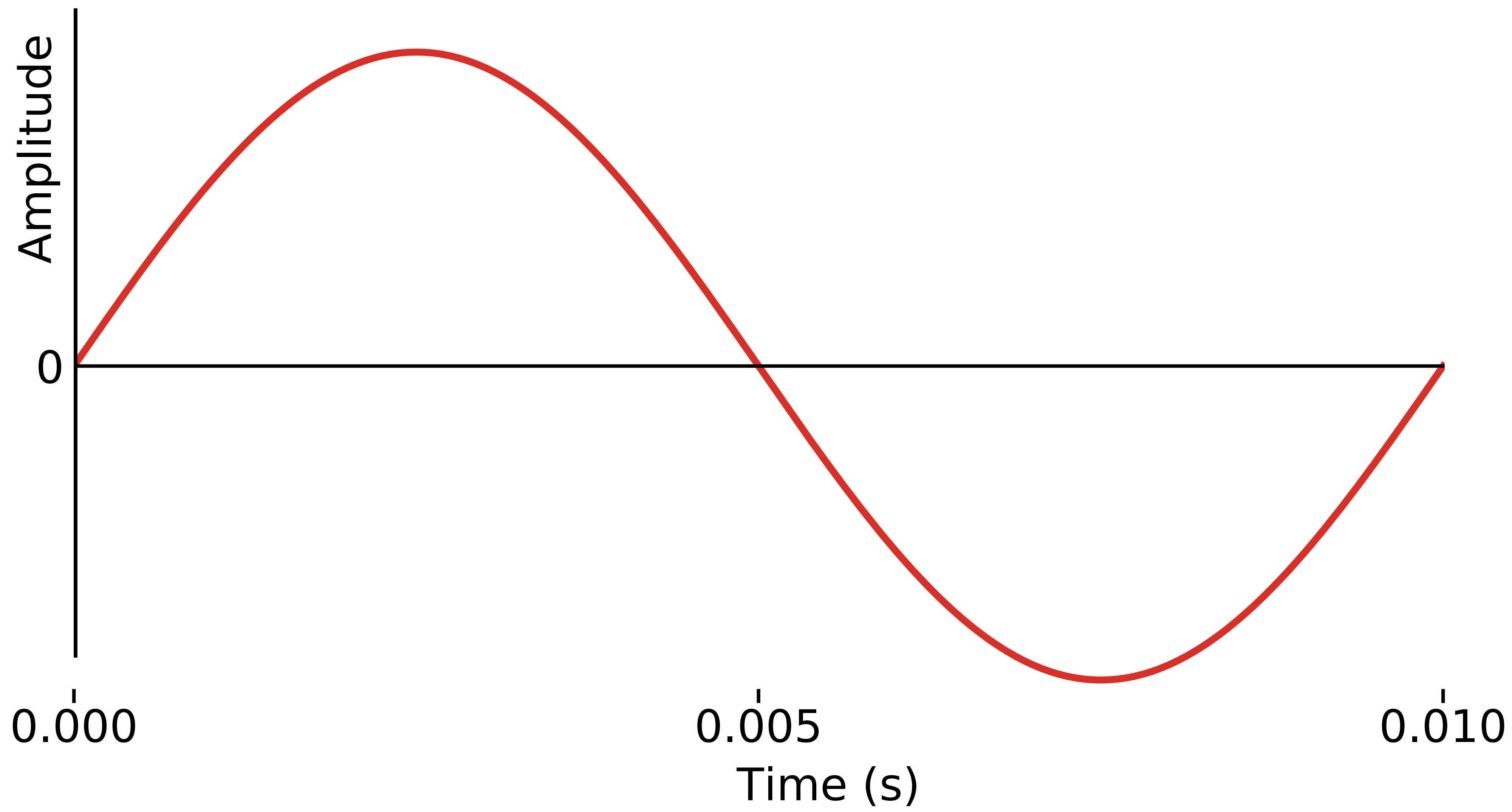




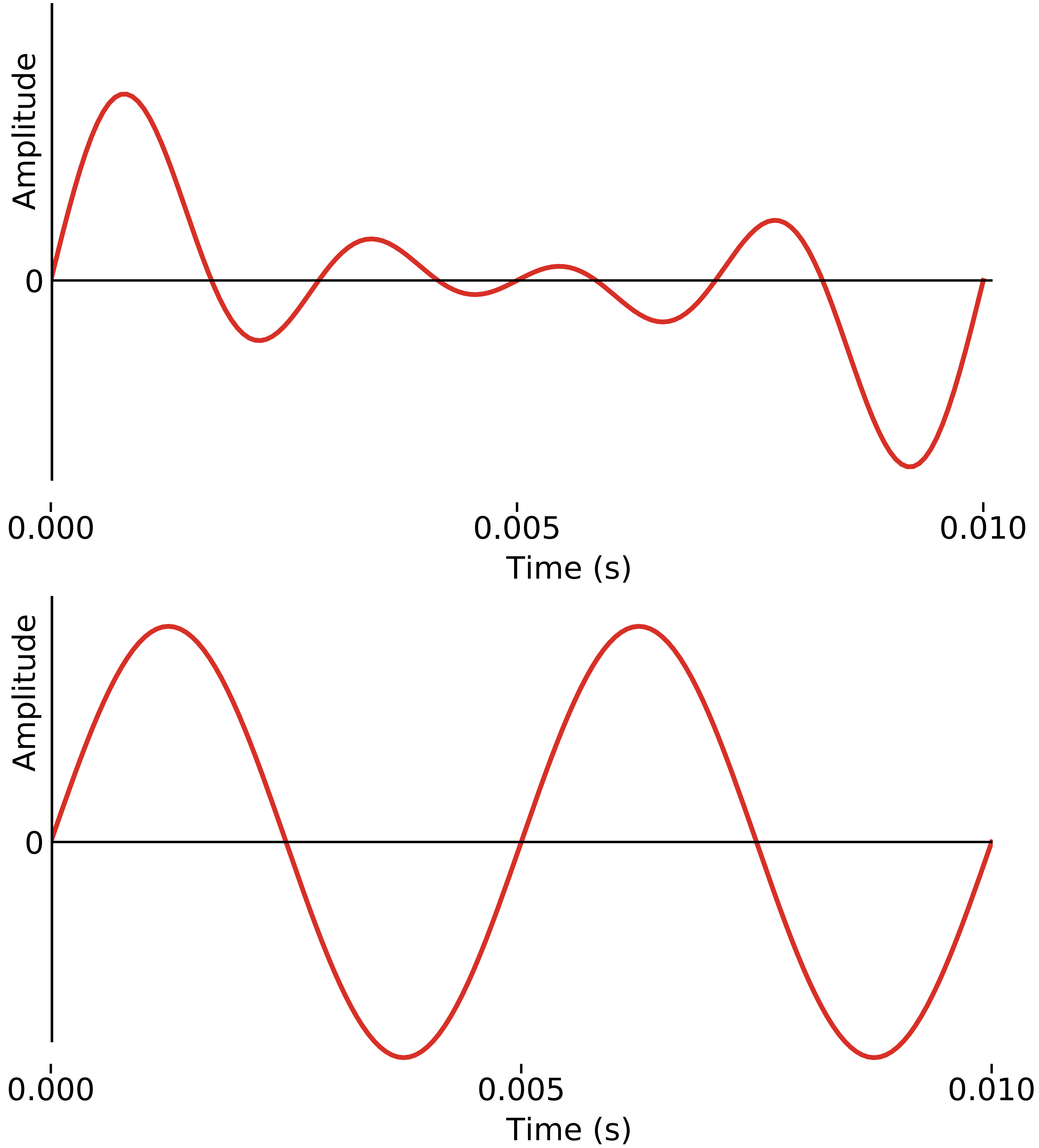


Fourier analysis = finding the coefficients

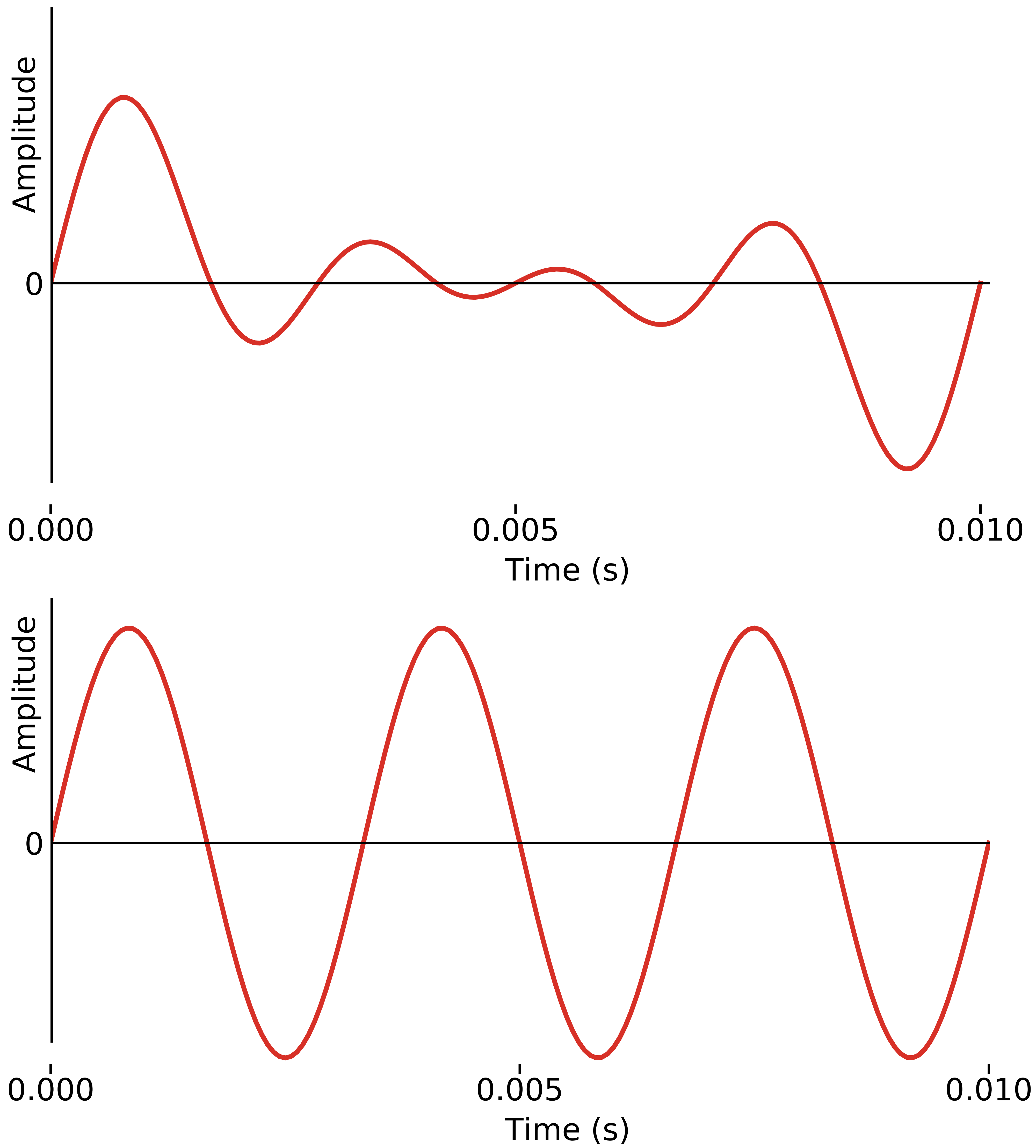




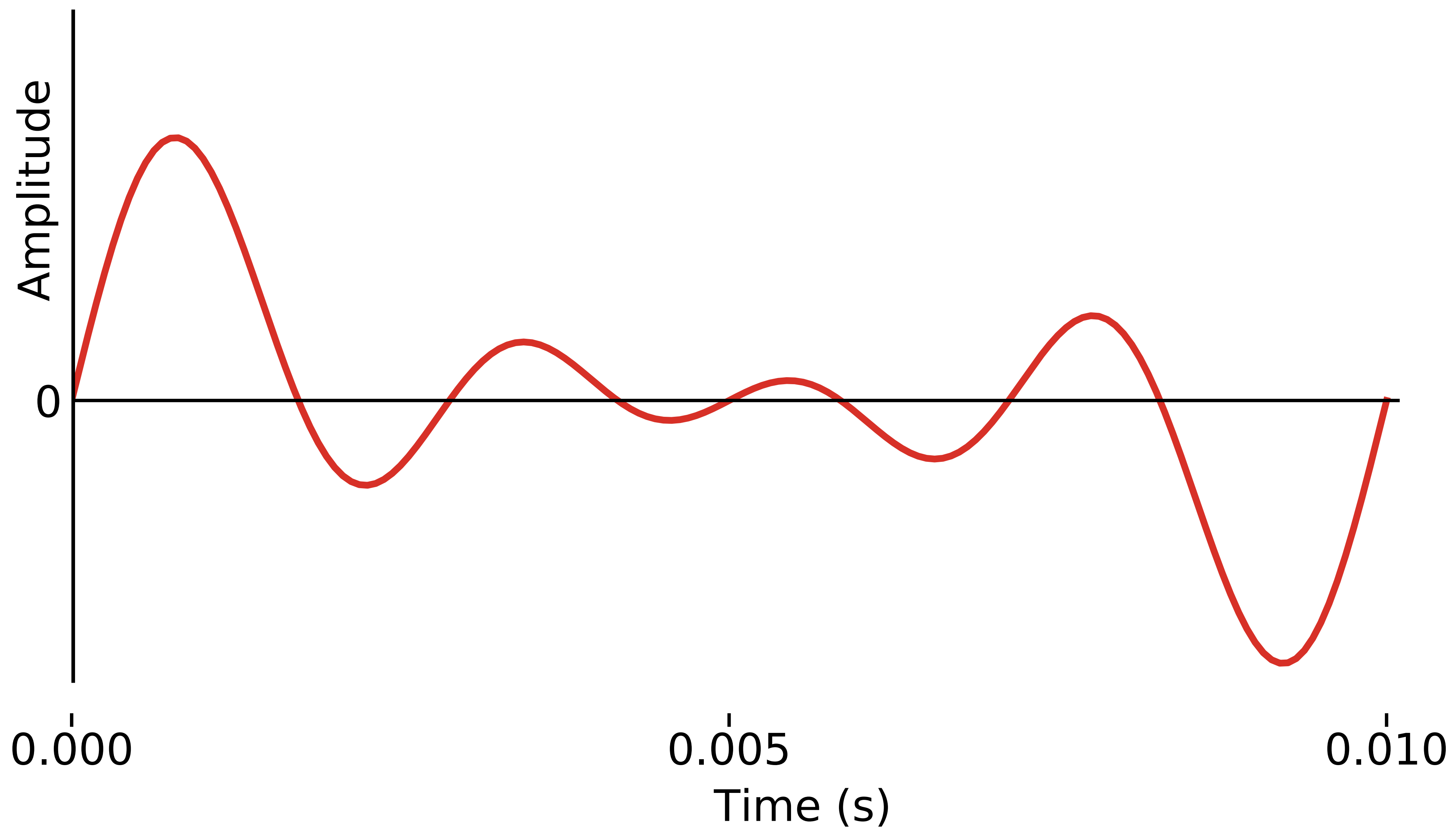
Fourier analysis = finding the coefficients

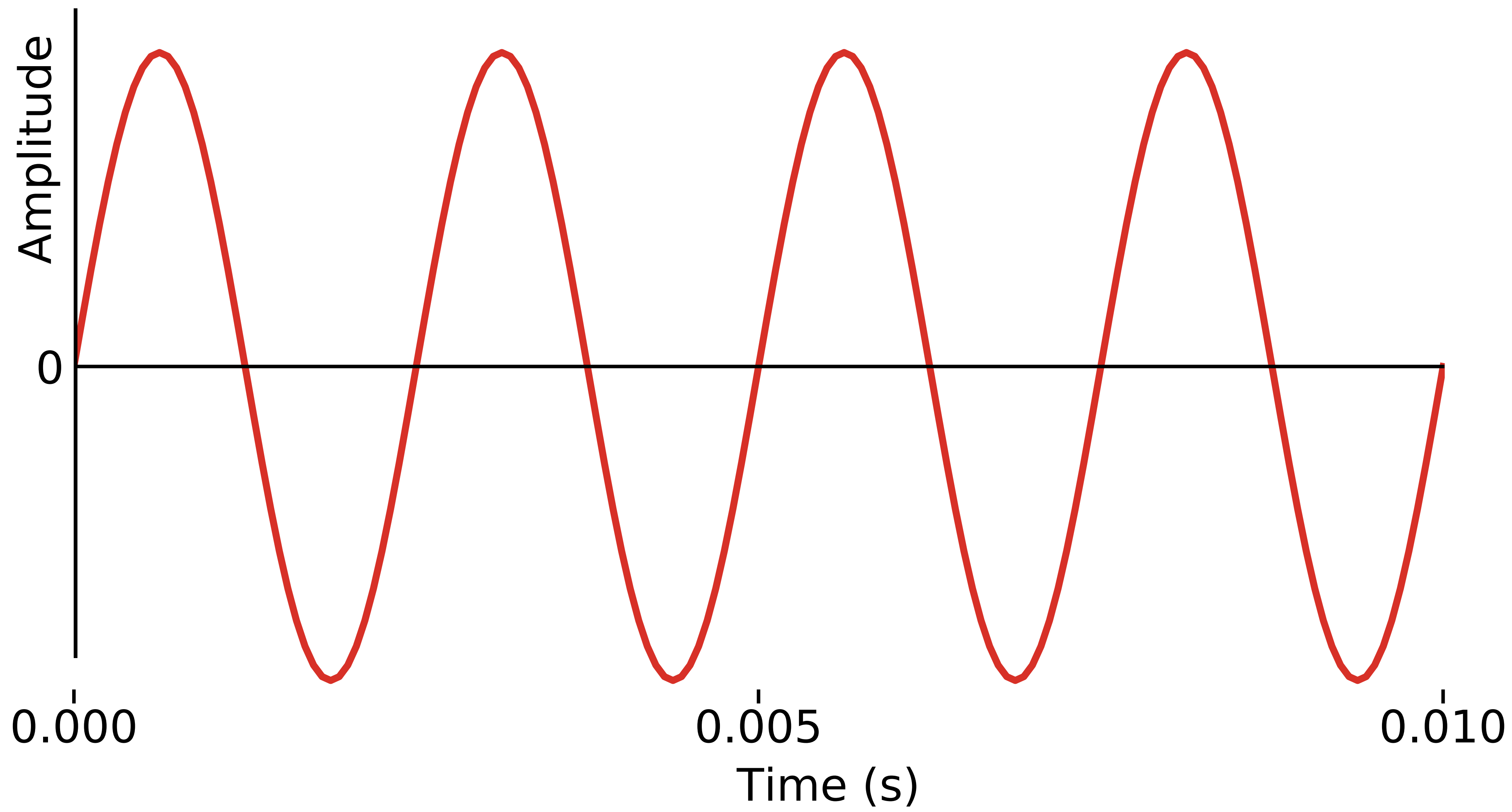


Fourier analysis = finding the coefficients

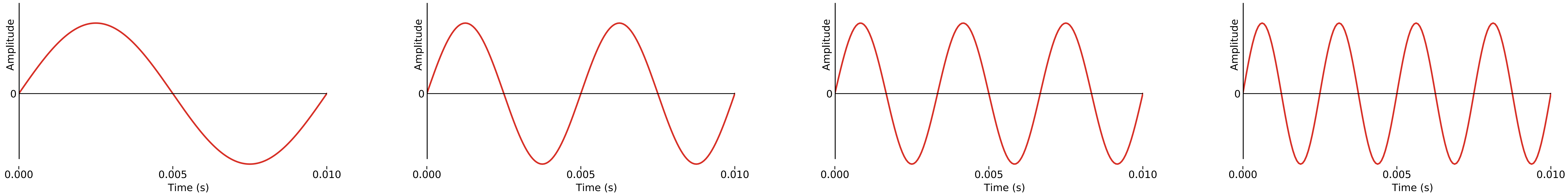


Fourier analysis = finding the coefficients

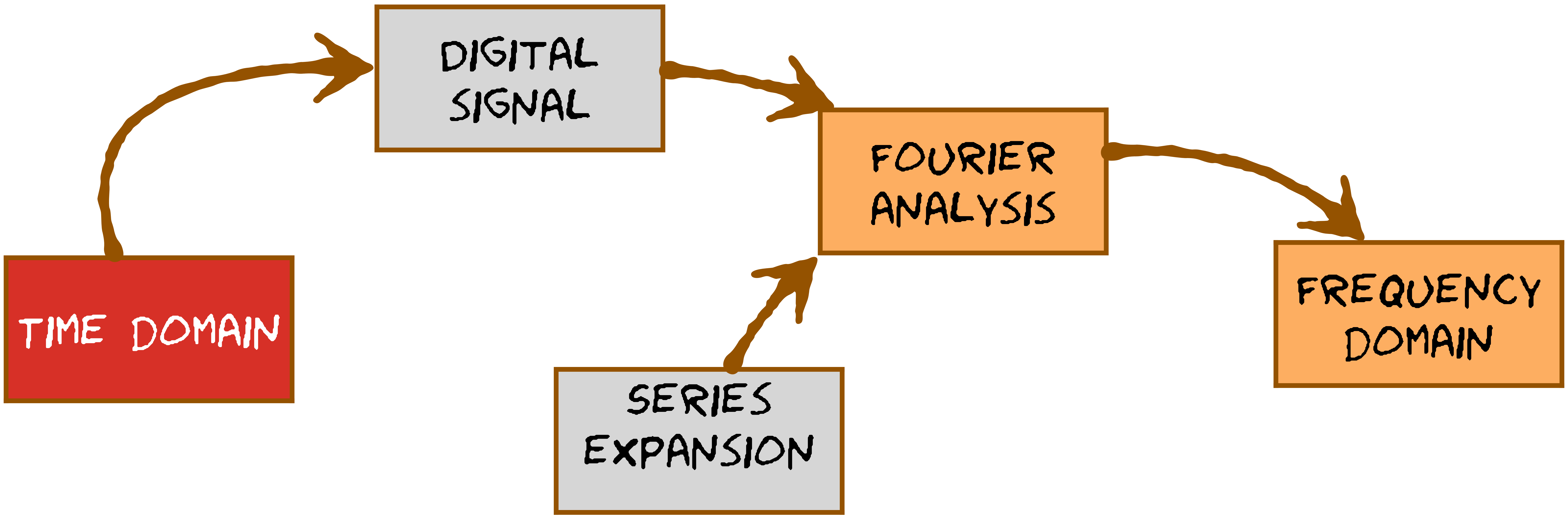




The basis functions are orthogonal

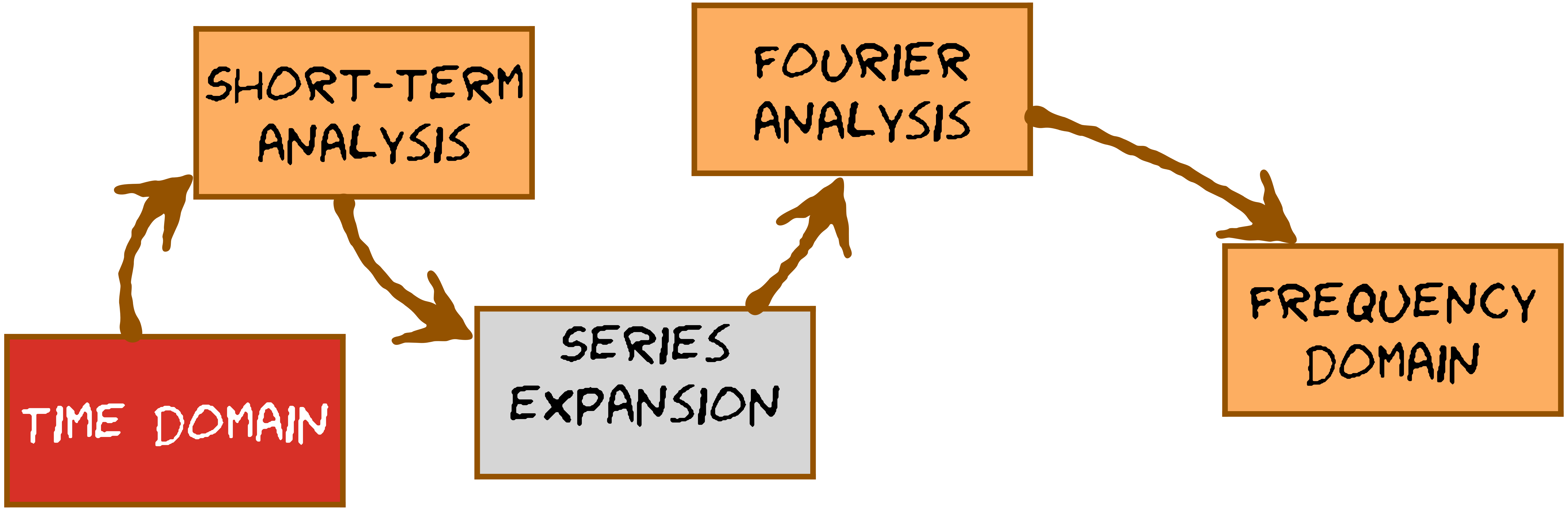


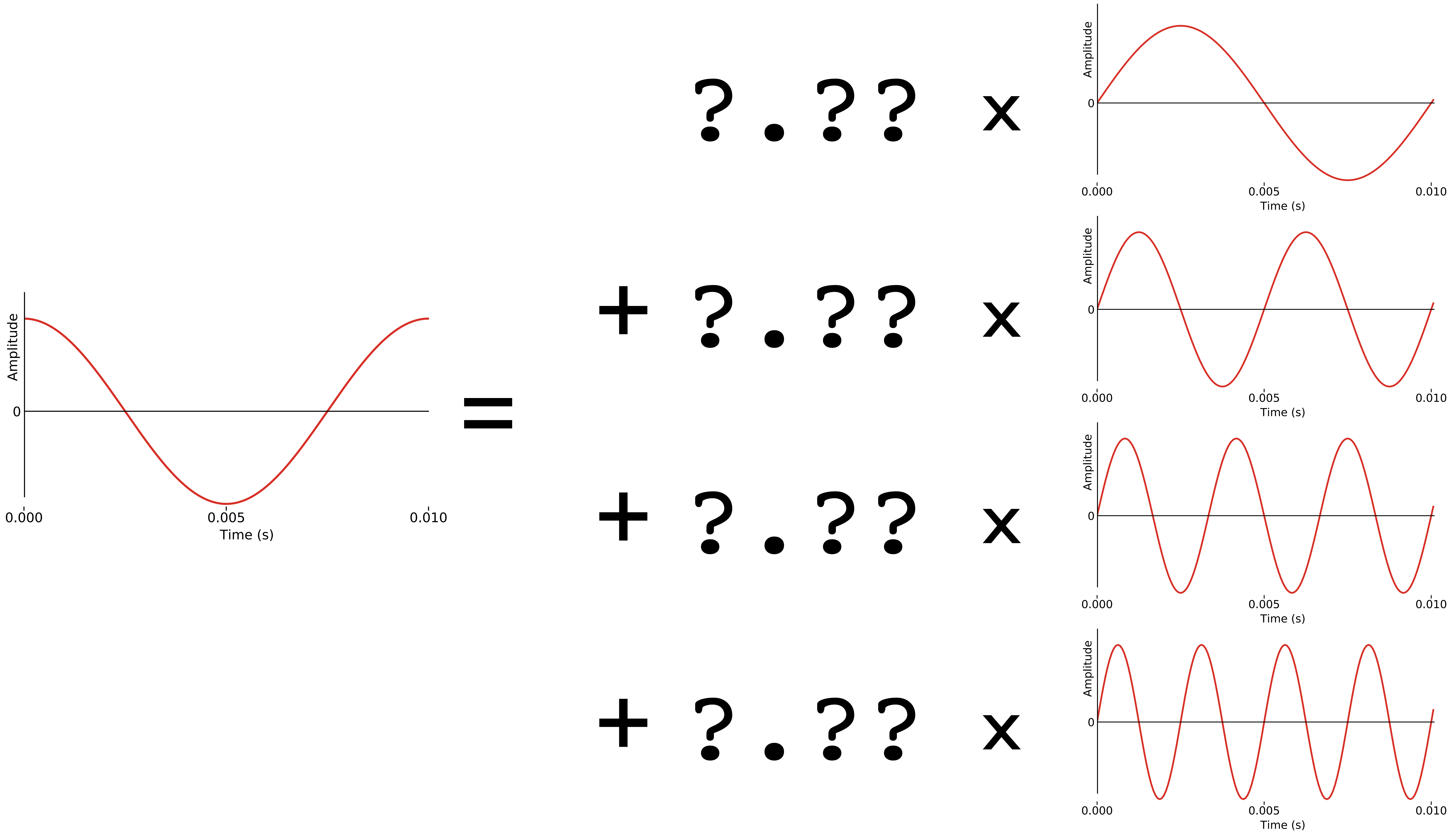
What you can learn next



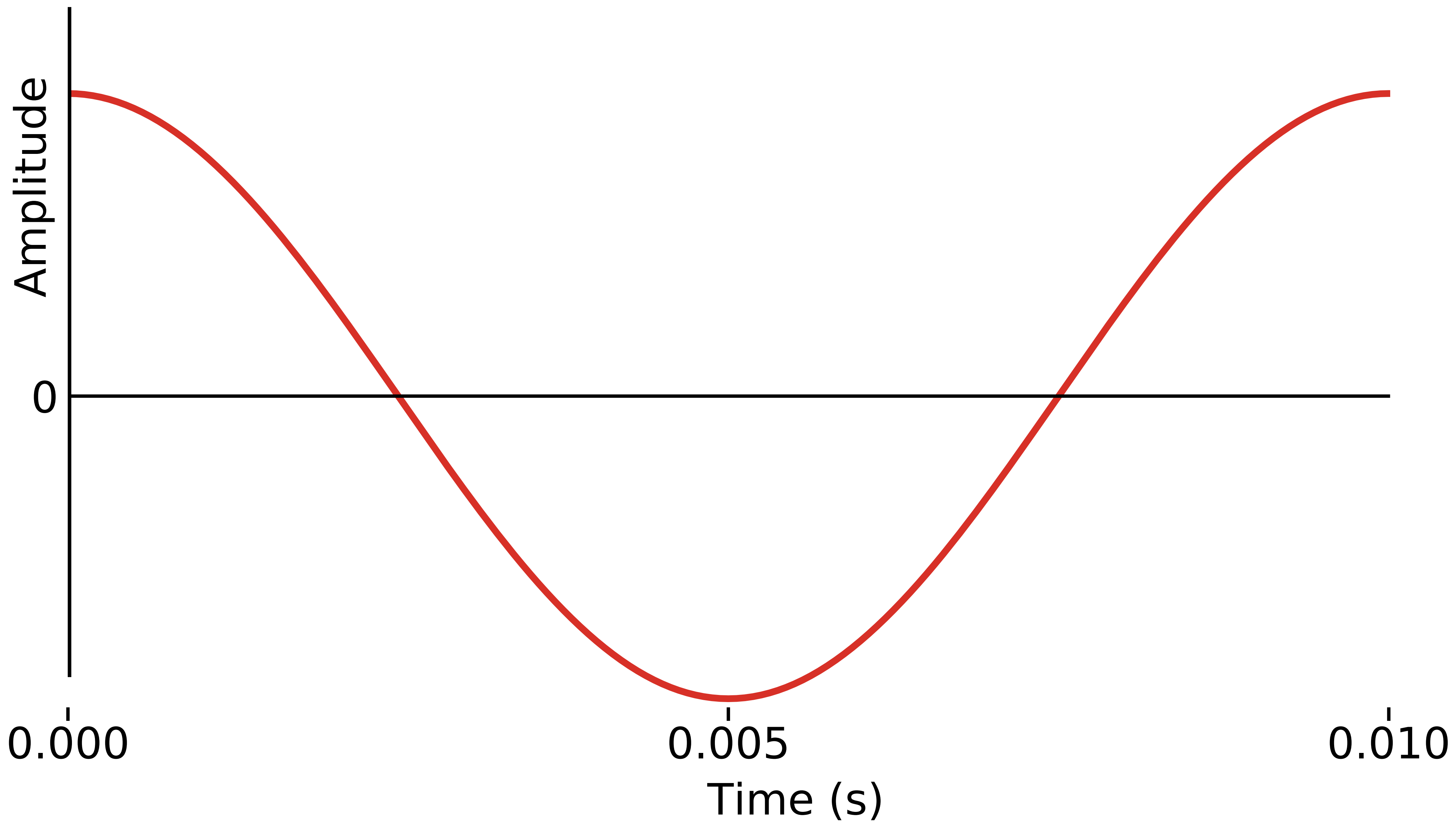
FREQUENCY DOMAIN

What you need to know already

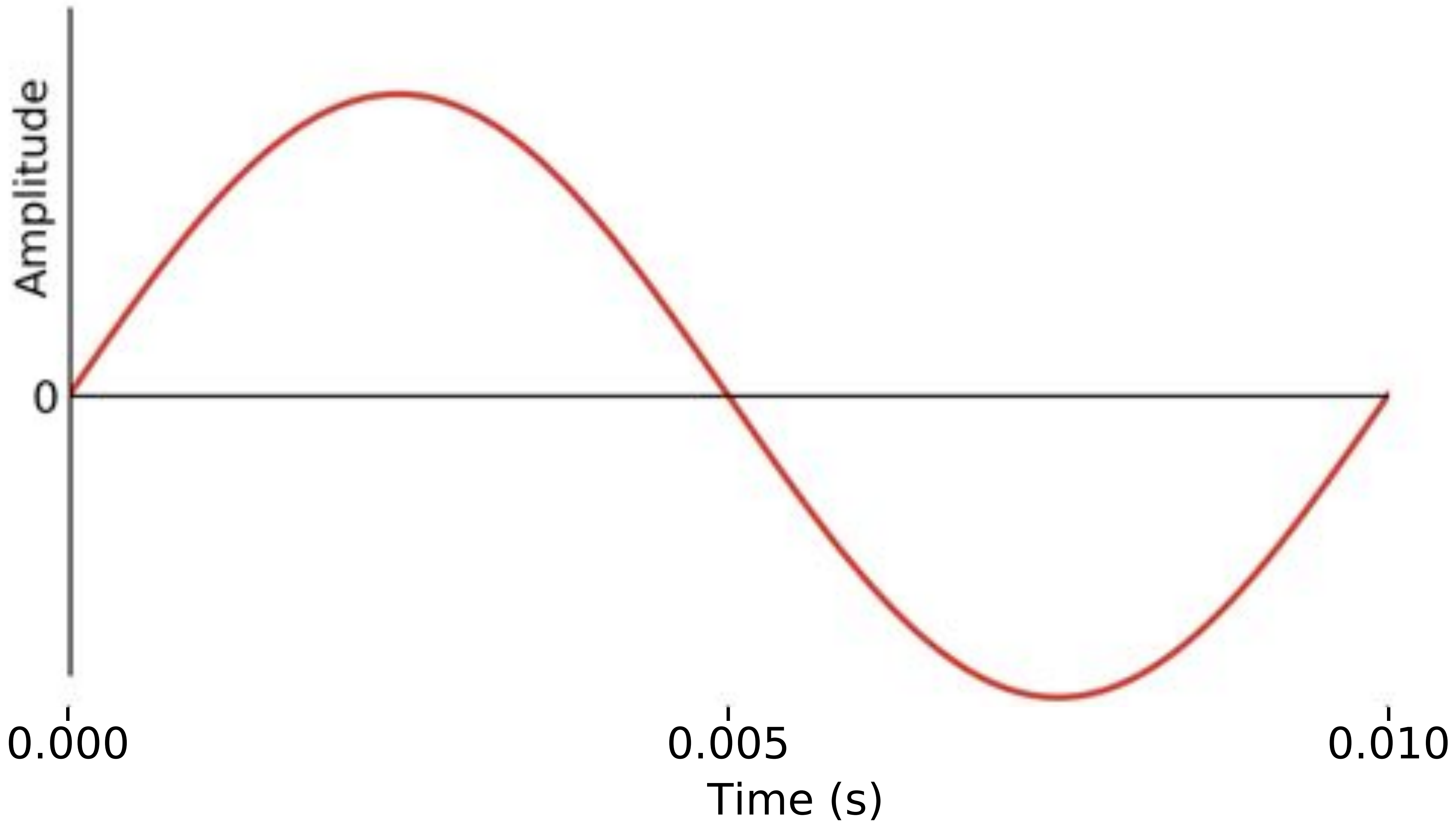


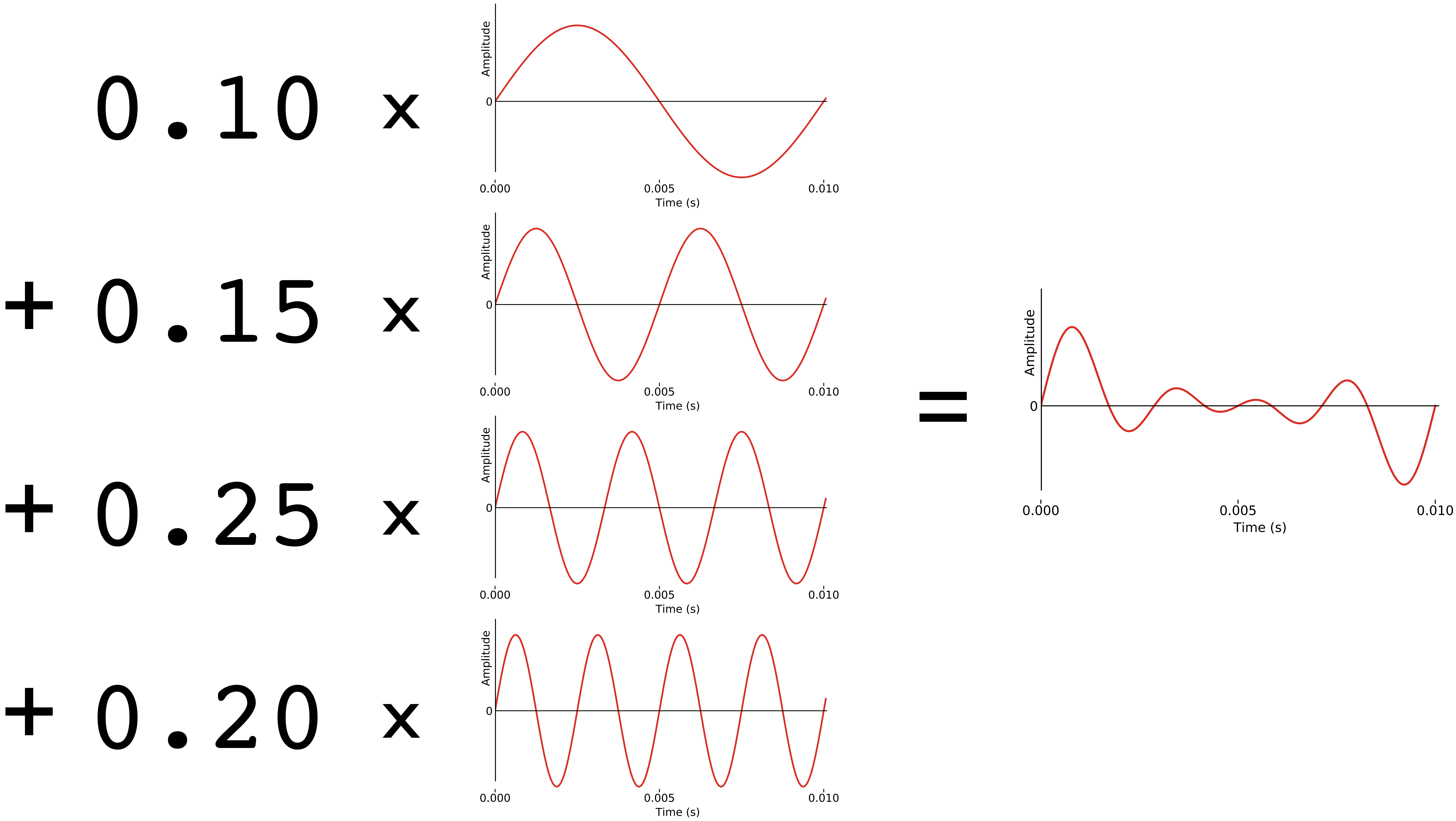

Amplitude
0
0.000
0.005
0.010
Time (s)
=
?.?? x
+ ?.?? x
+ ?.?? x
+ ?.?? x
Amplitude
0
0.000
0.005
0.010
Time (s)

Phase - what it is, and why we usually don't analyse it

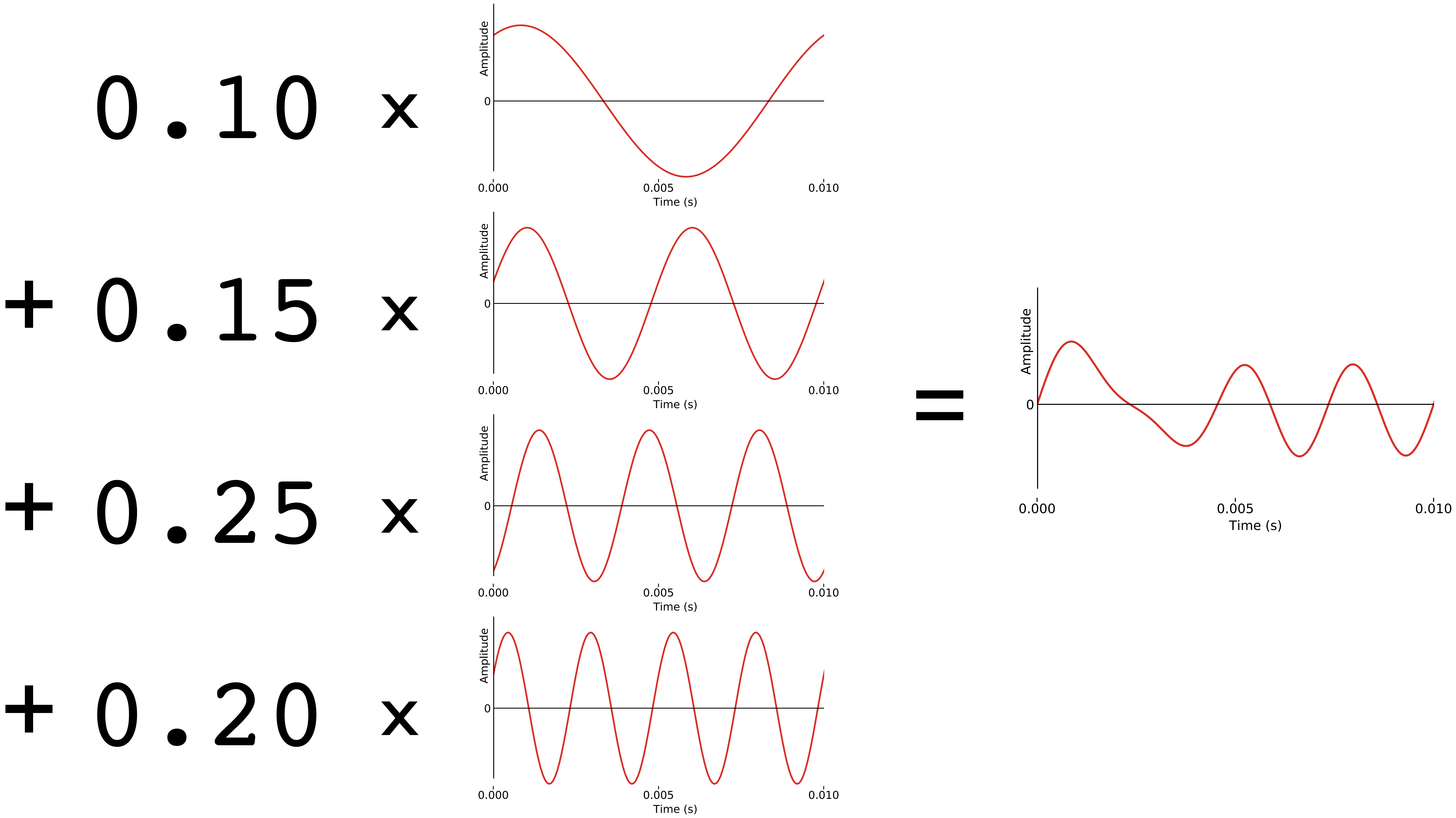


=

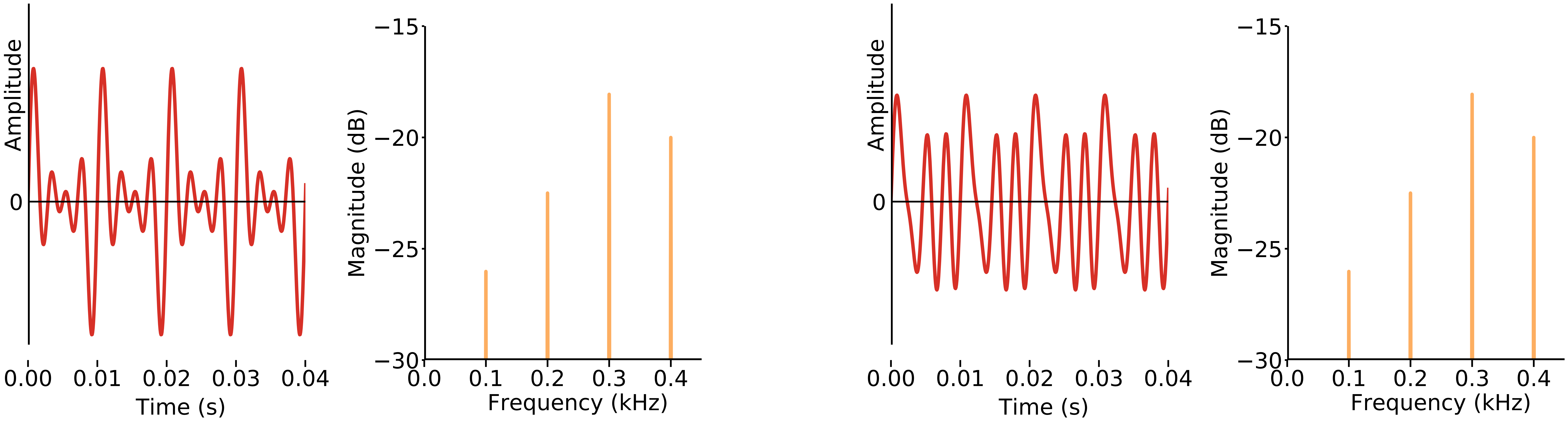




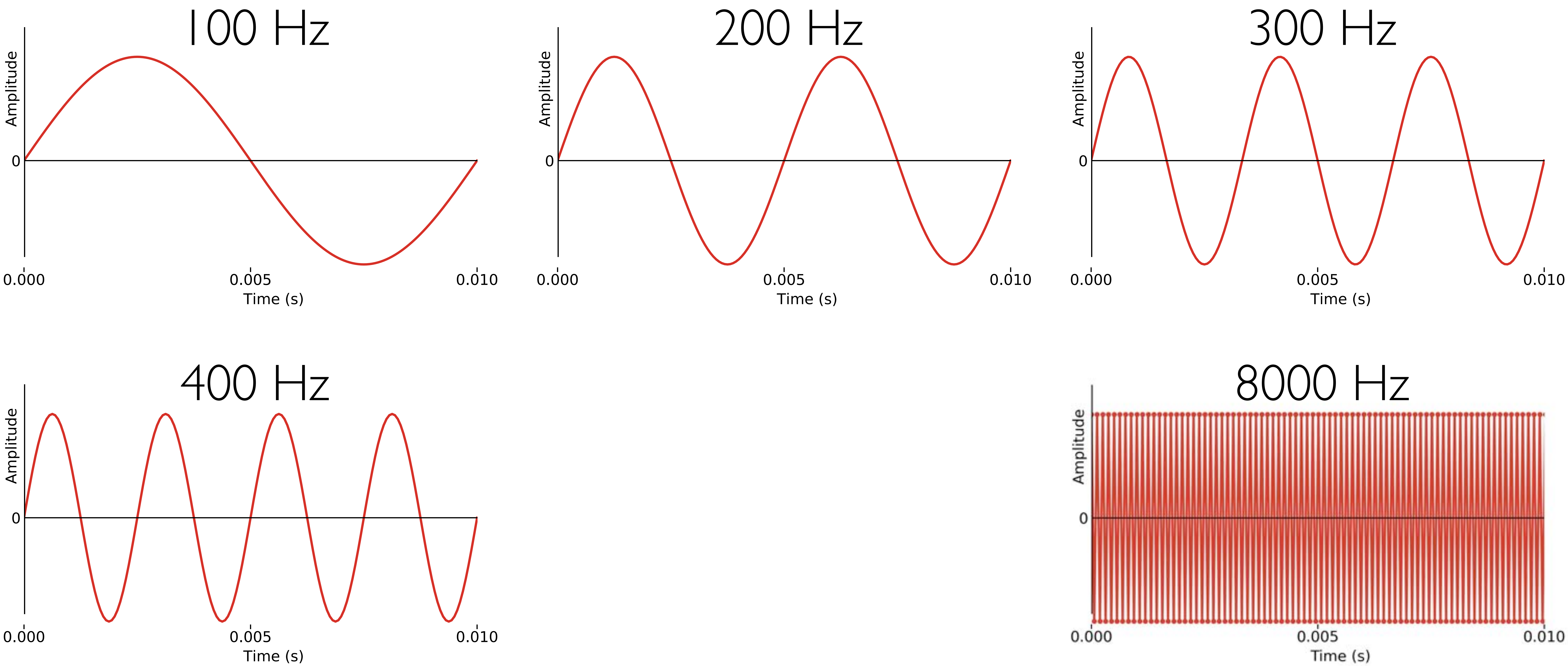
0.10 x
+ 0.15 x
+ 0.25 x
+ 0.20 x
=
Amplitude
0
0.000
0.005
0.010
Time (s)


0.10 x
+ 0.15 x
+ 0.25 x
+ 0.20 x
=
Amplitude
0
0.000
0.005
0.010
Time (s)

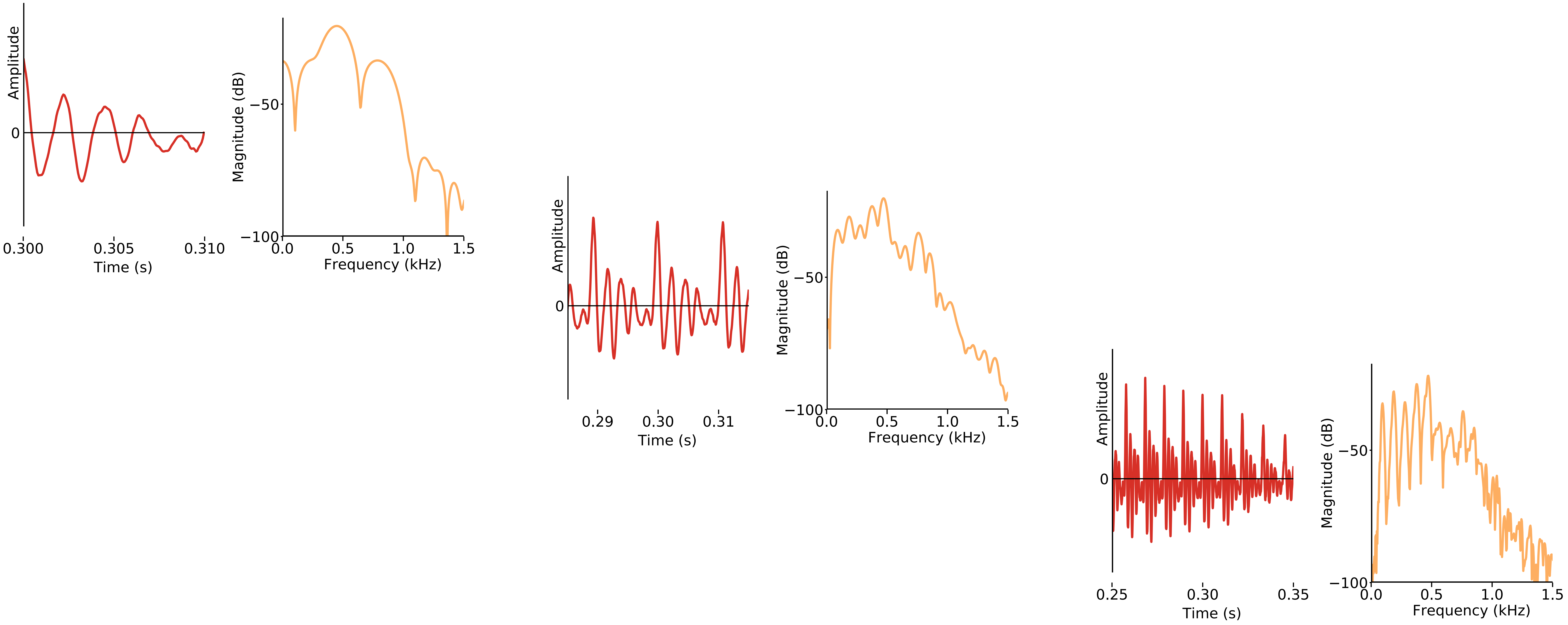
The magnitude spectrum



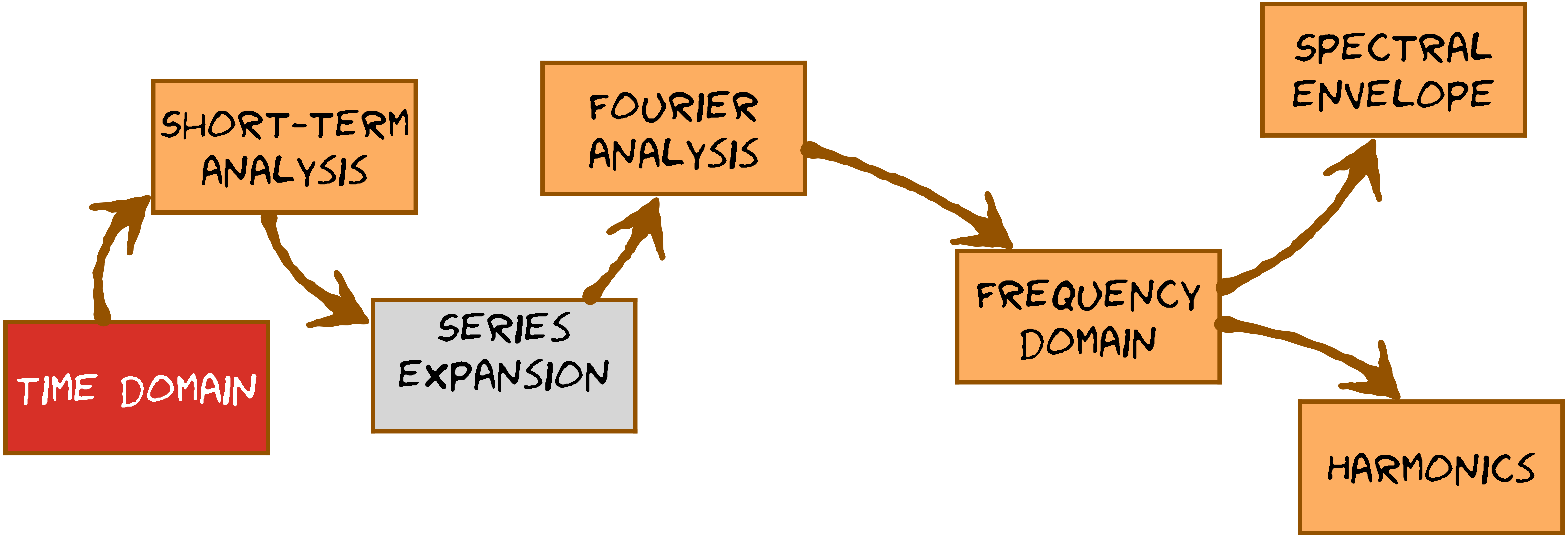
The effect of analysis frame size



Larger analysis frame = more components = higher frequency resolution

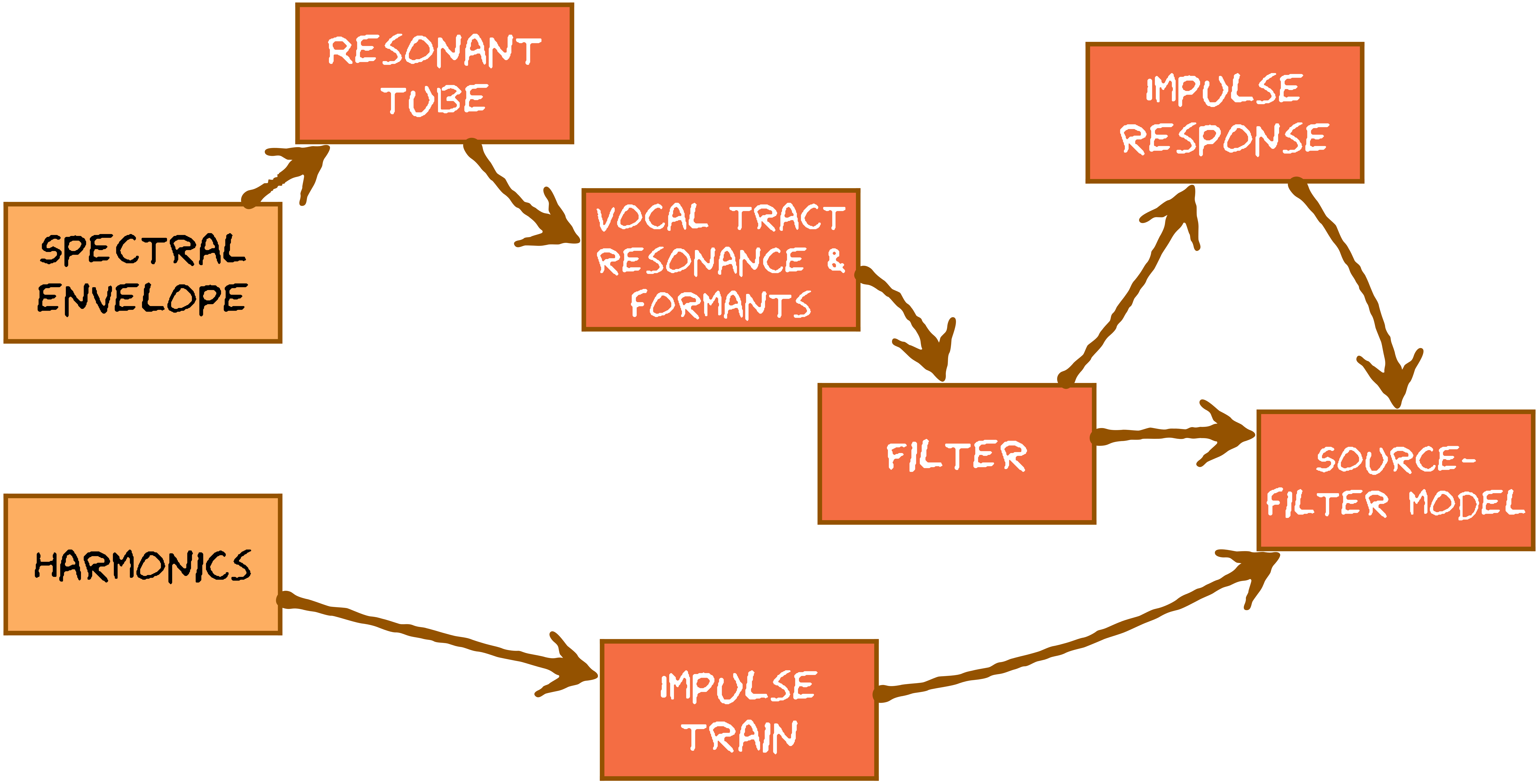


What you can learn next



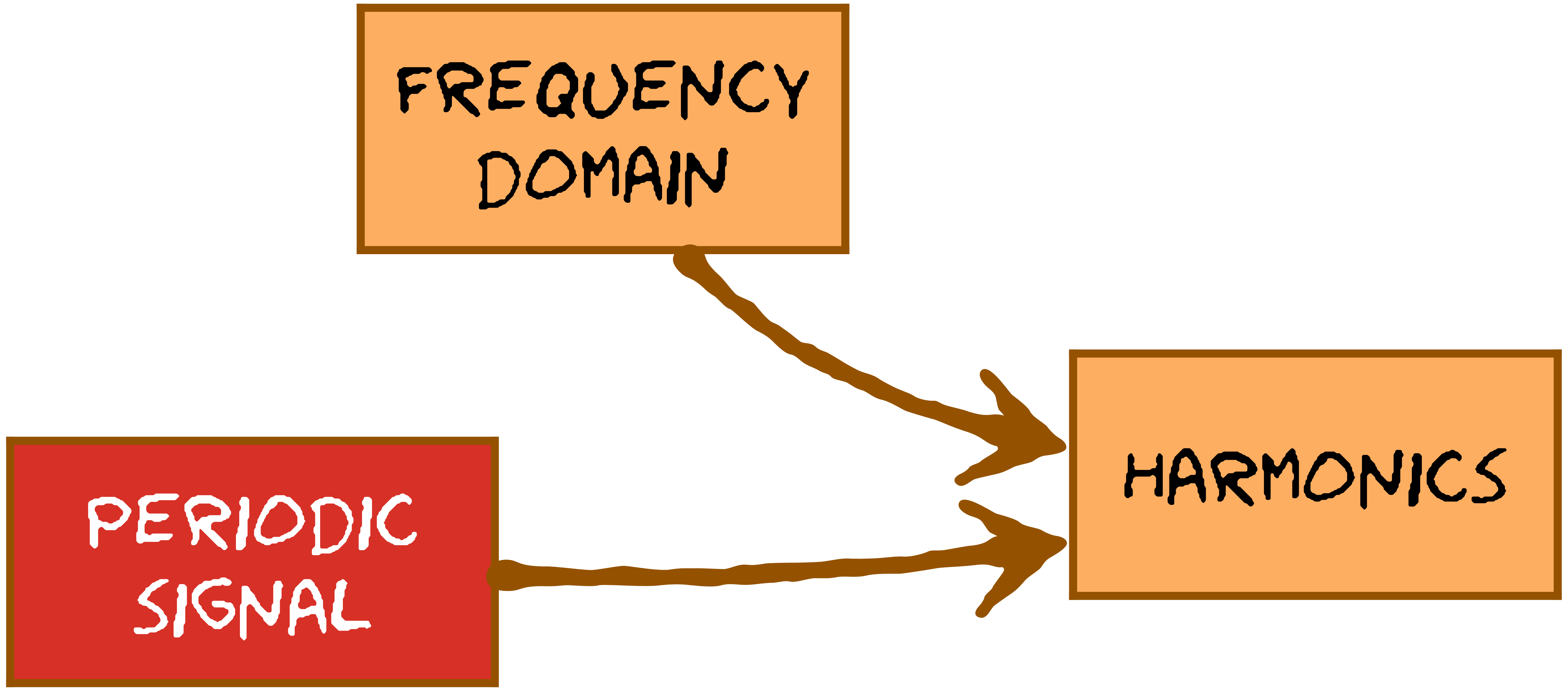
Module 2

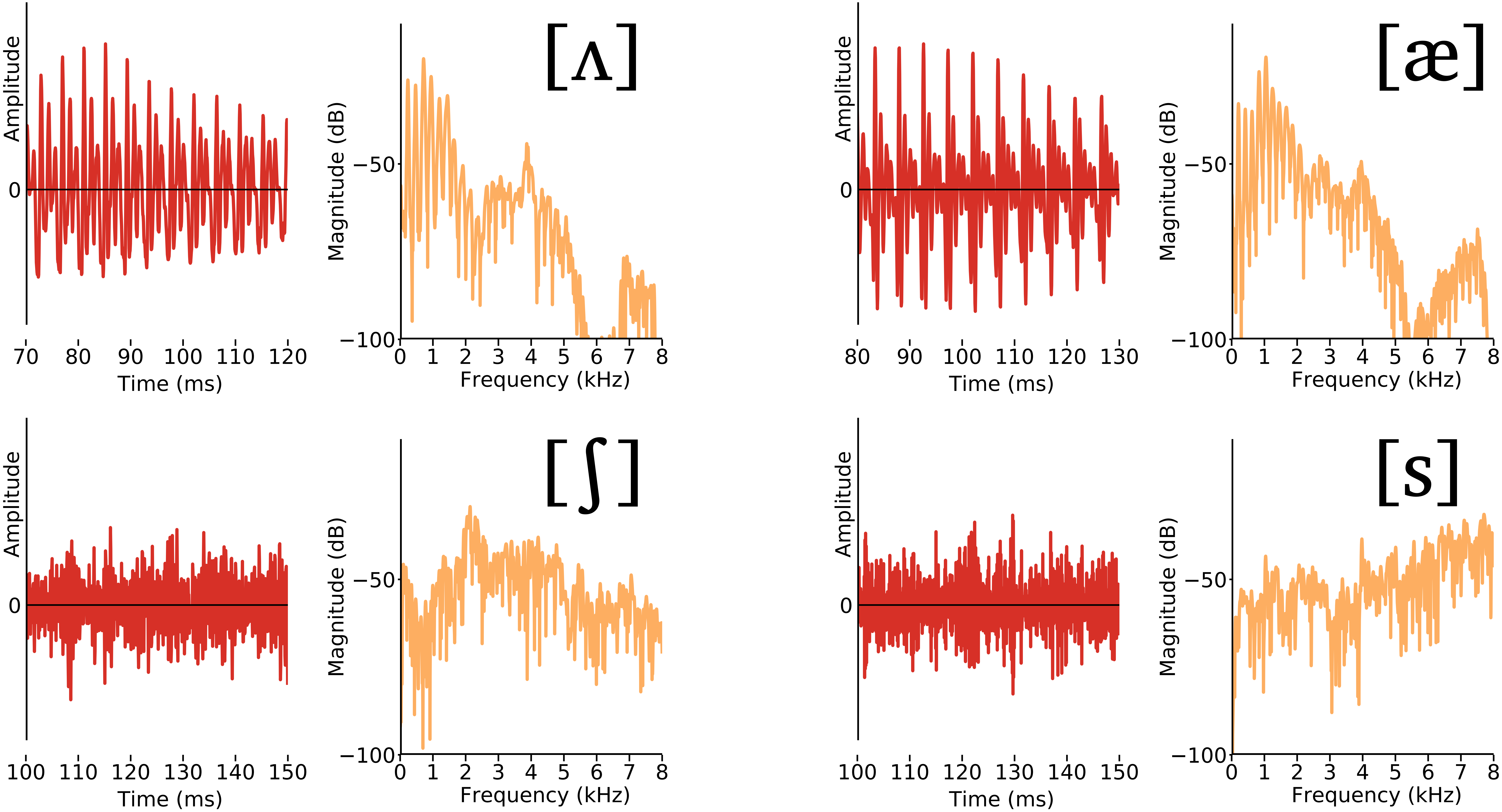
Speech production


SPECTRAL ENVELOPE
RESONANT TUBE
VOCAL TRACT RESONANCE & FORMANTS
FILTER
IMPULSE RESPONSE
SOURCE-FILTER MODEL
HARMONICS
IMPULSE TRAIN

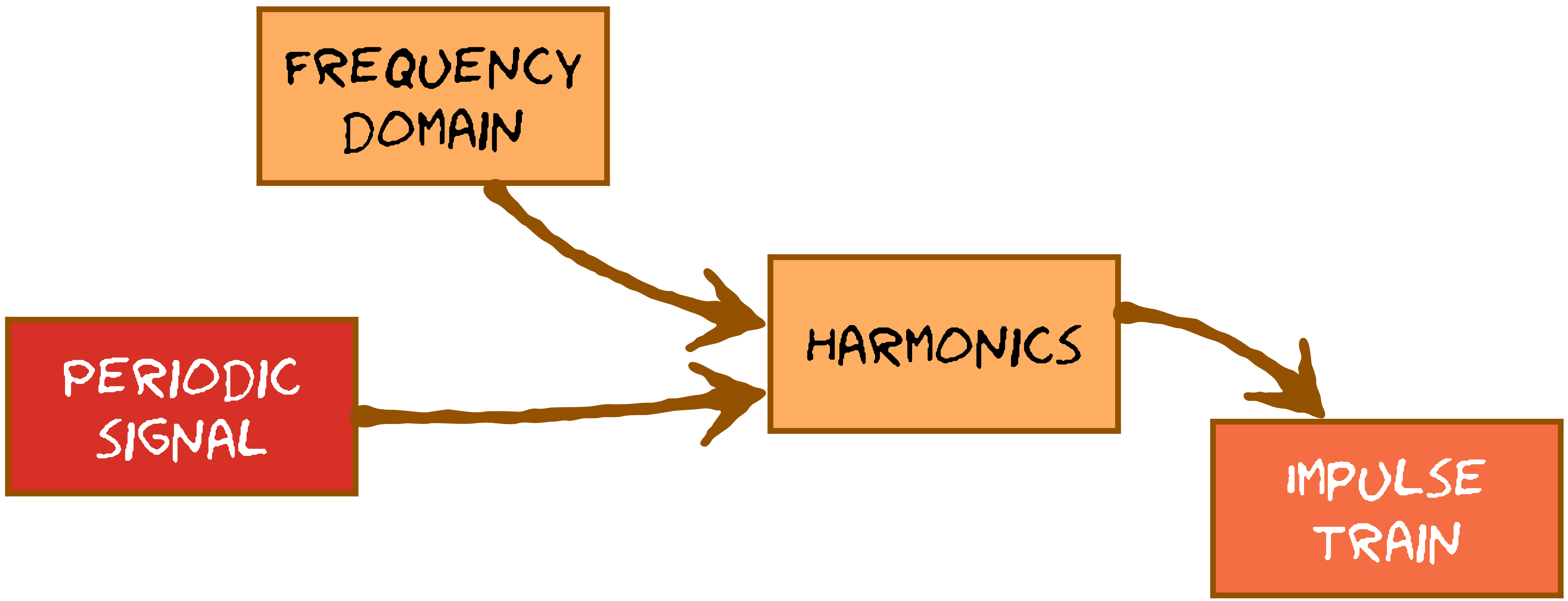
HARMONICS

What you need to know already



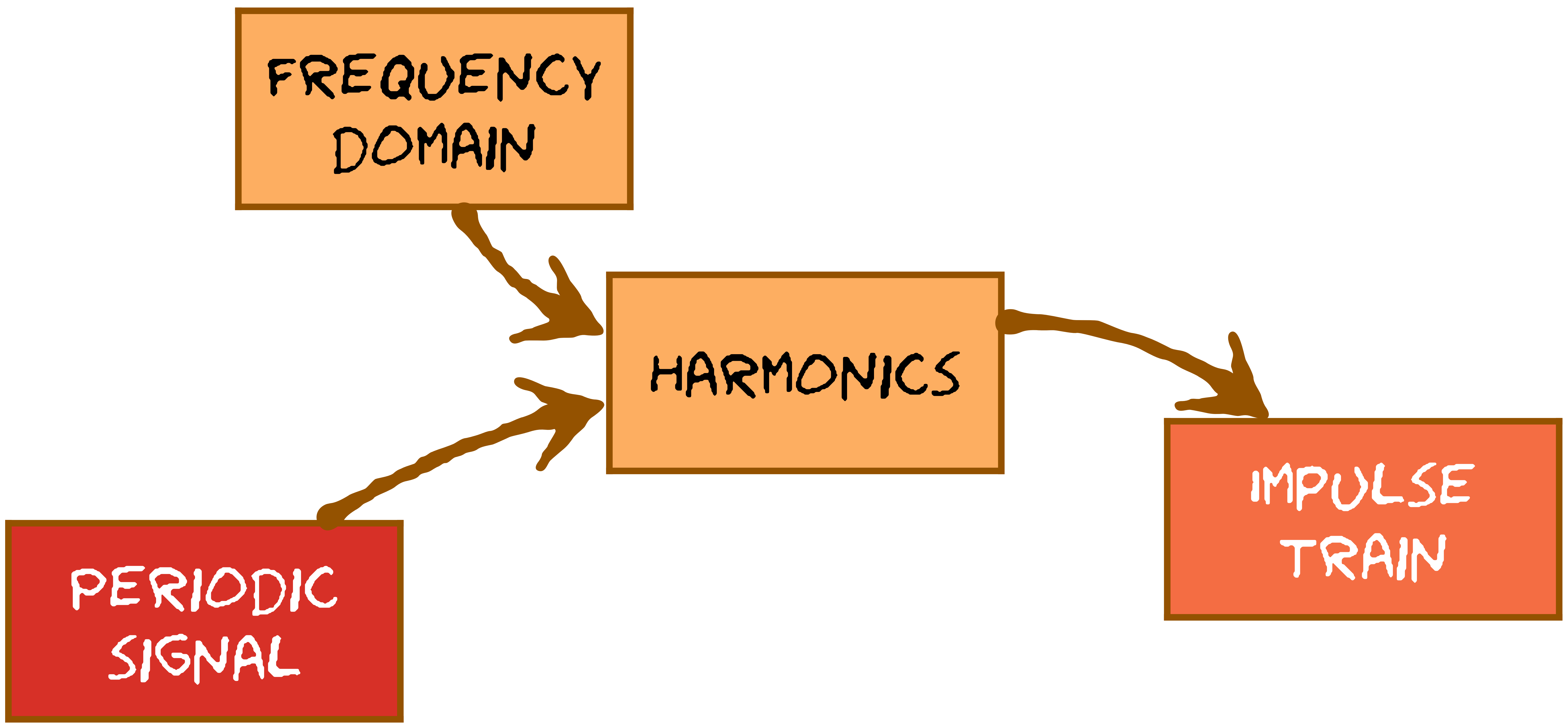

[ʌ]
[æ]
[ʃ]
[s]
Amplitude
0
Time (ms)
70 80 90 100 110 120
80 90 100 110 120 130
100 110 120 130 140 150
Magnitude (dB)
−50
−100
Frequency (kHz)
0 1 2 3 4 5 6 7 8

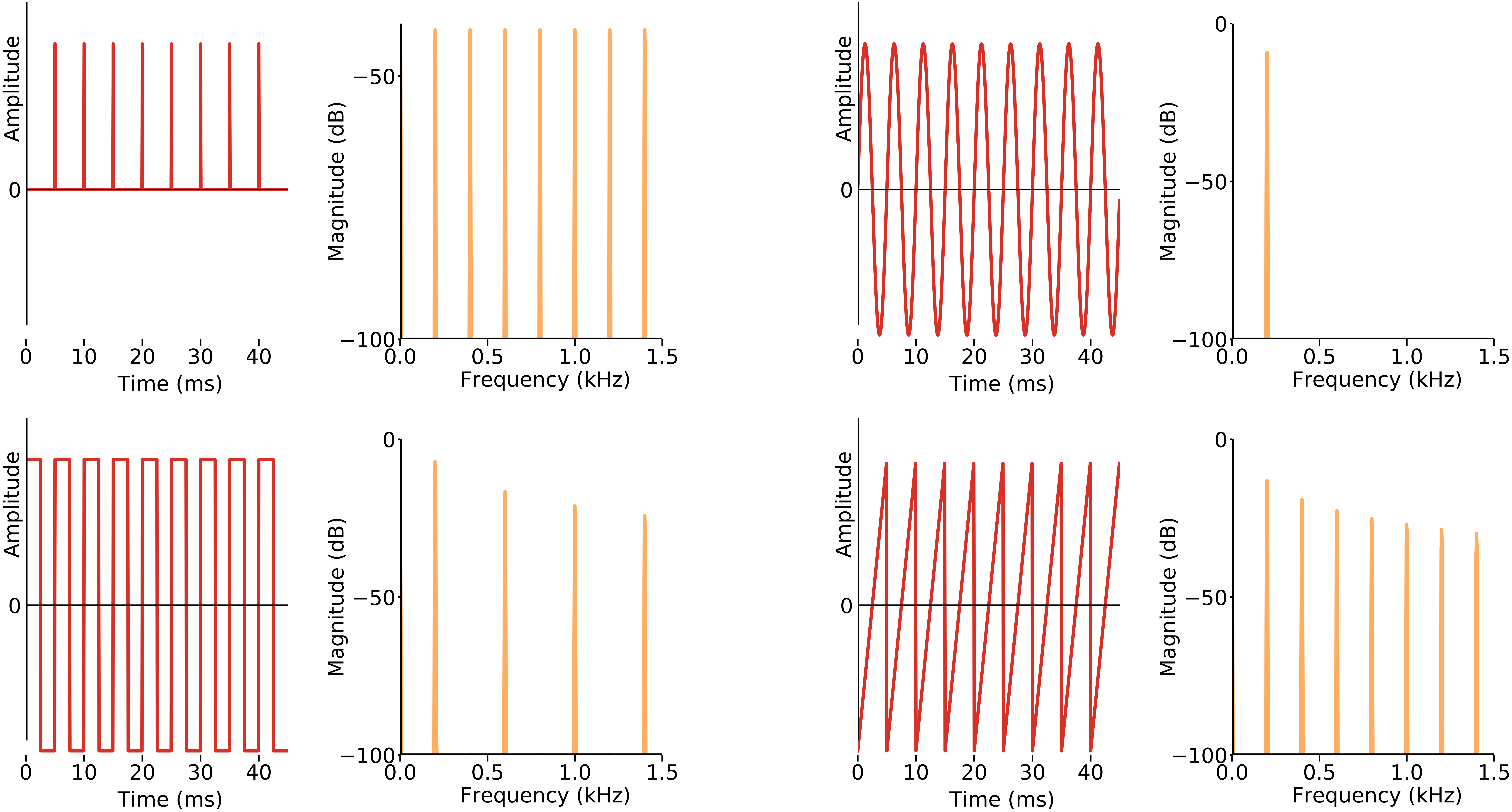
What you can learn next



IMPULSE TRAIN

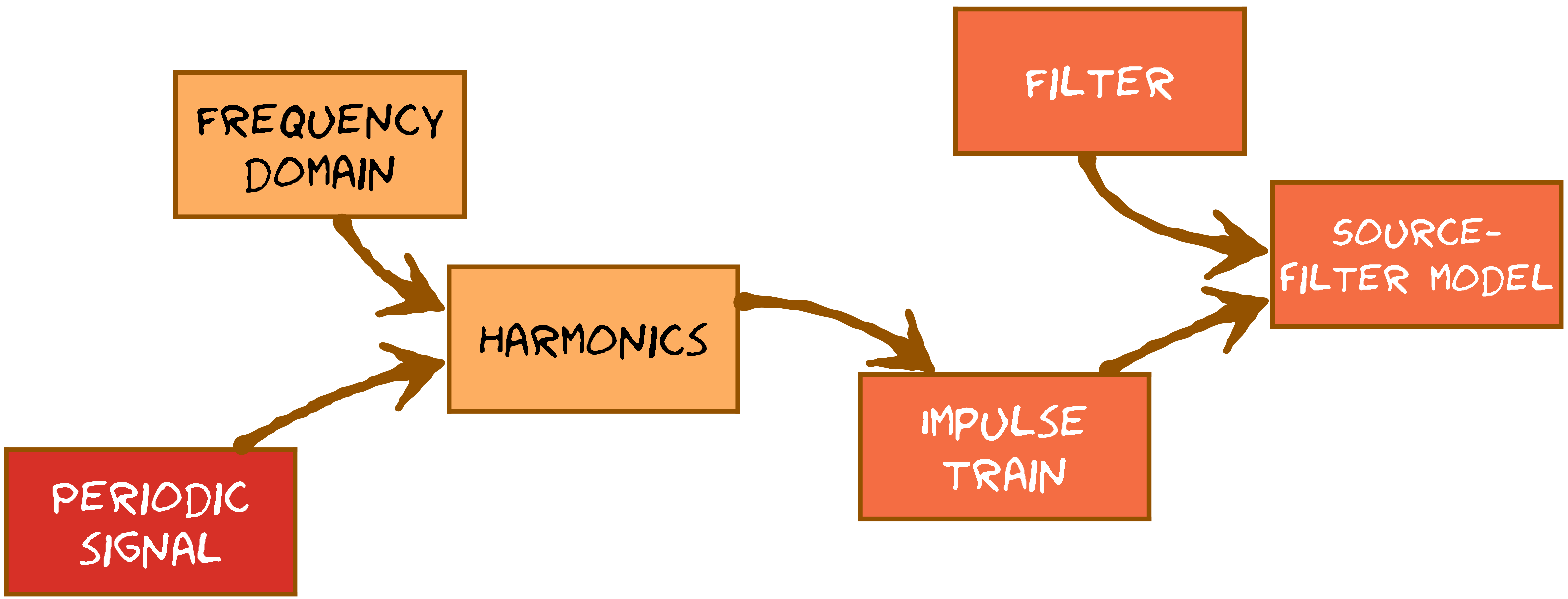
What you need to know already





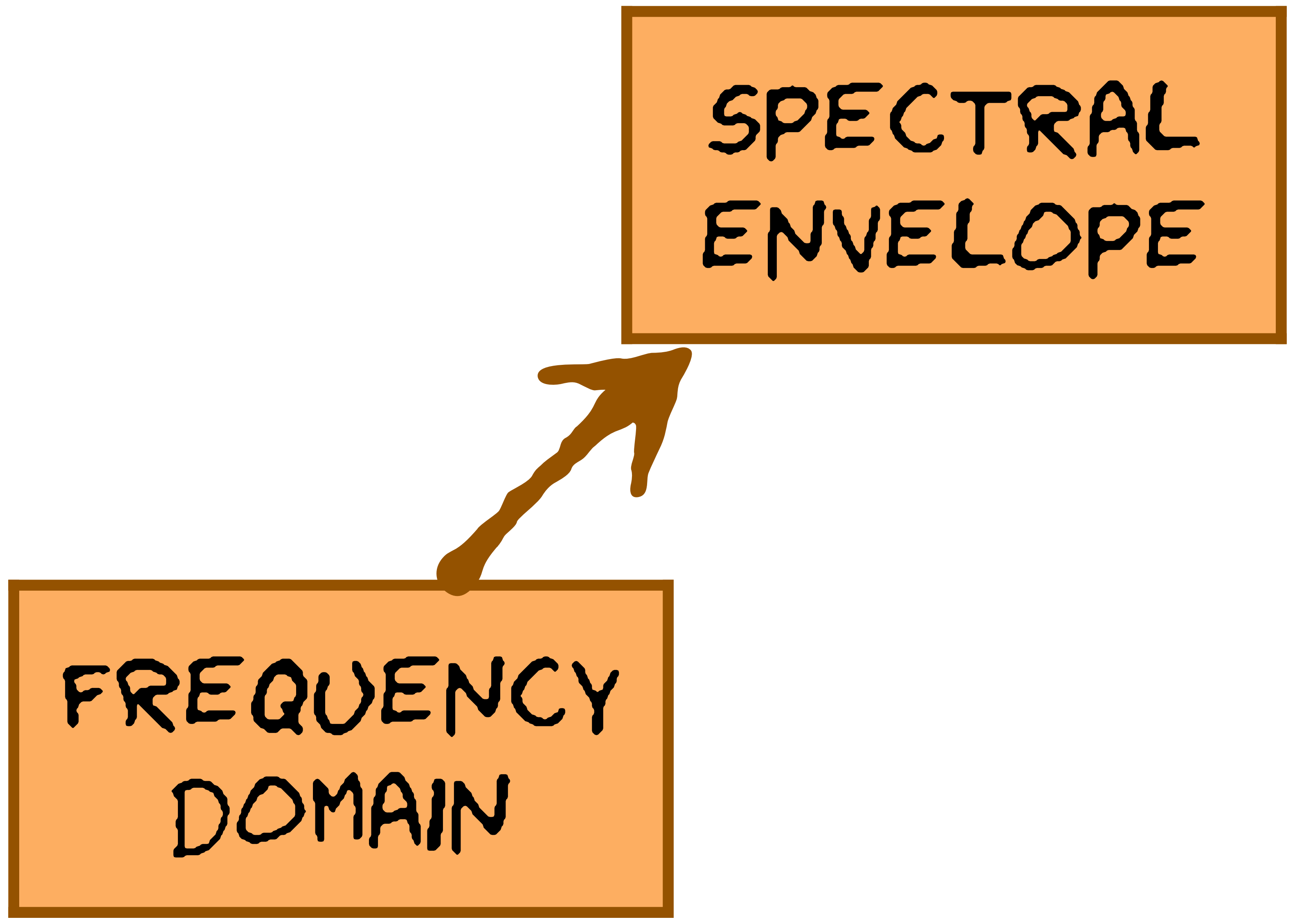
Amplitude
0
0 10 20 30 40
Time (ms)
Magnitude (dB)
−50
−100
0.0 0.5 1.0 1.5
Frequency (kHz)
Amplitude
0
0 10 20 30 40
Time (ms)
Magnitude (dB)
0
−50
−100
0.0 0.5 1.0 1.5
Frequency (kHz)
Amplitude
0
0 10 20 30 40
Time (ms)
Magnitude (dB)
0
−50
−100
0.0 0.5 1.0 1.5
Frequency (kHz)
Amplitude
0
0 10 20 30 40
Time (ms)
Magnitude (dB)
0
−50
−100
0.0 0.5 1.0 1.5
Frequency (kHz)

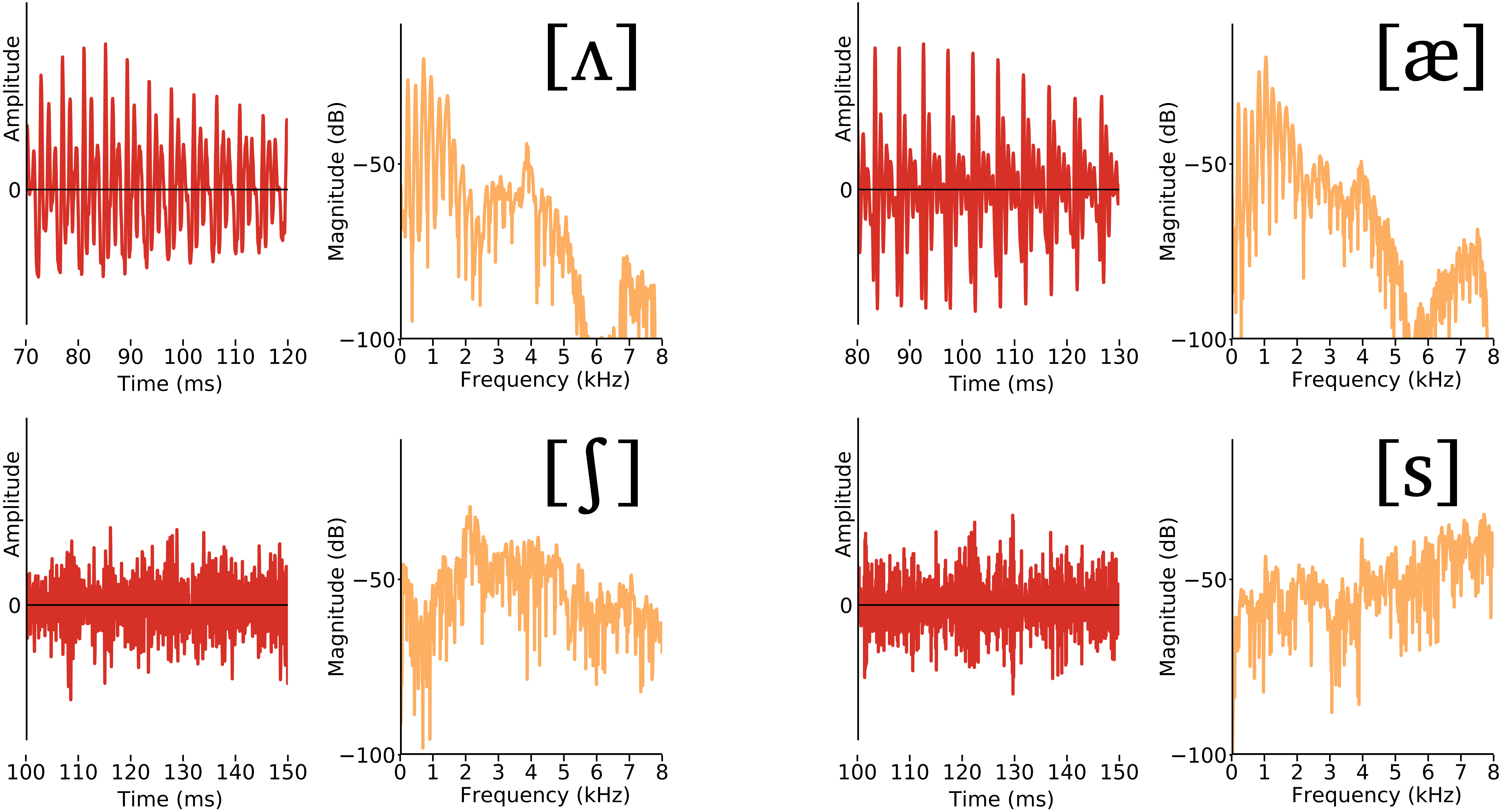
What you can learn next



SPECTRAL ENVELOPE

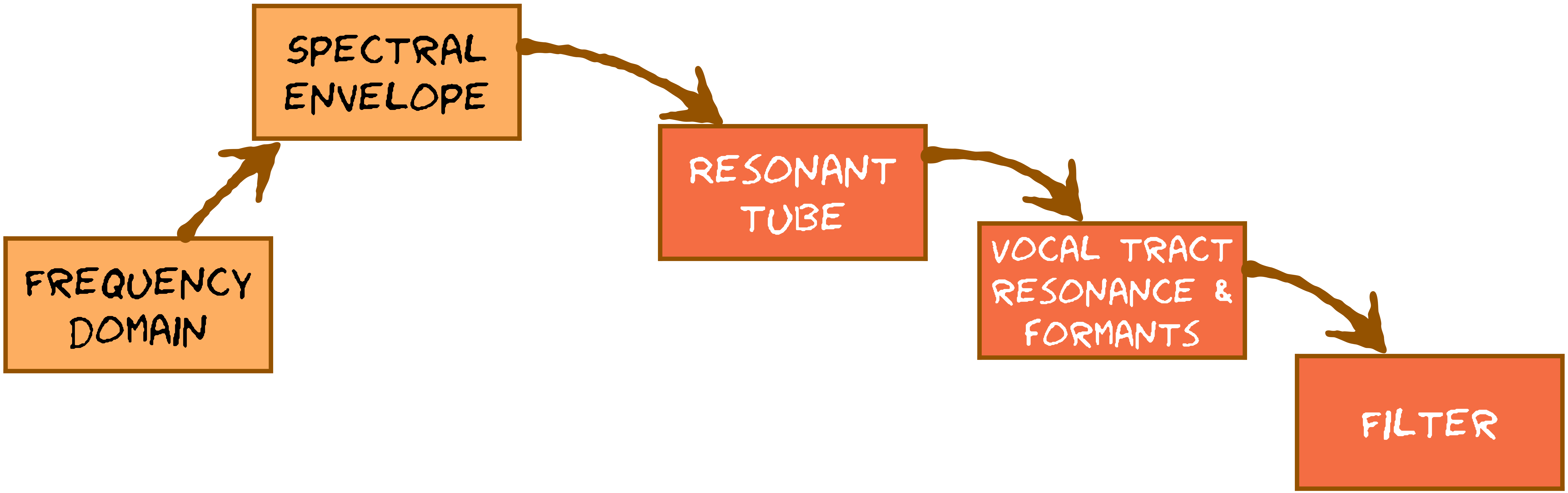
What you need to know already





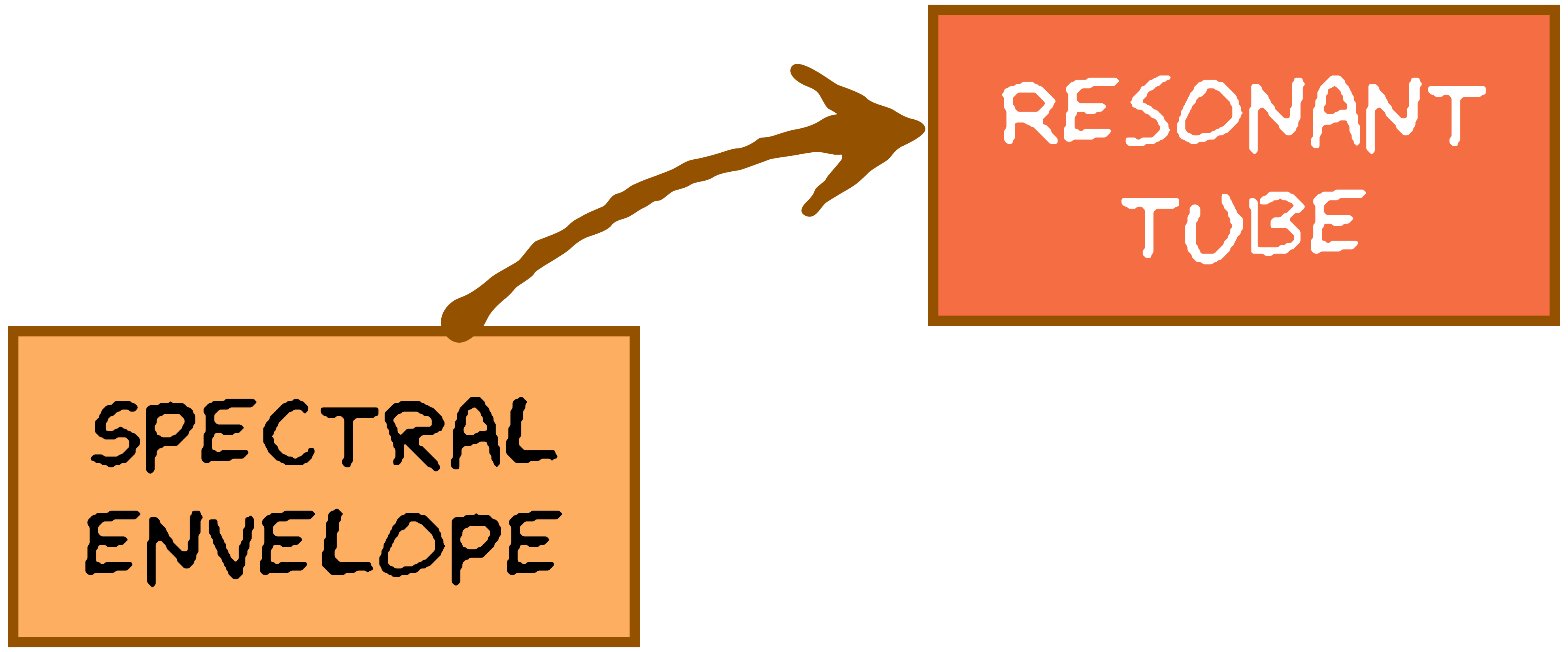
[ʌ]
[æ]
[ʃ]
[s]
Amplitude
0
Time (ms)
70
80
90
100
110
120
130
140
150
Magnitude (dB)
−50
−100
Frequency (kHz)
0
1
2
3
4
5
6
7
8

What you can learn next

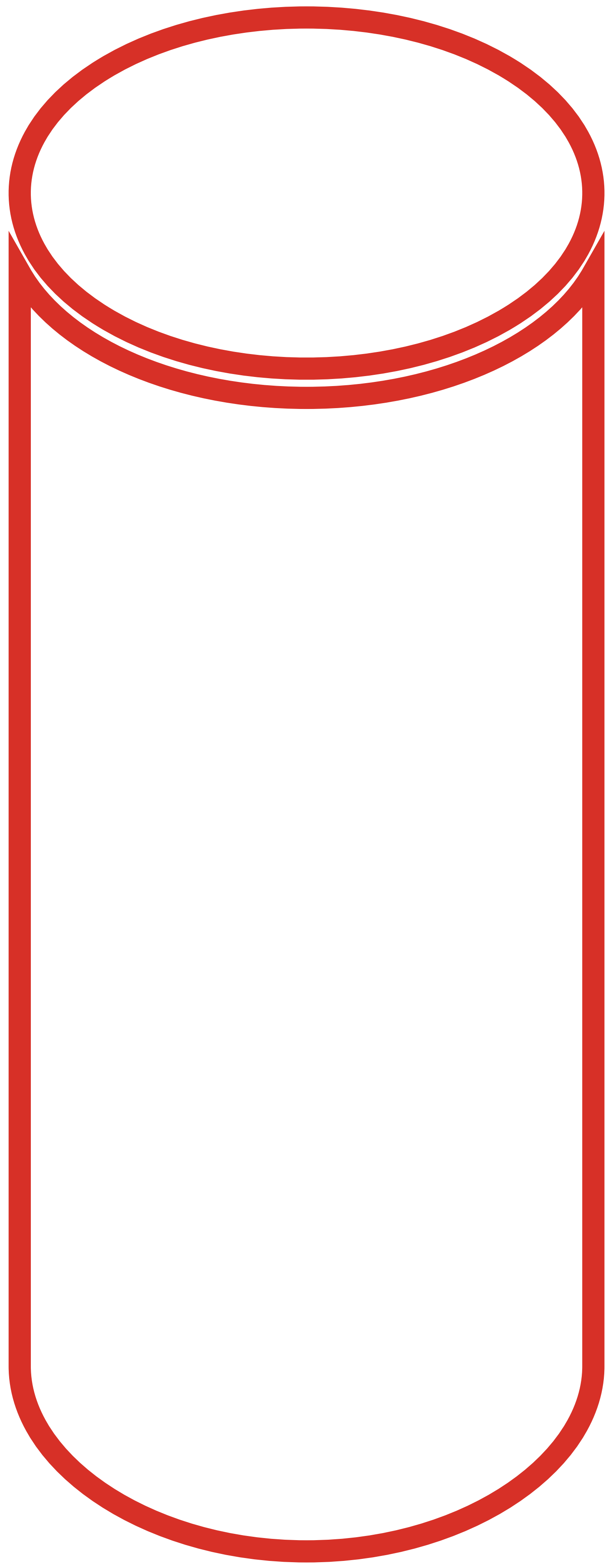


RESONANT TUBE

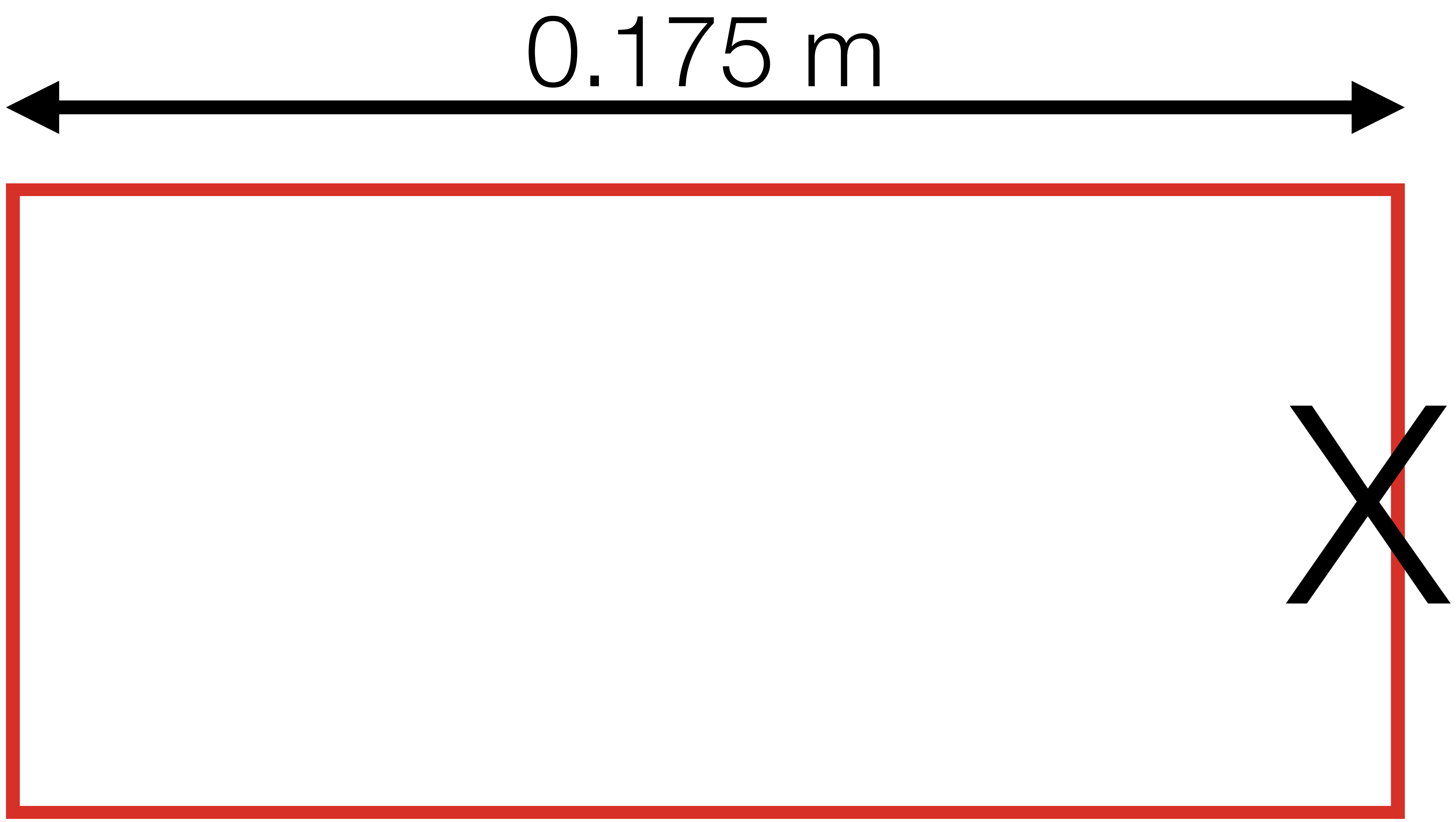
What you need to know already



The vocal tract is tube

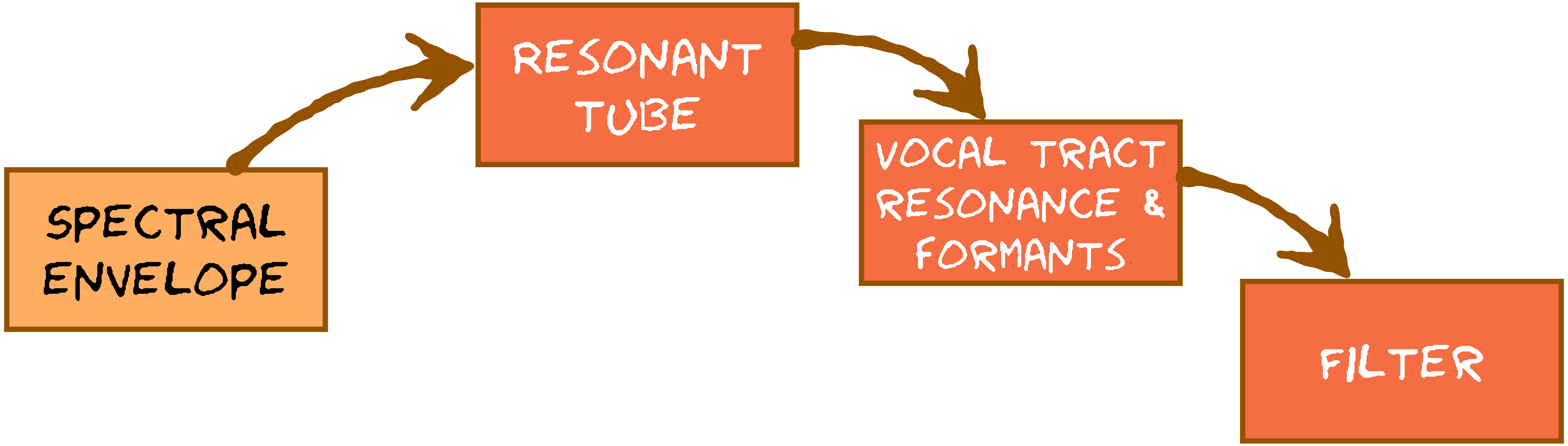


A tube is a resonator



speed of sound is 350 ms^{-1}

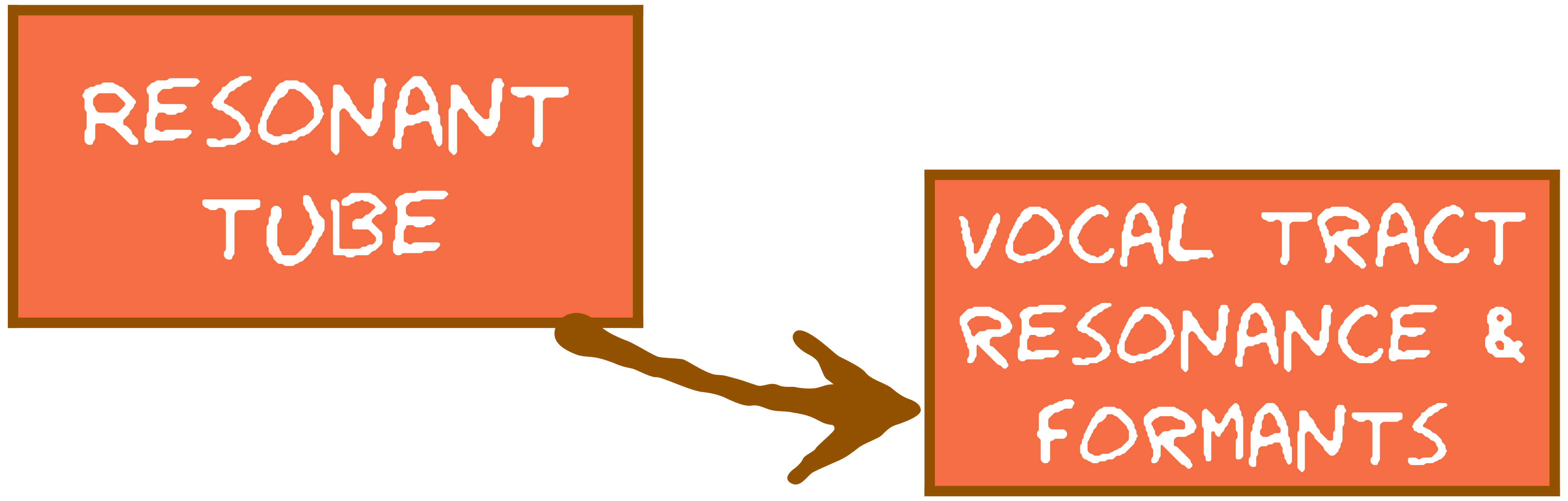
What you can learn next



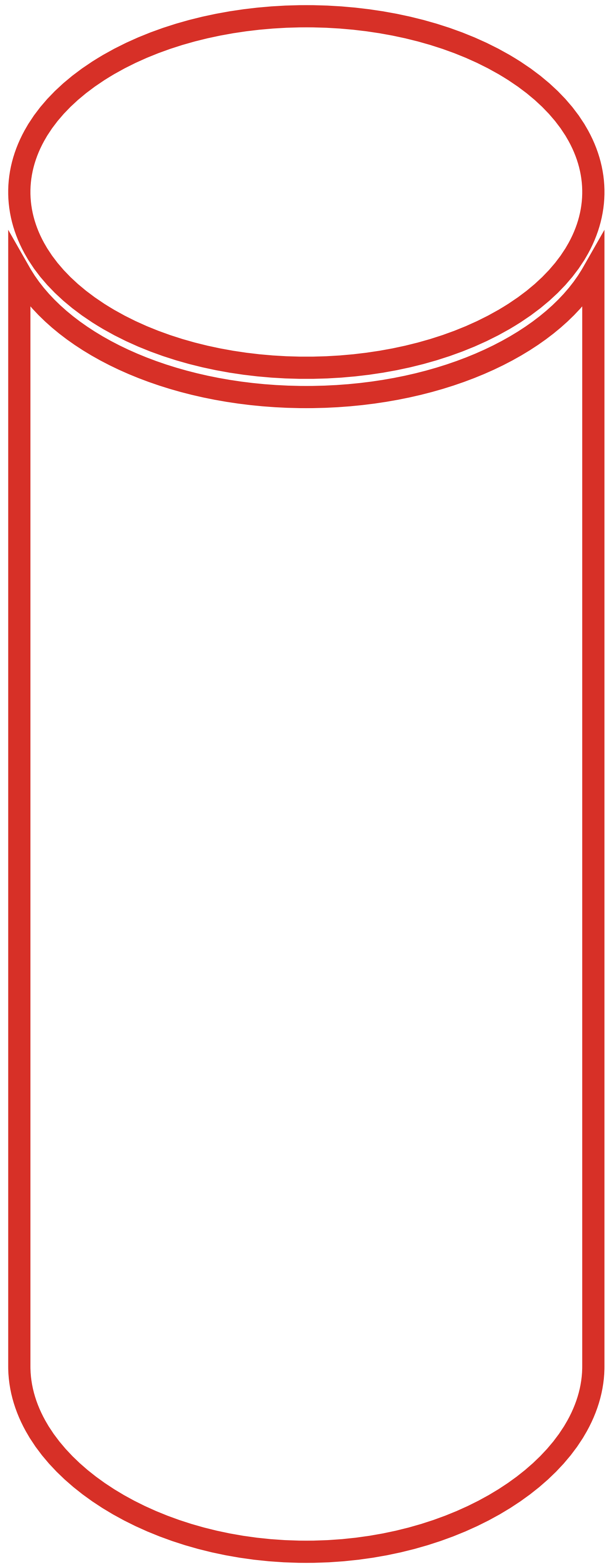
VOCAL TRACT RESONANCE & FORMANTS

THE VOCAL TRACT IS A FILTER

What you need to know already

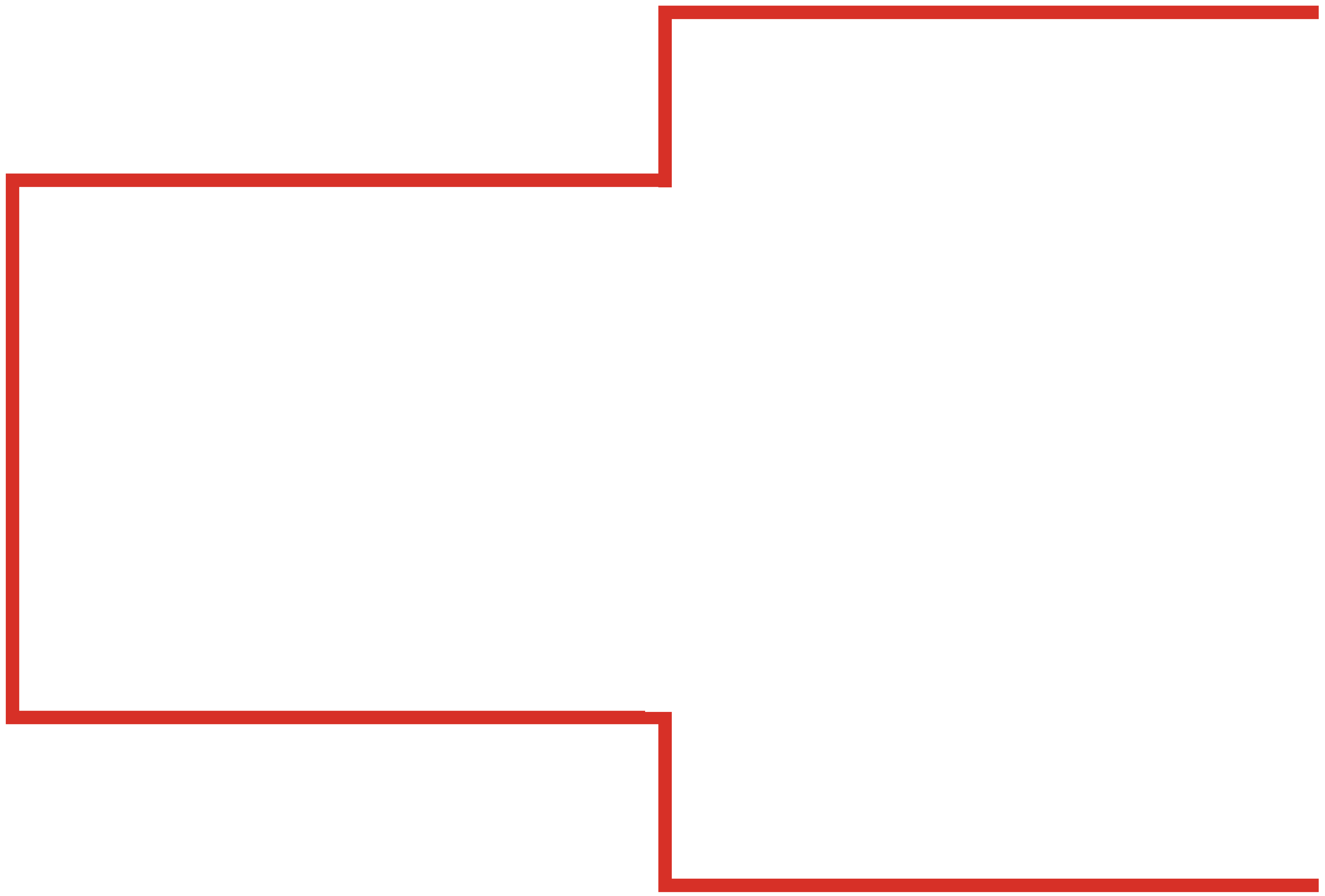


Multiple resonant frequencies

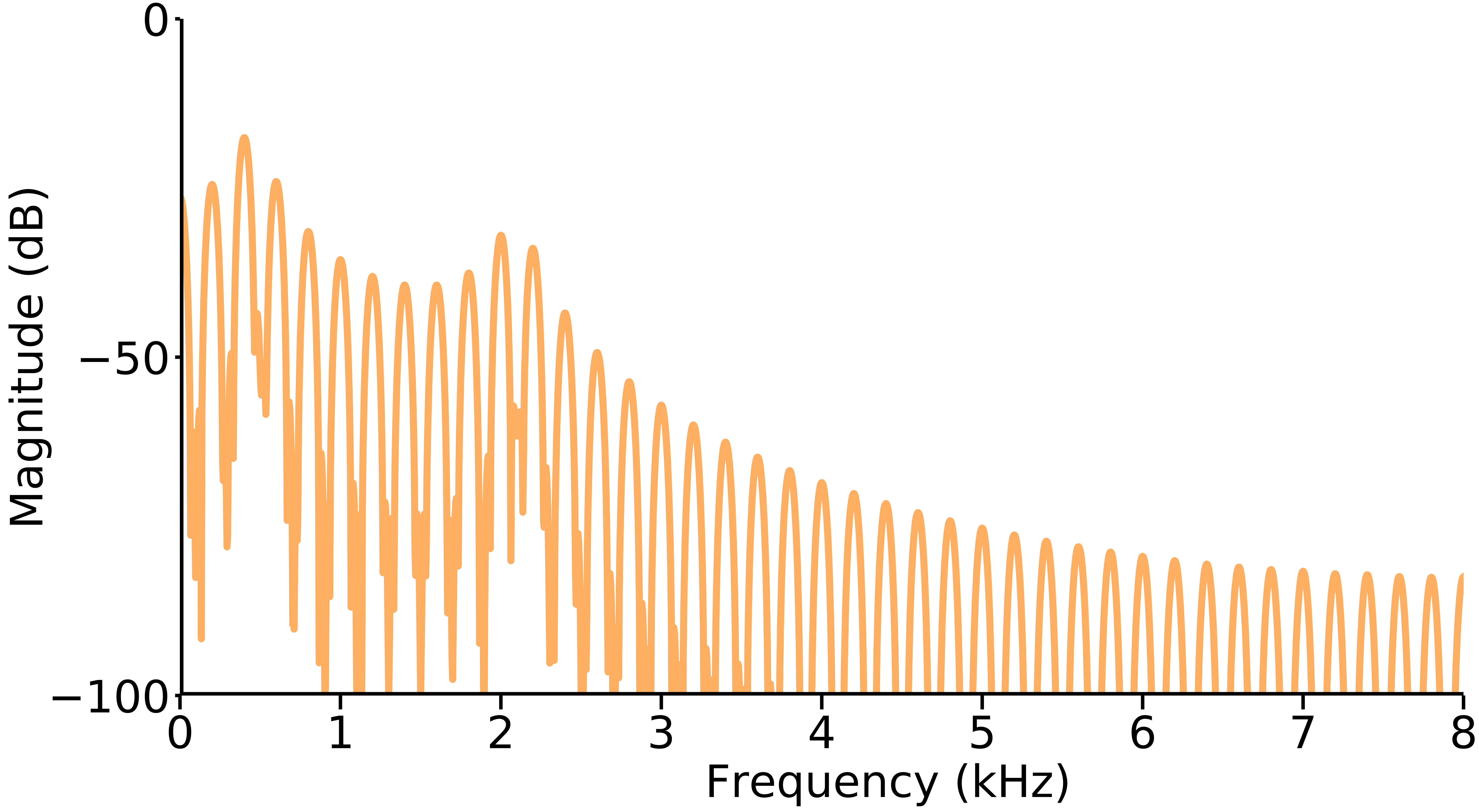


Simplify our model into one dimension: only the length matters

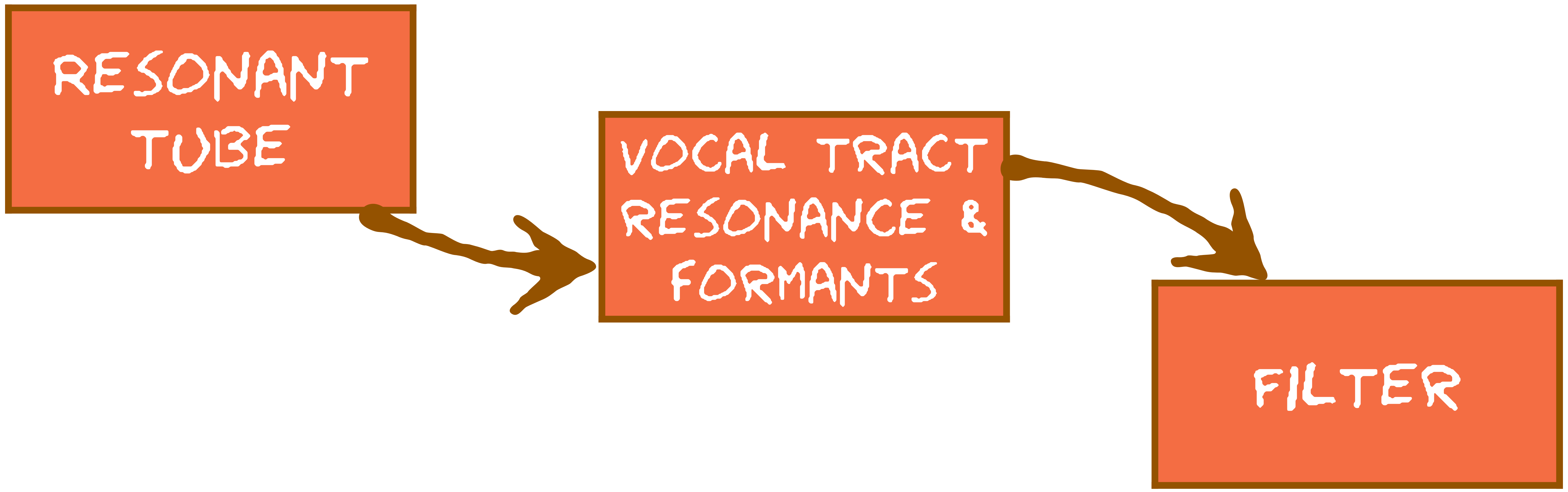
The tube can vary in shape



Formants are the resonant frequencies of the vocal tract



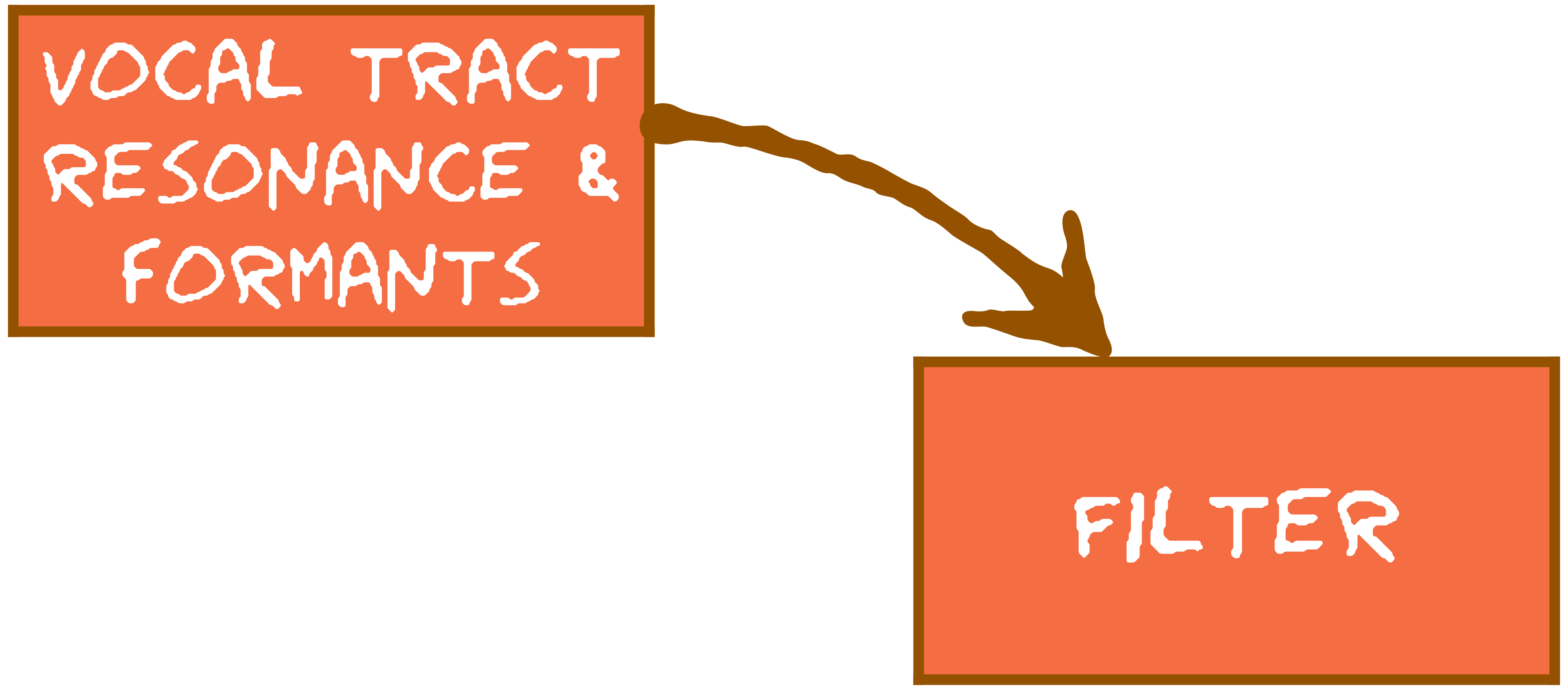
What you can learn next



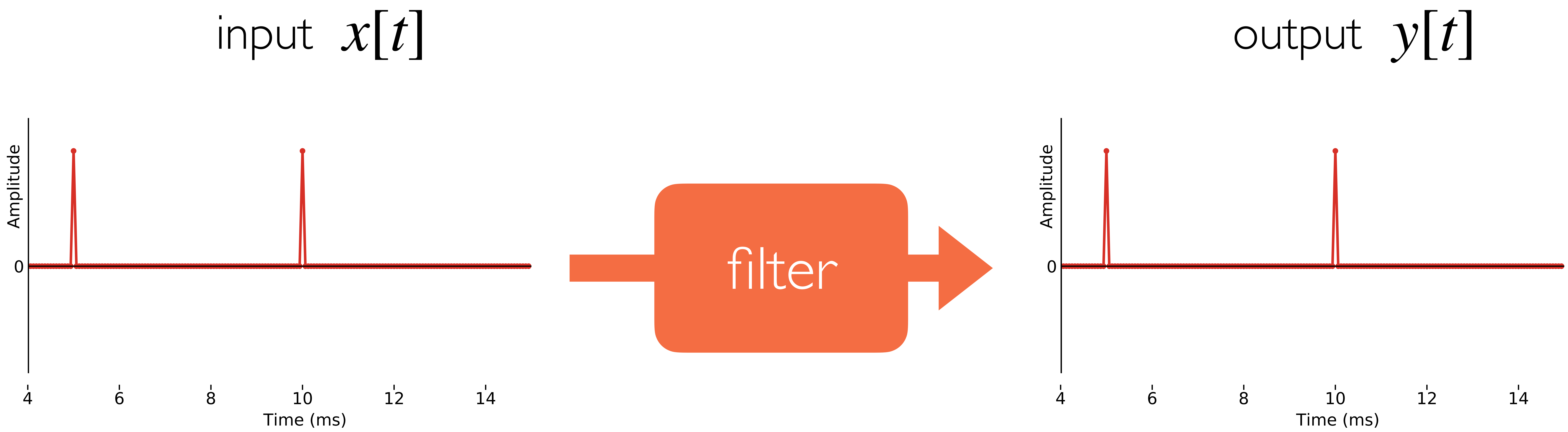
FILTER

THE VOCAL TRACT IS A FILTER

What you need to know already



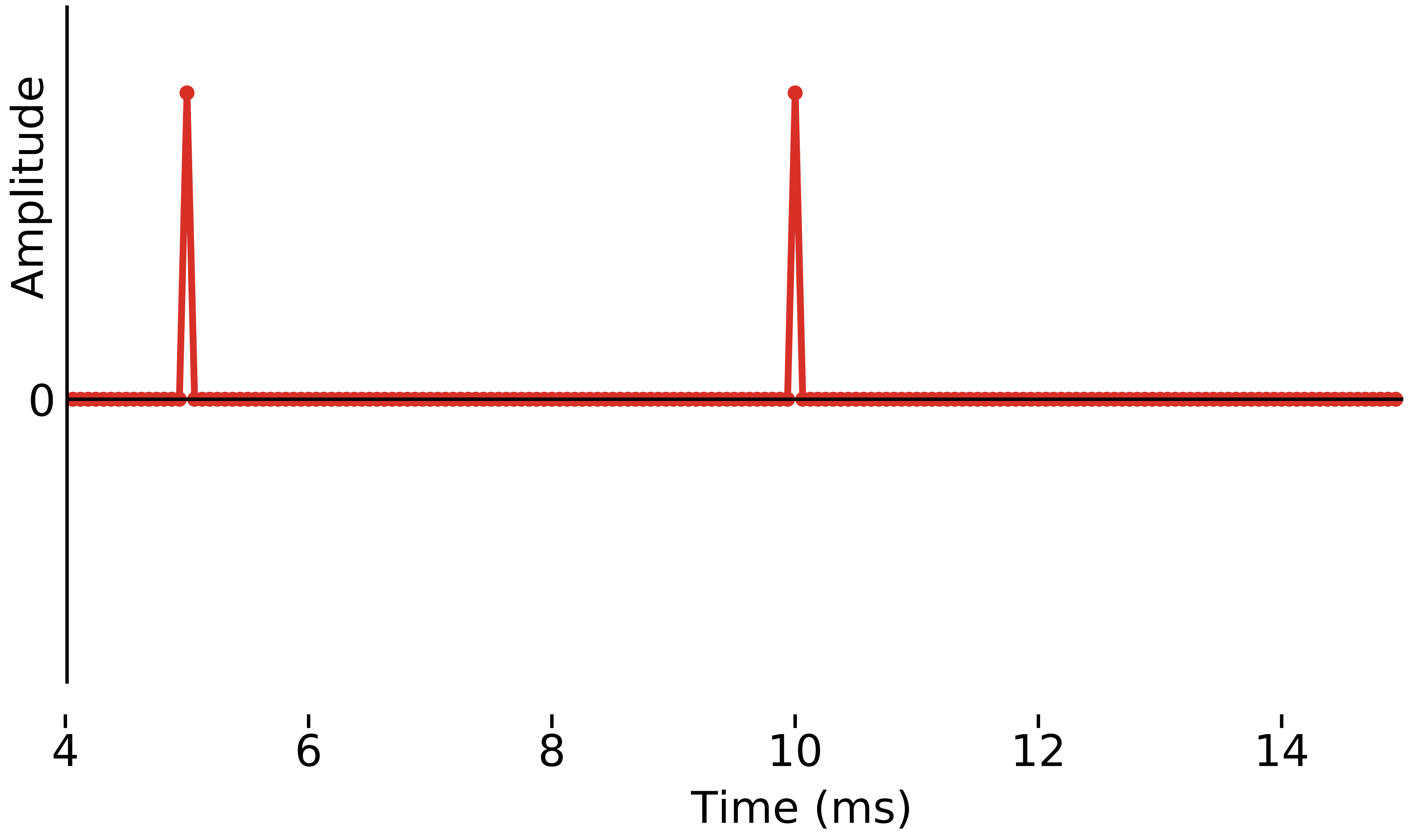
Defining a filter



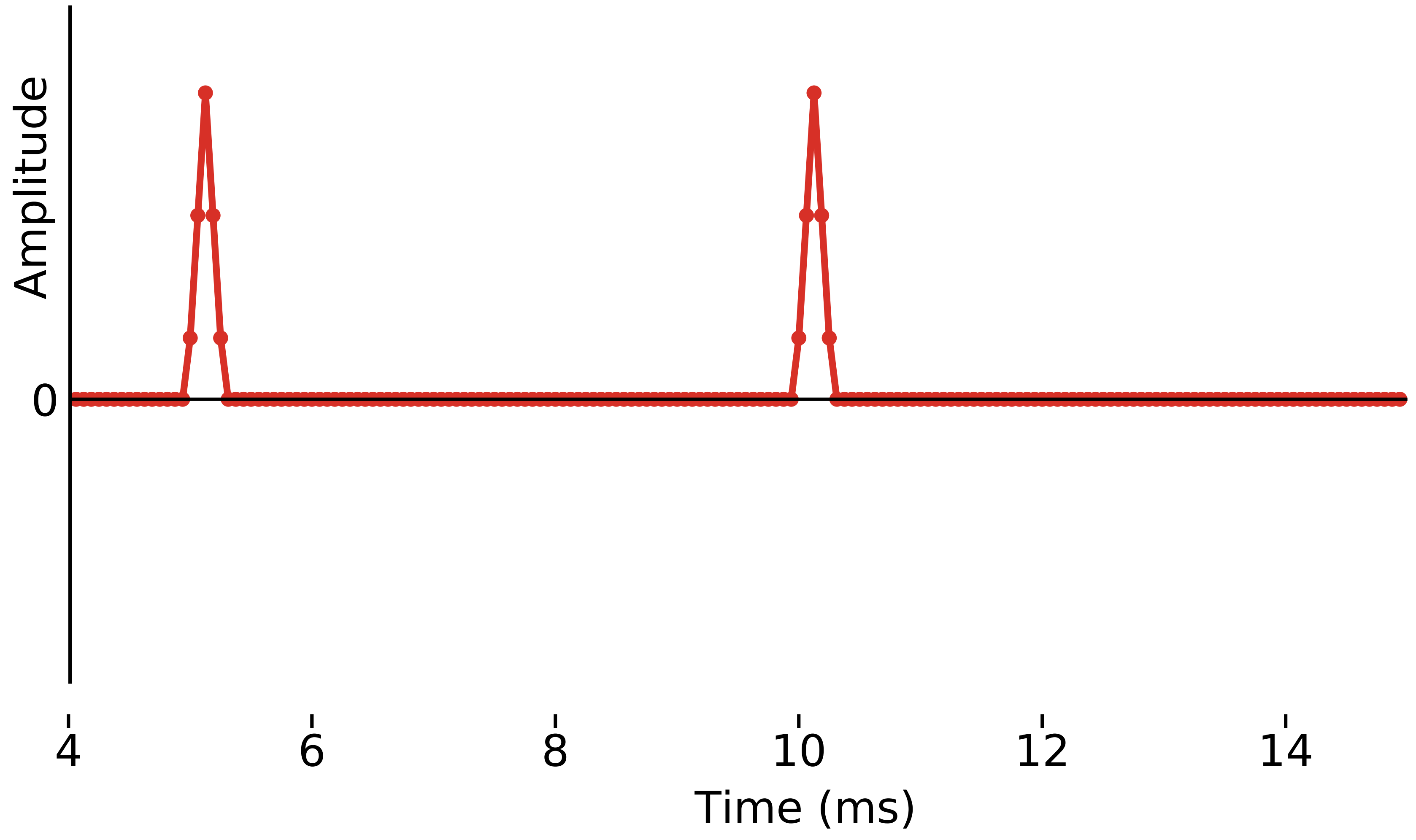
$$y[t] = 1.0x[t]$$

Filtering in action

input $x[t]$

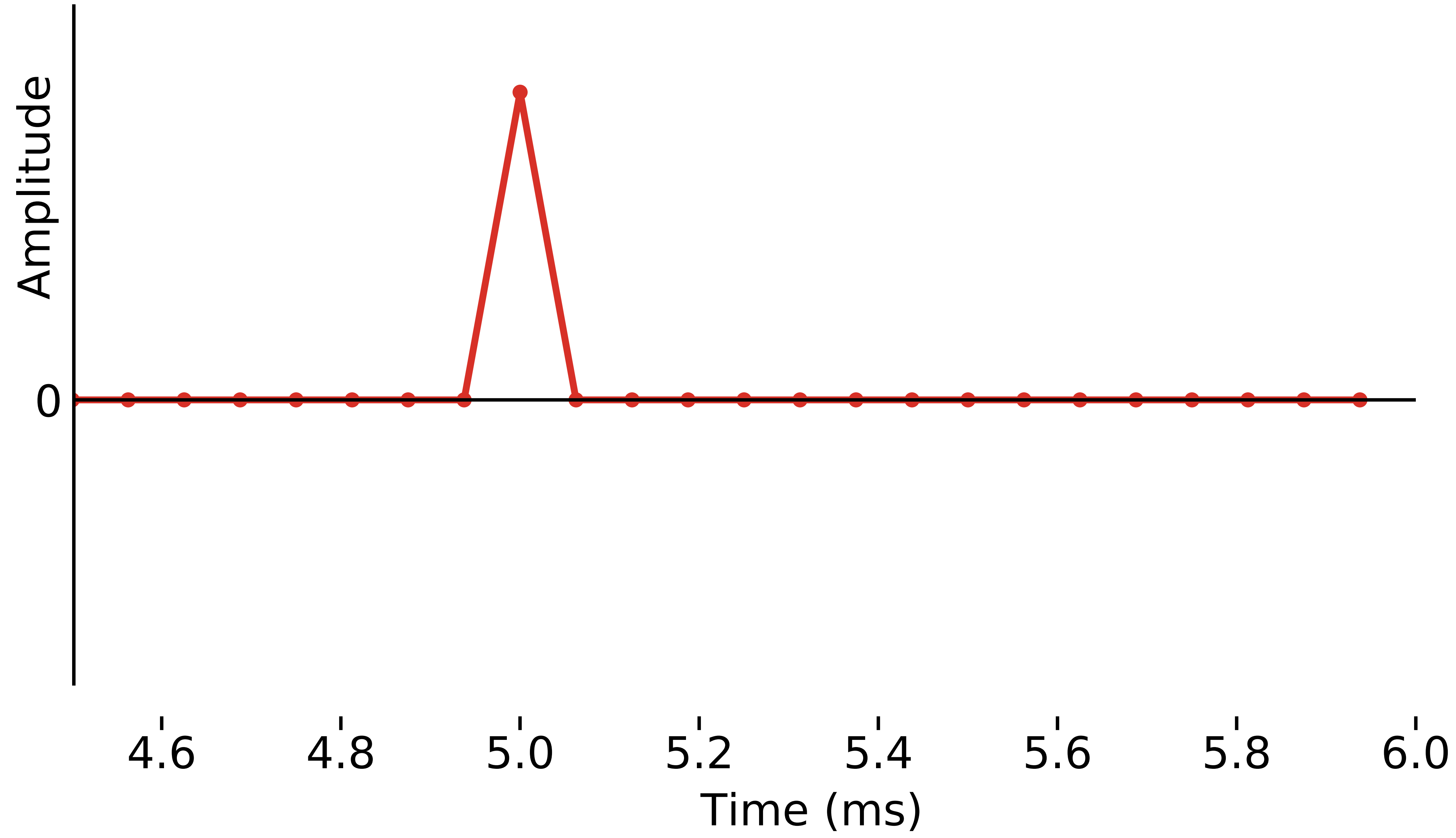


output $y[t]$



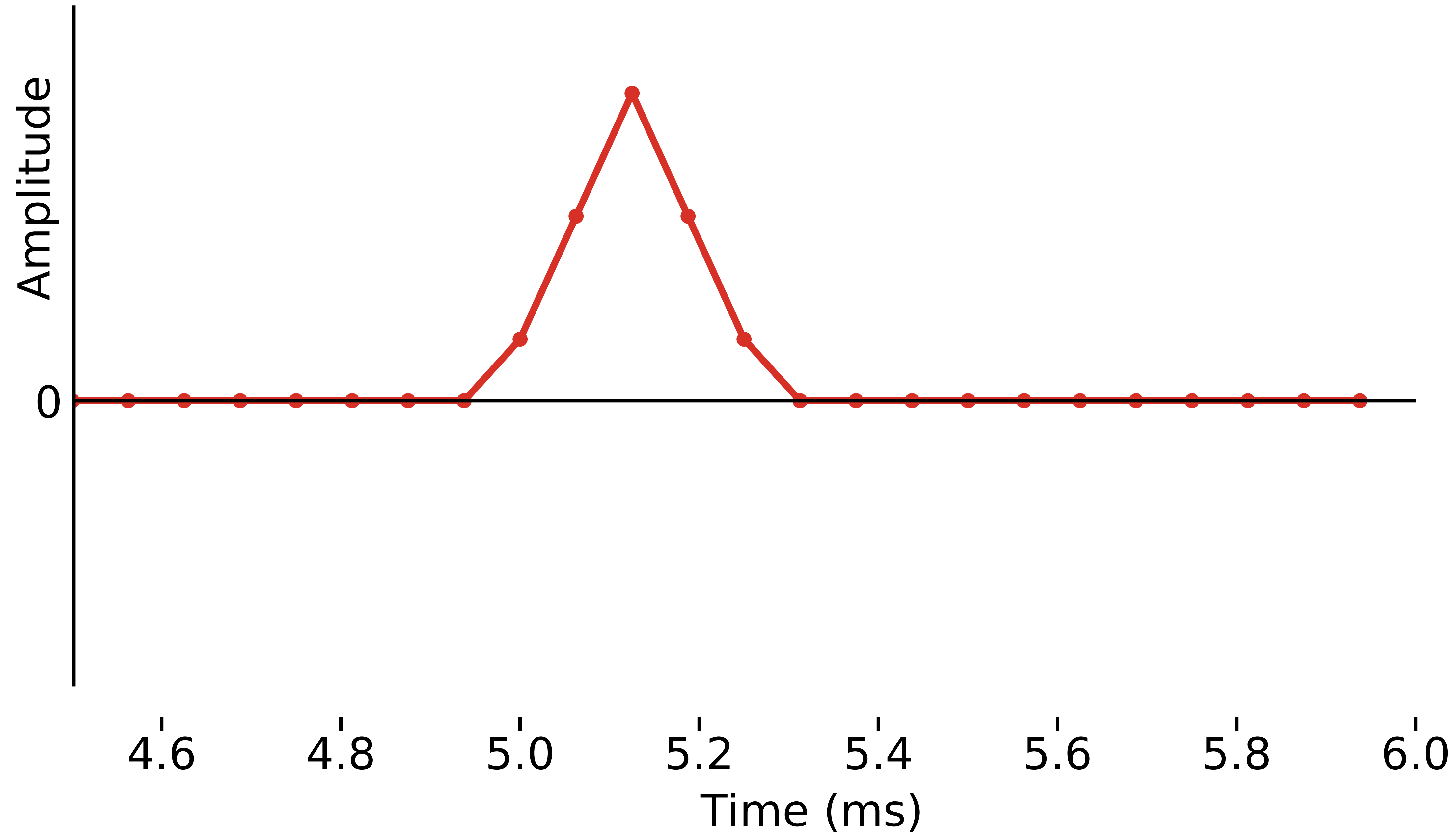
$$y[t] = 0.1x[t-4] + 0.3x[t-3] + 0.5x[t-2] + 0.3x[t-1] + 0.1x[t]$$

$x[t]$

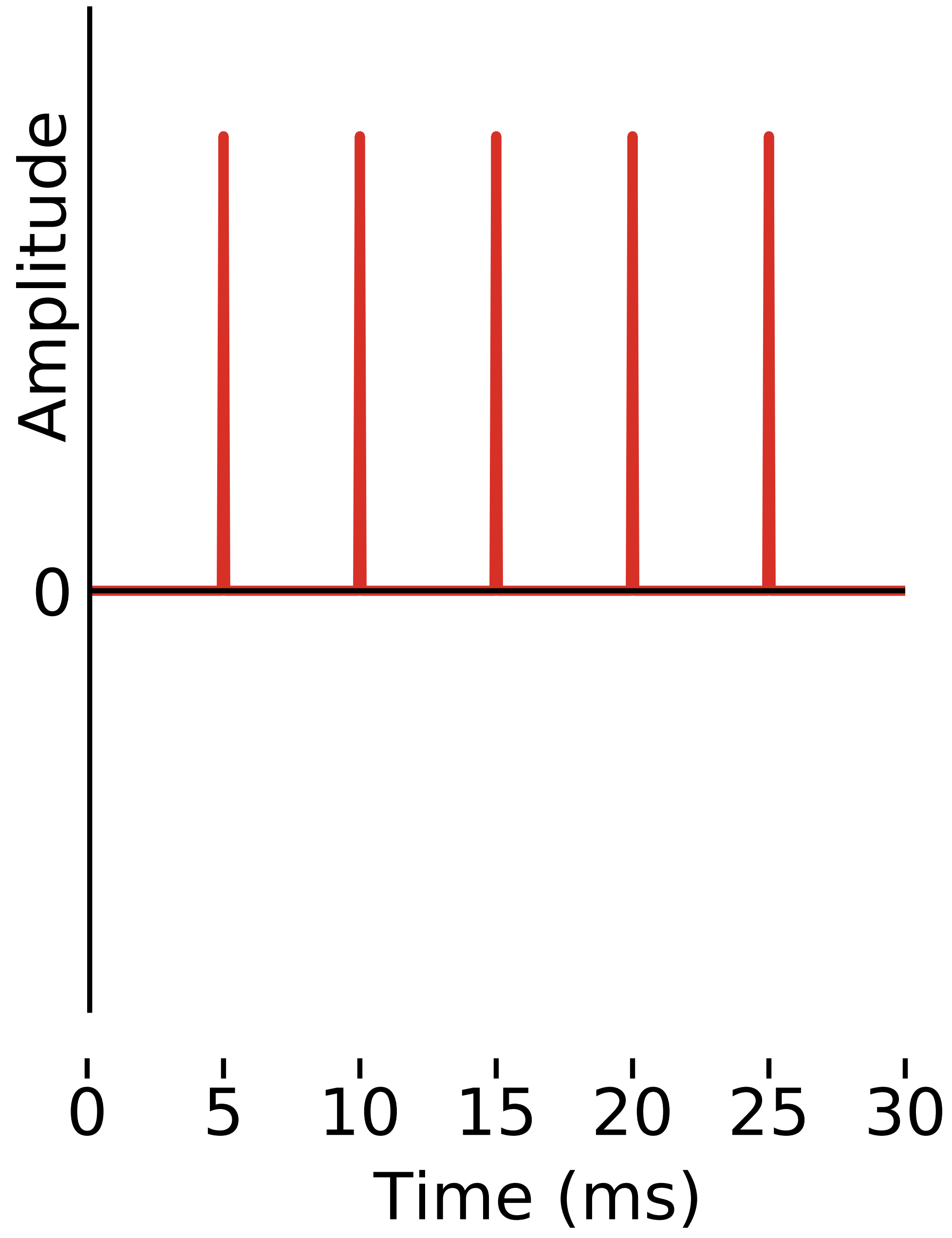


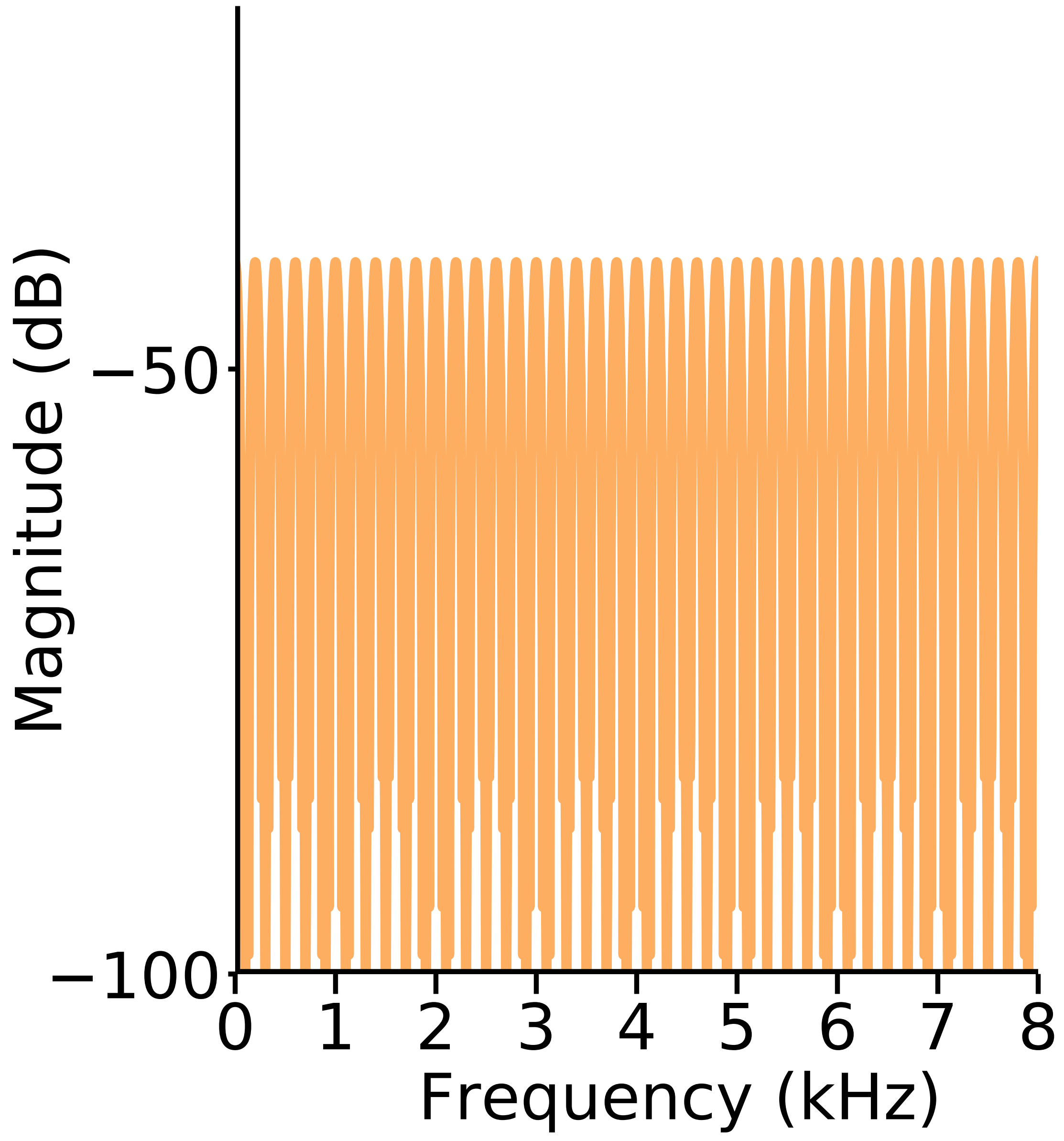
$$y[t] = 0.1x[t-4] + 0.3x[t-3] + 0.5x[t-2] + 0.3x[t-1] + 0.1x[t]$$

$y[t]$



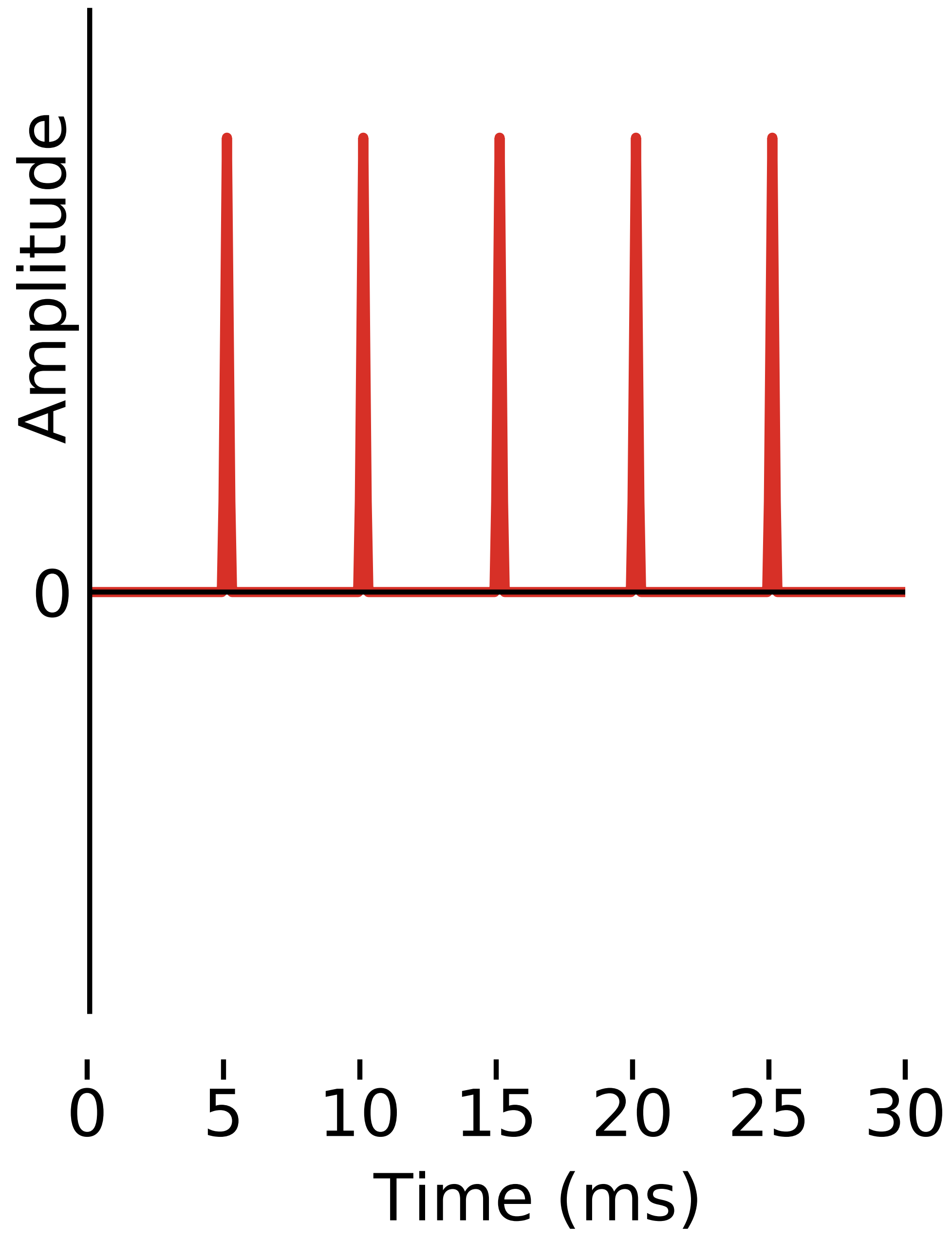
$x[t]$

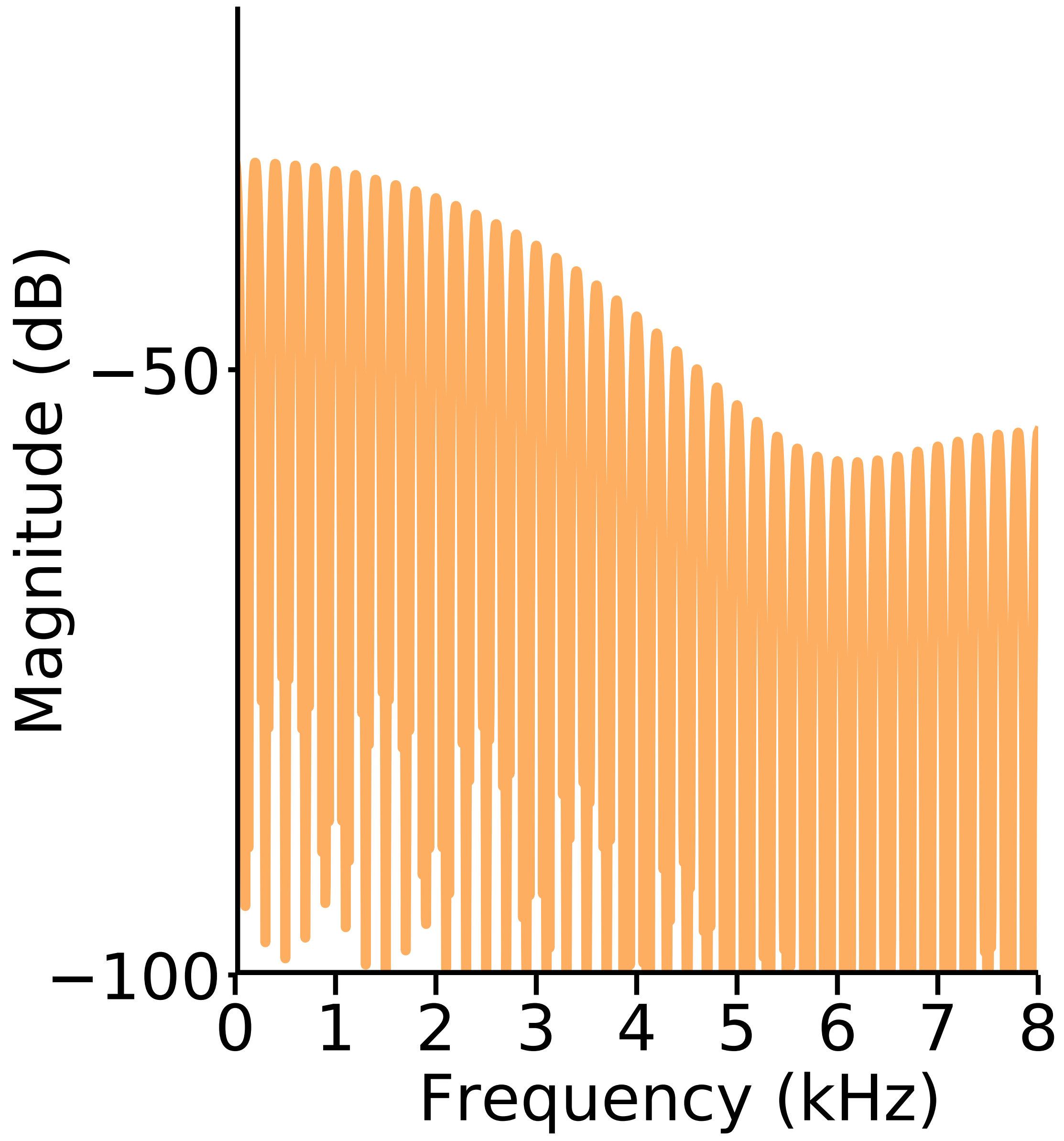




$$y[t] = 0.1x[t-4] + 0.3x[t-3] + 0.5x[t-2] + 0.3x[t-1] + 0.1x[t]$$

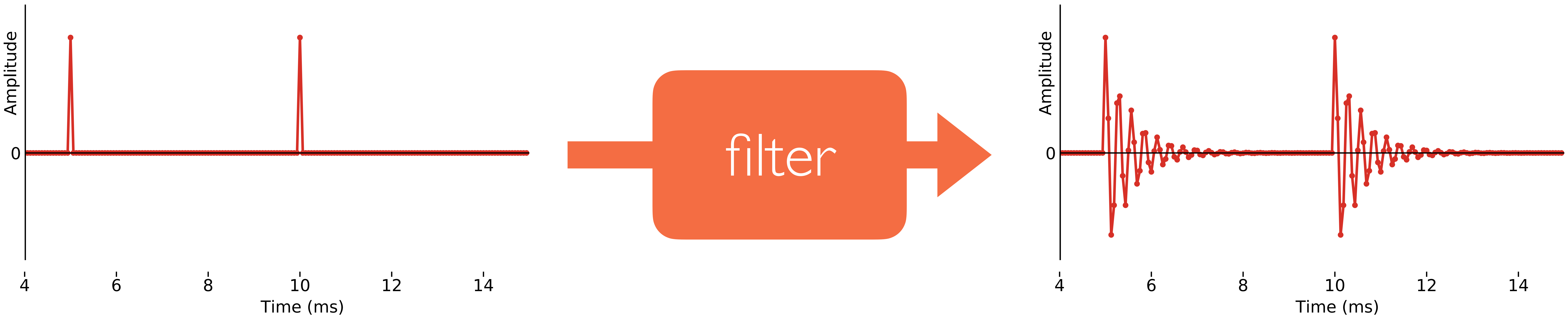
$y[t]$





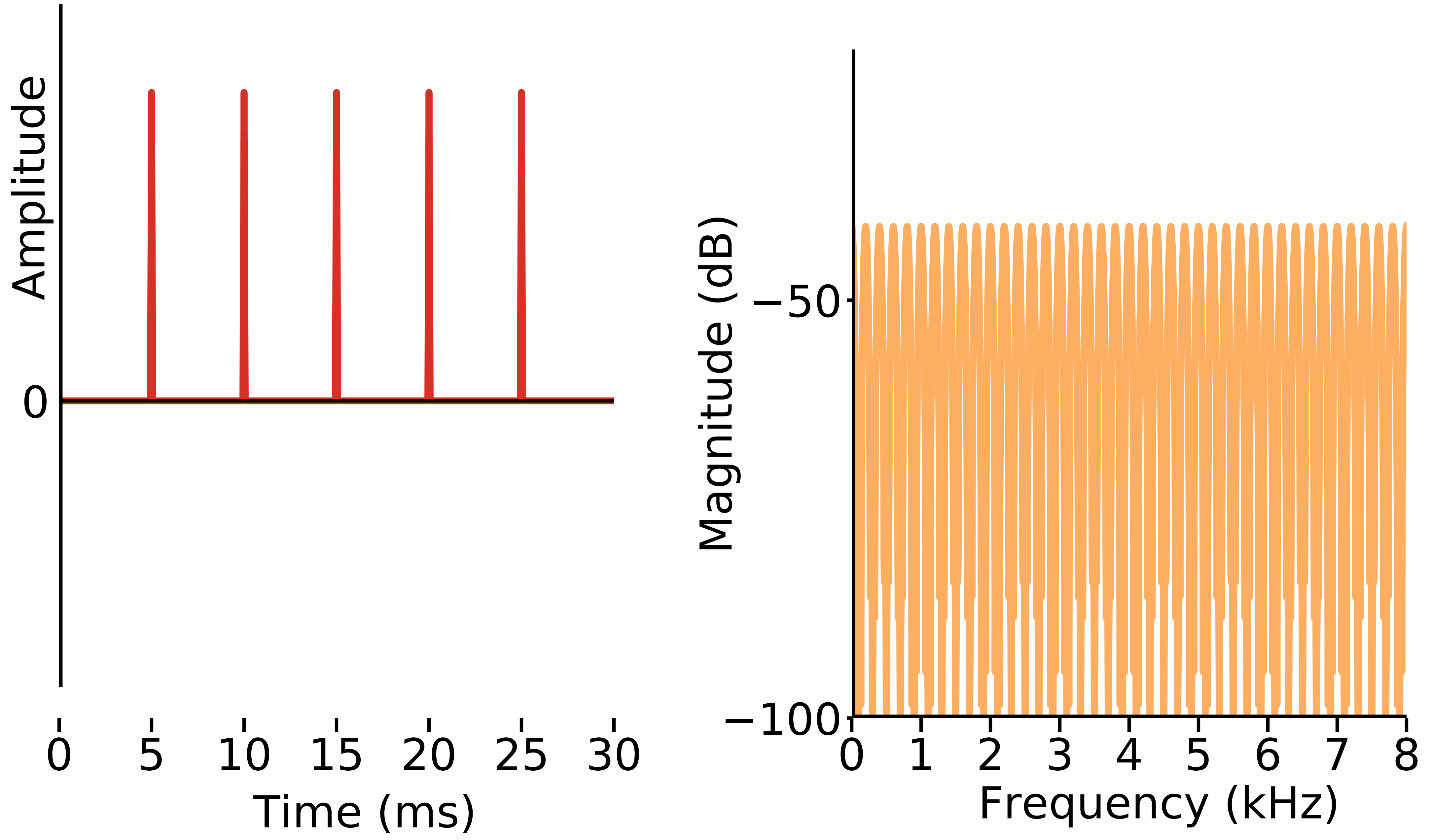
Filtering in action

input $x[t]$ output $y[t]$



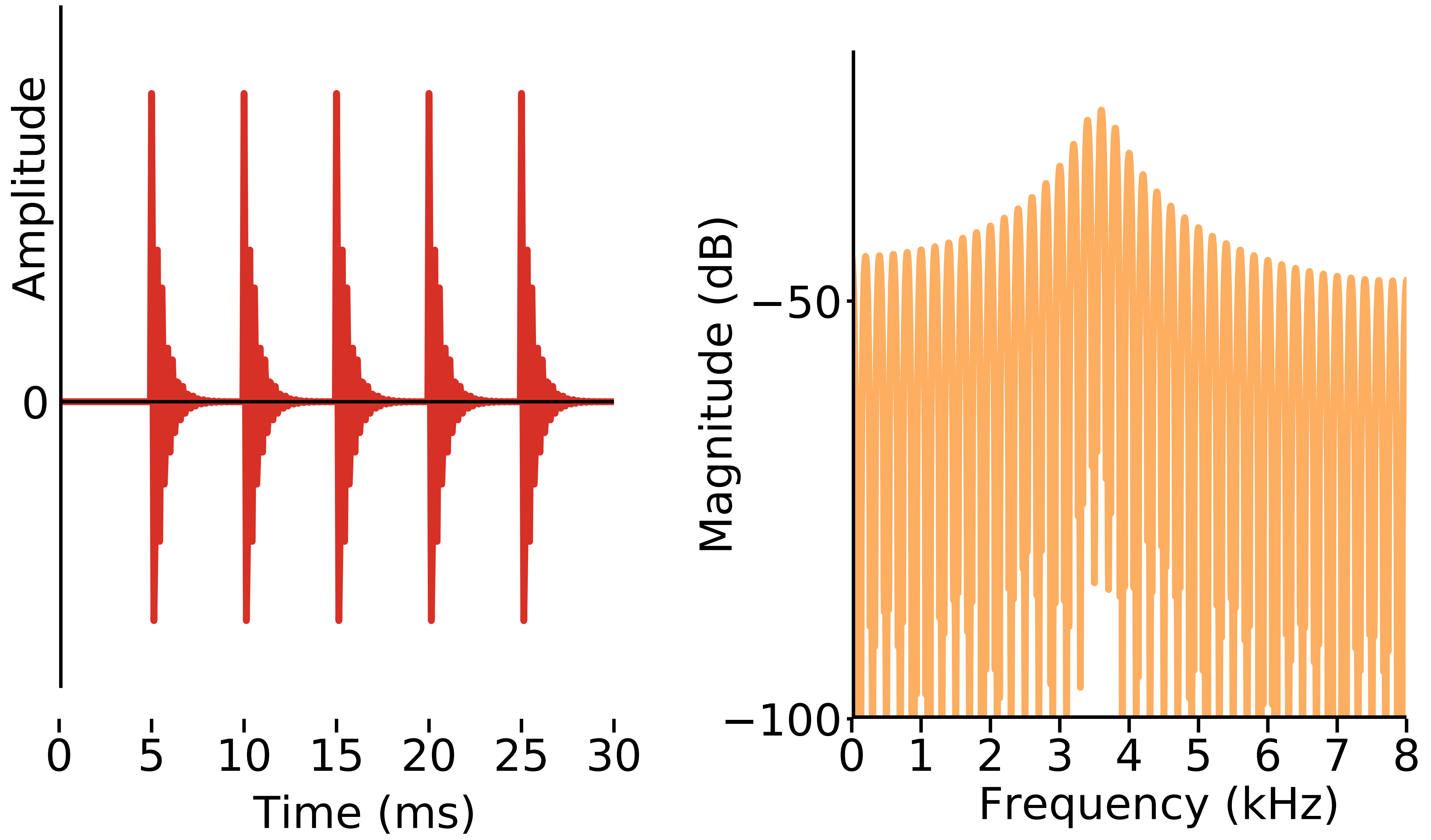
$$y[t] = 1.0x[t] - 1.0y[t-1] + 0.3y[t-2] - 0.8y[t-3]$$

$x[t]$



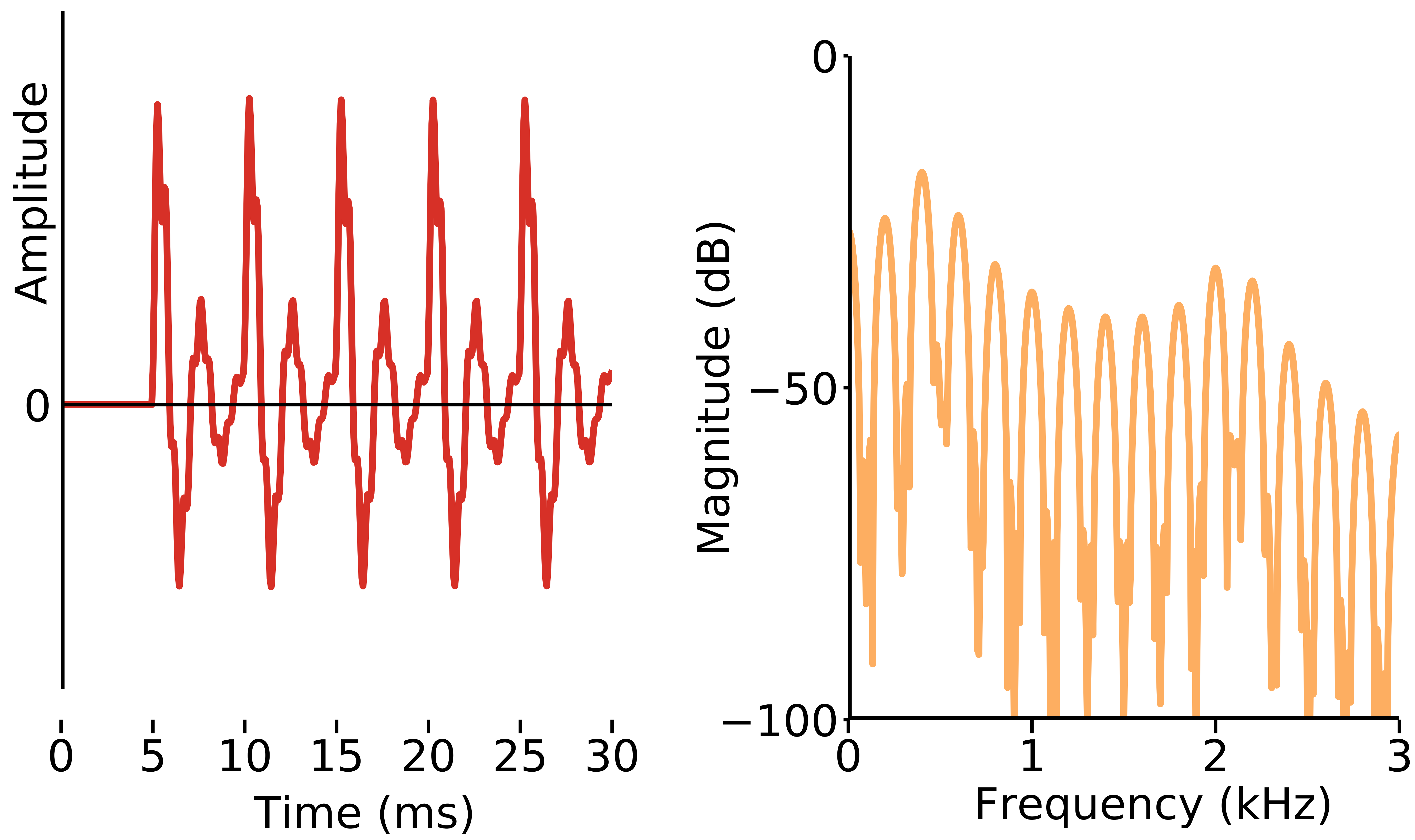
$$y[t] = 1.0x[t] - 1.0y[t-1] + 0.3y[t-2] - 0.8y[t-3]$$

$y[t]$

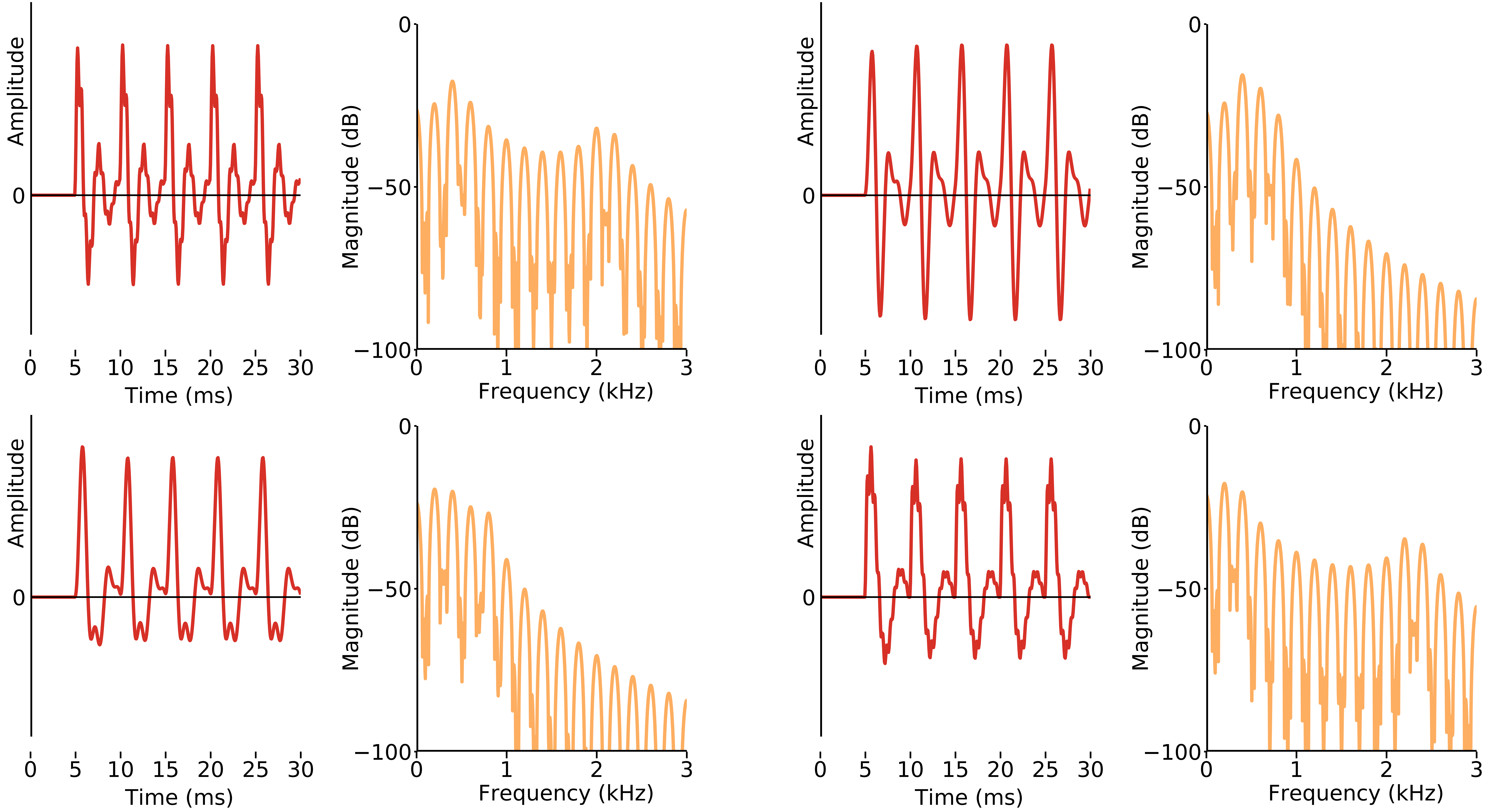


A filter with similar properties to the vocal tract

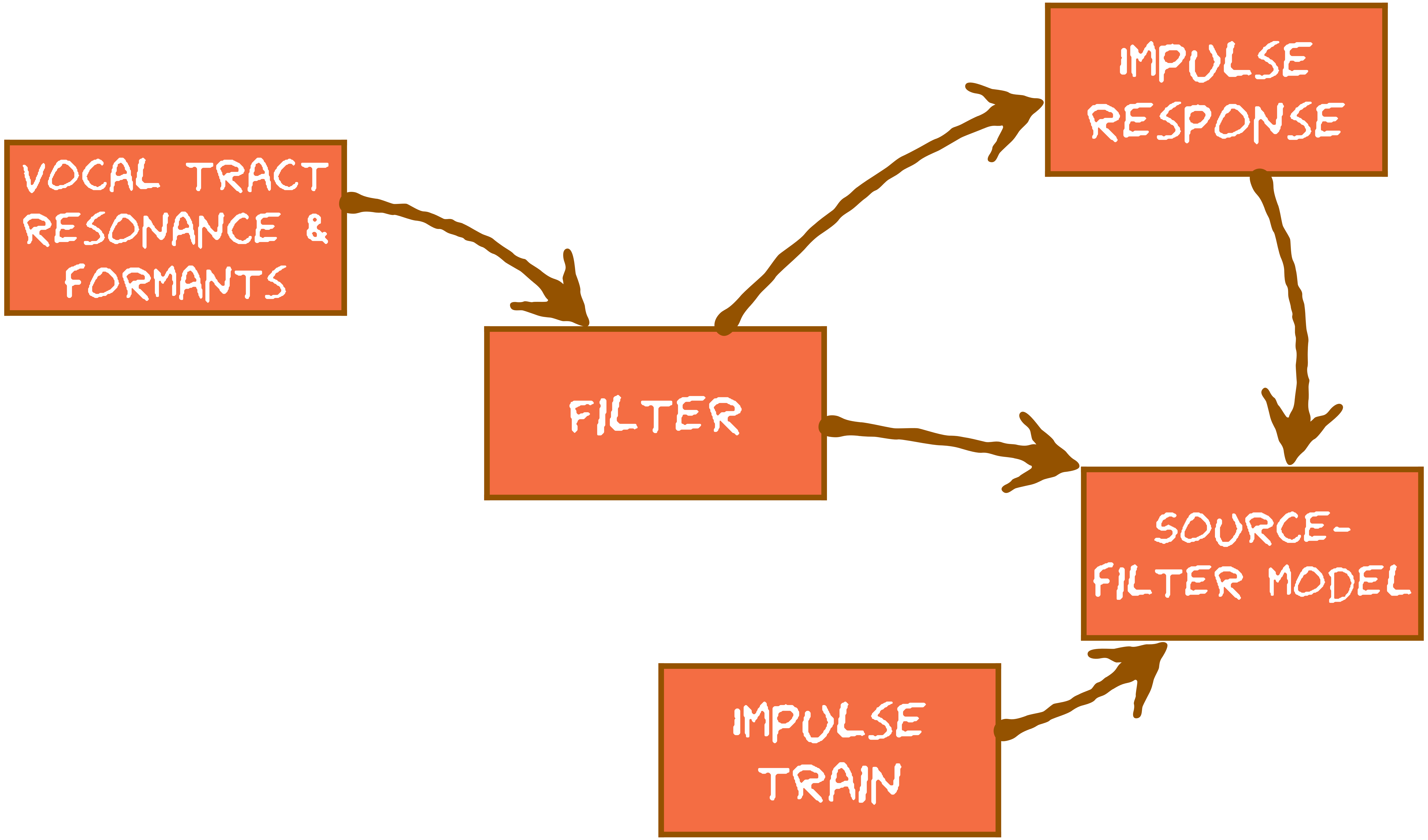
$y[t]$



$$y[t] = 1.0x[t] + 3.2y[t-1] - 4.4y[t-2] + 3.0y[t-3] - 0.9y[t-4]$$

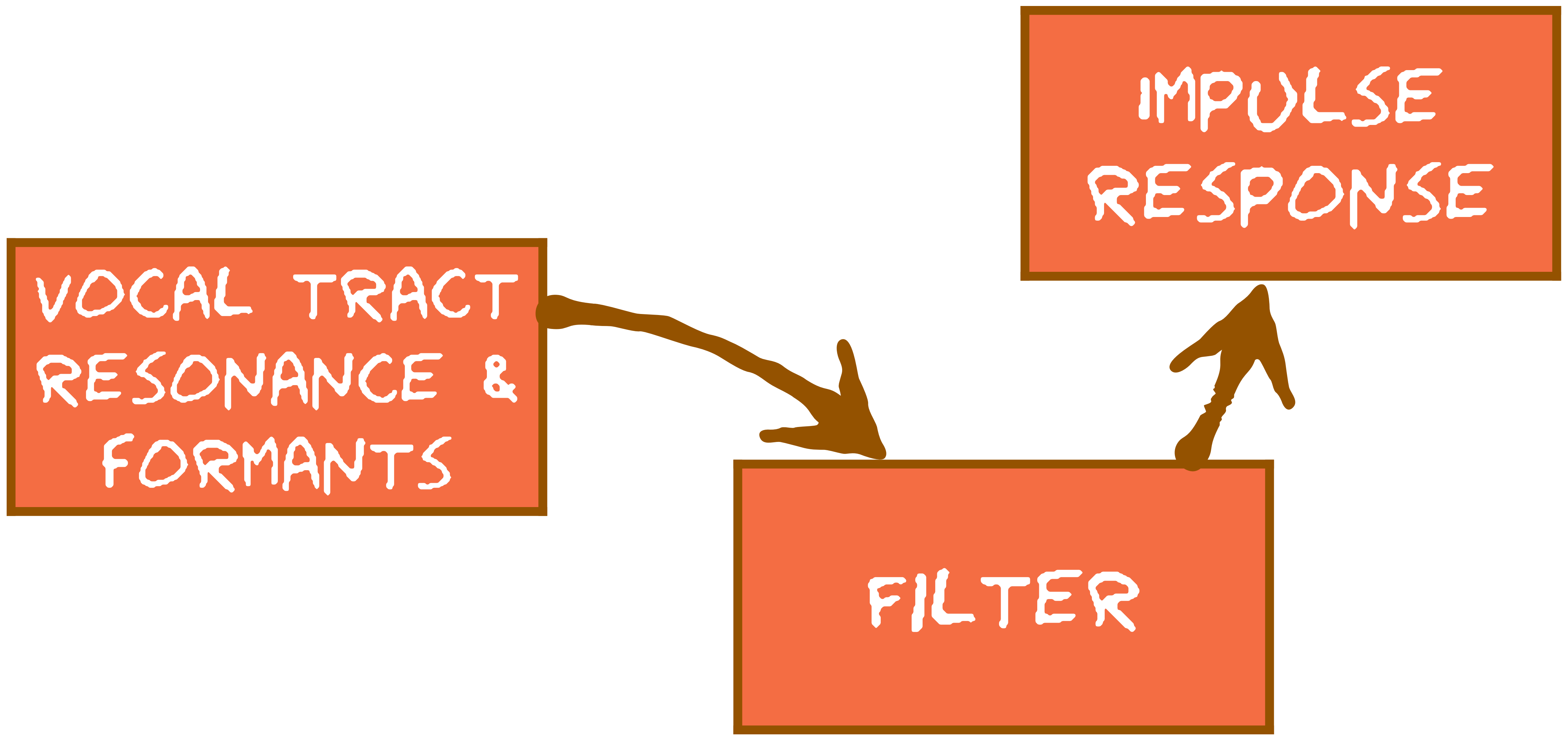

Amplitude
0
0 5 10 15 20 25 30
Time (ms)
Magnitude (dB)
0
−50
−100
0 1 2 3
Frequency (kHz)

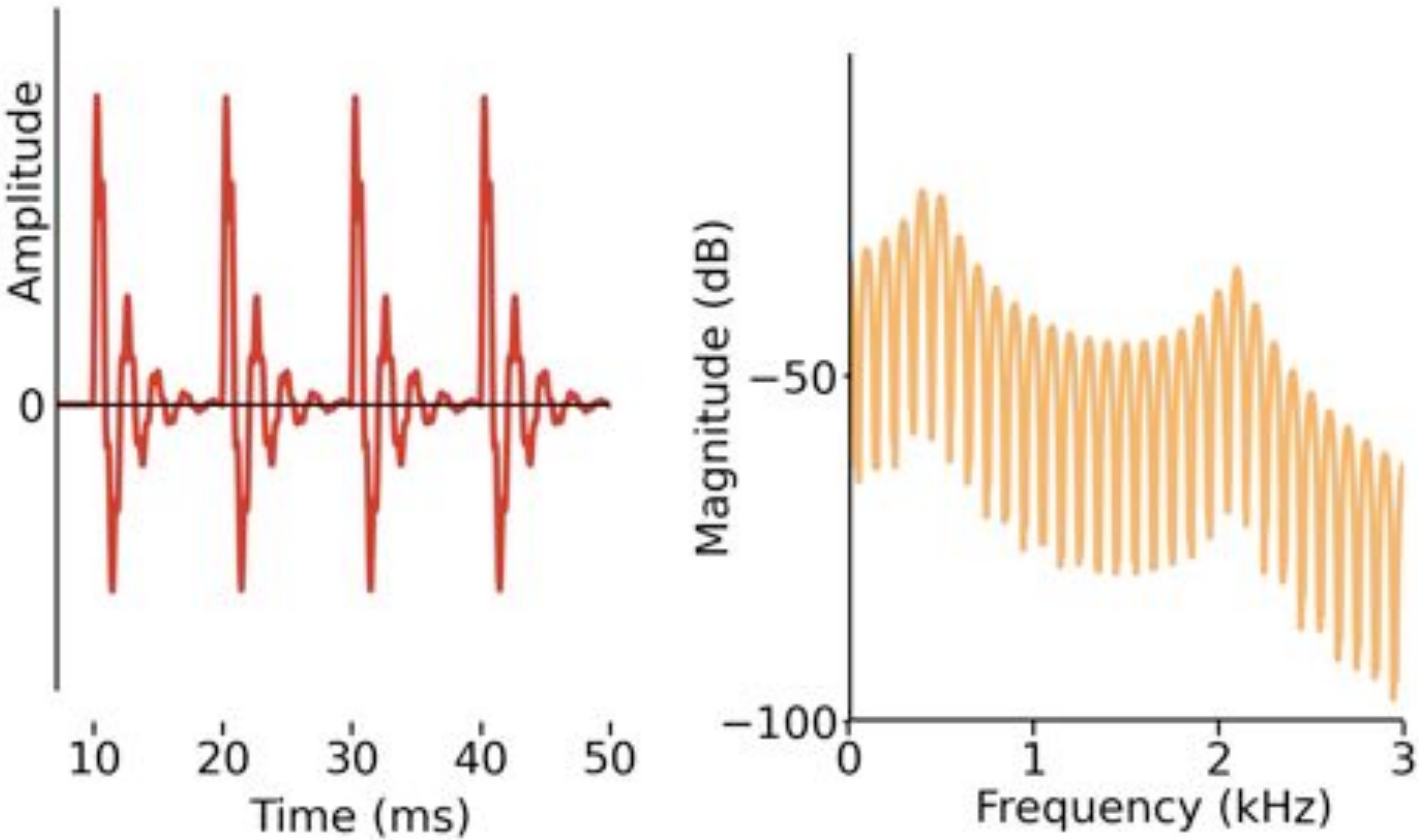
What you can learn next



IMPULSE RESPONSE

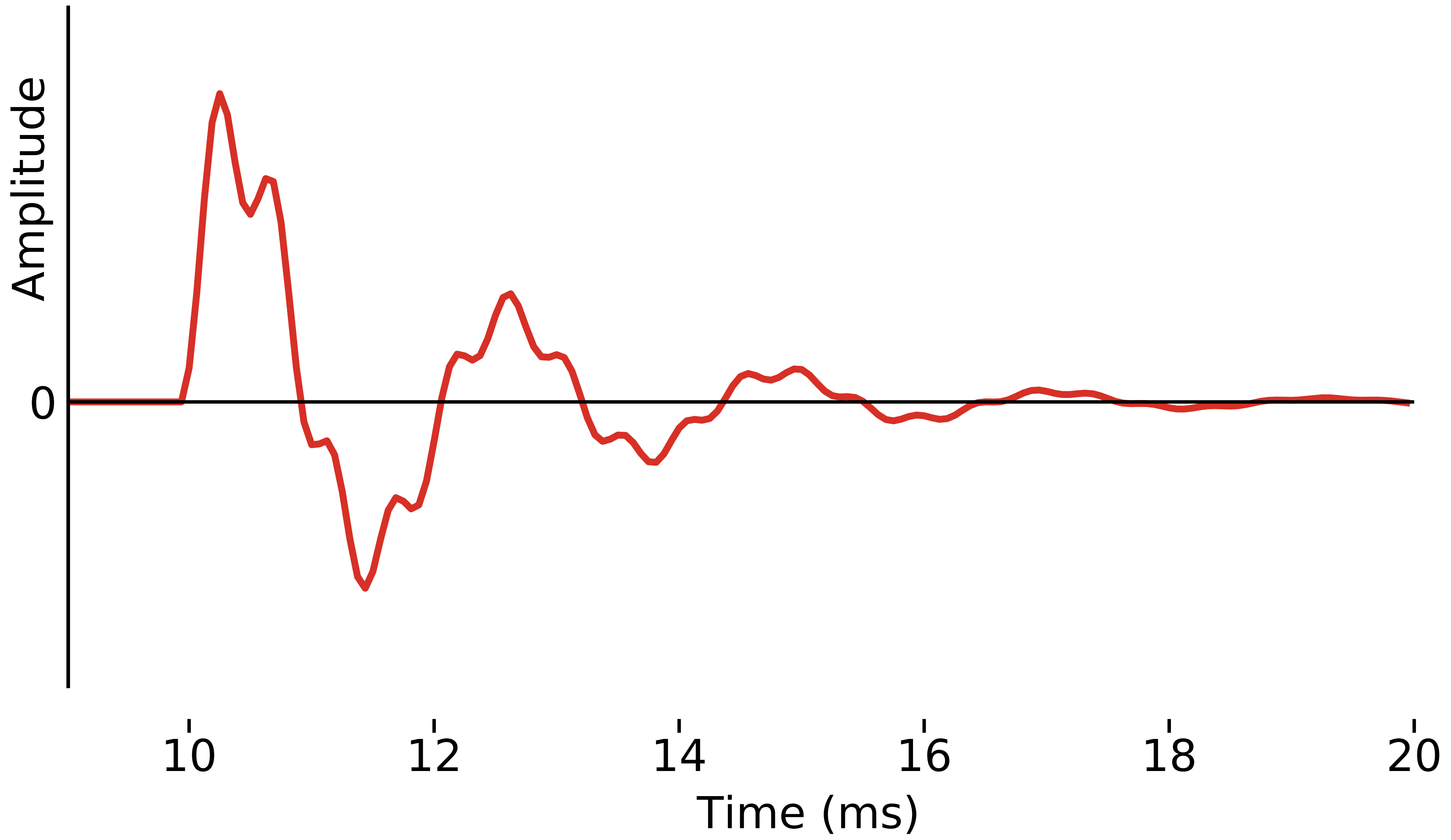
What you need to know already

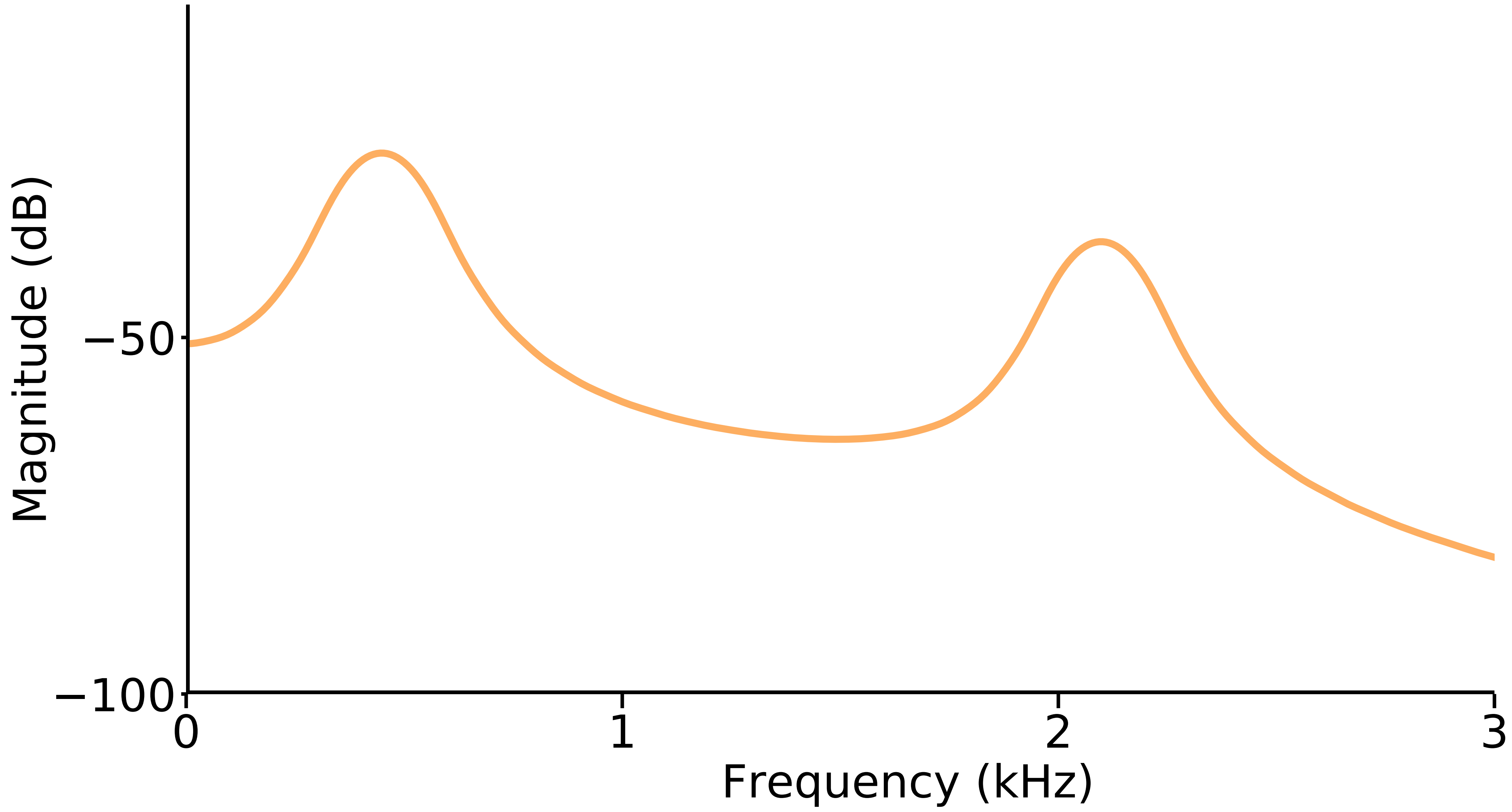




Amplitude
0
10
20
30
40
50
Time (ms)
Magnitude (dB)
−50
−100
0
1
2
3
Frequency (kHz)

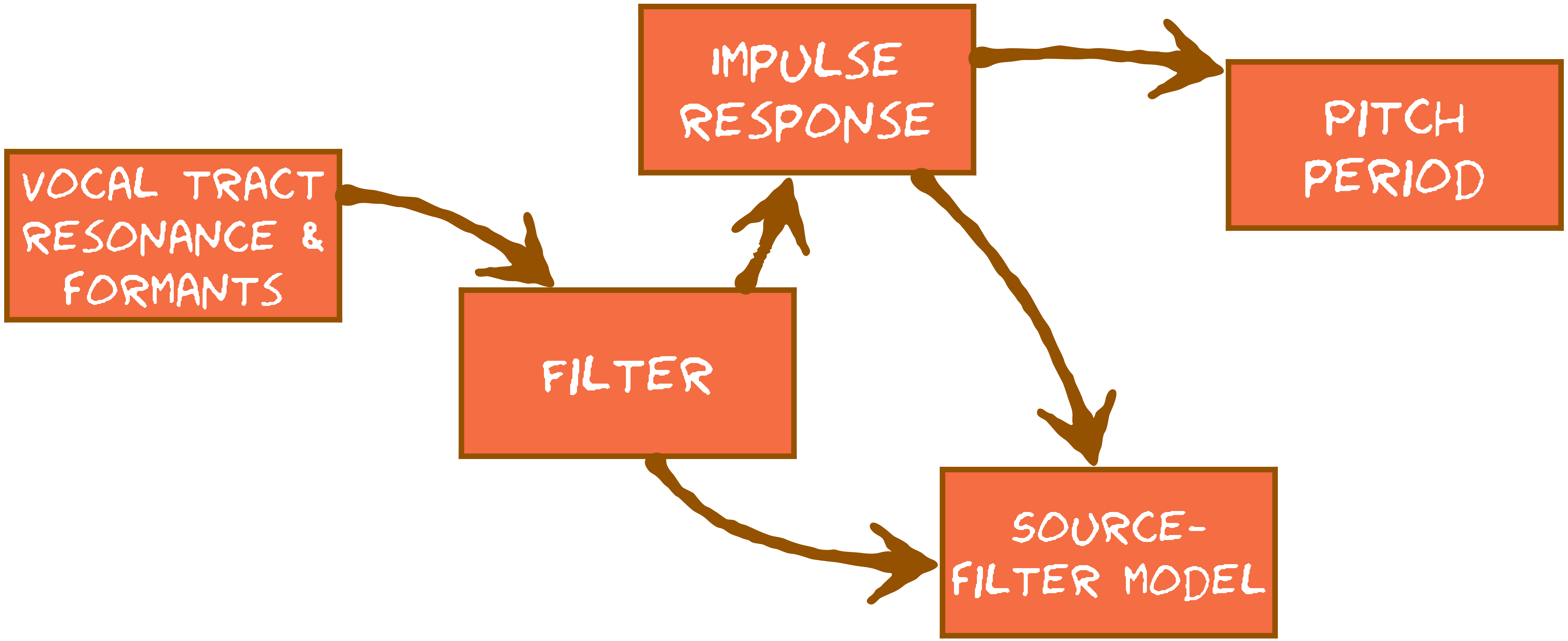
Describing a linear filter in different domains





$$y[t] = 1.0x[t] + 3.2y[t-1] - 4.4y[t-2] + 3.0y[t-3] - 0.9y[t-4]$$

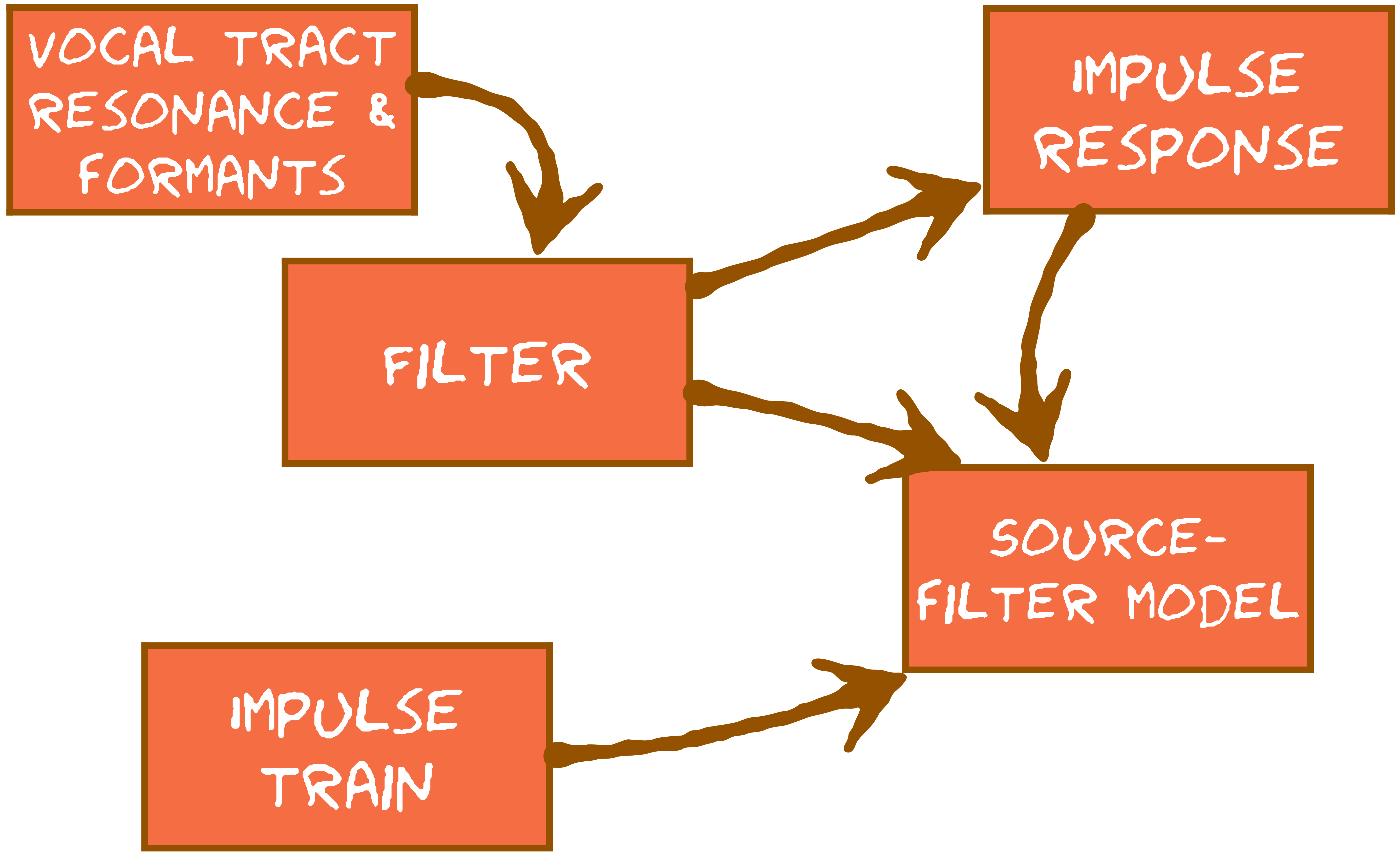
What you can learn next



SOURCE-FILTER MODEL

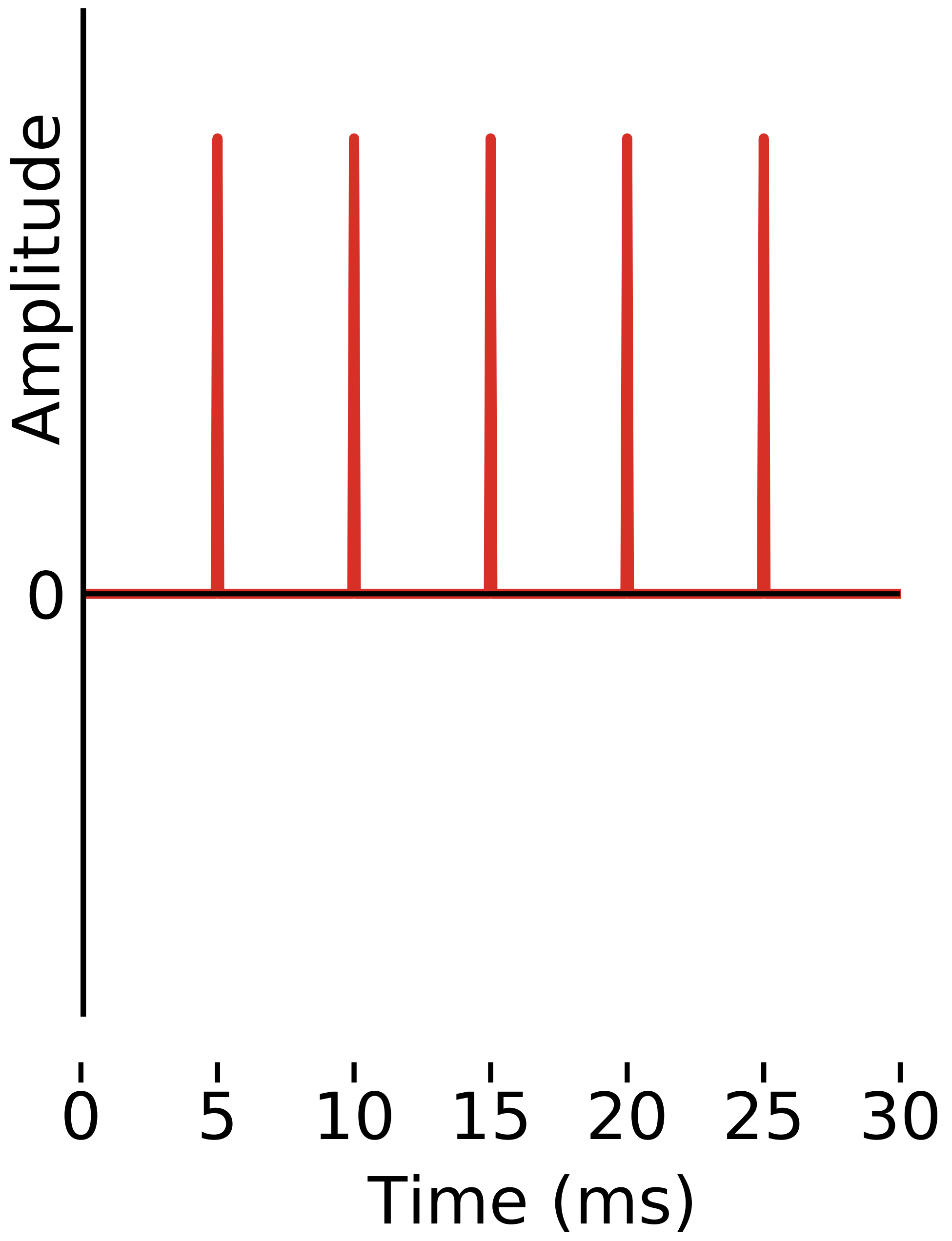
THE VOCAL TRACT IS A FILTER

What you need to know already



Source-filter model

$$s[t] = e[t] + \sum_{k=1}^{p} a_k \, s[t-k]$$

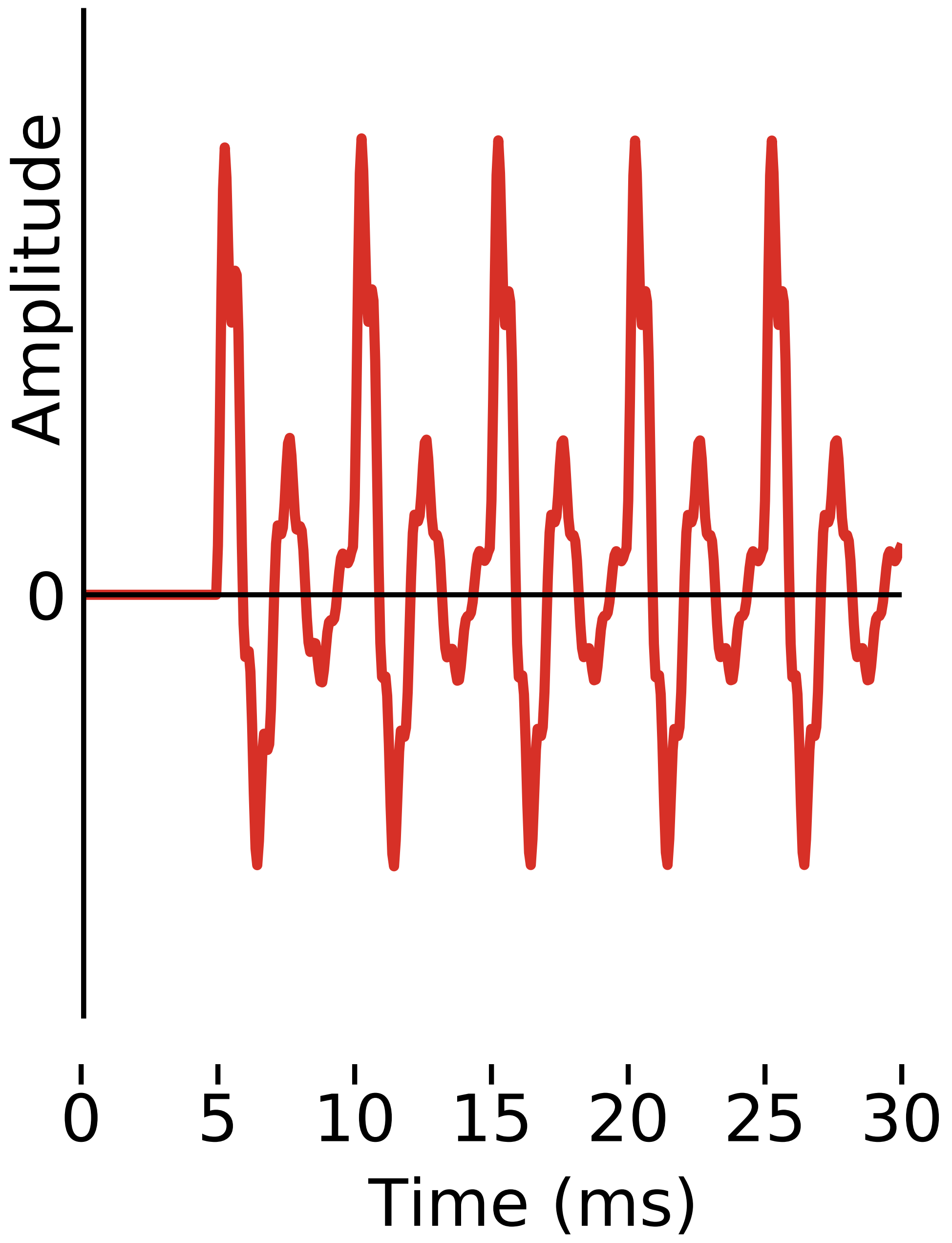


excitation signal

$e[t]$

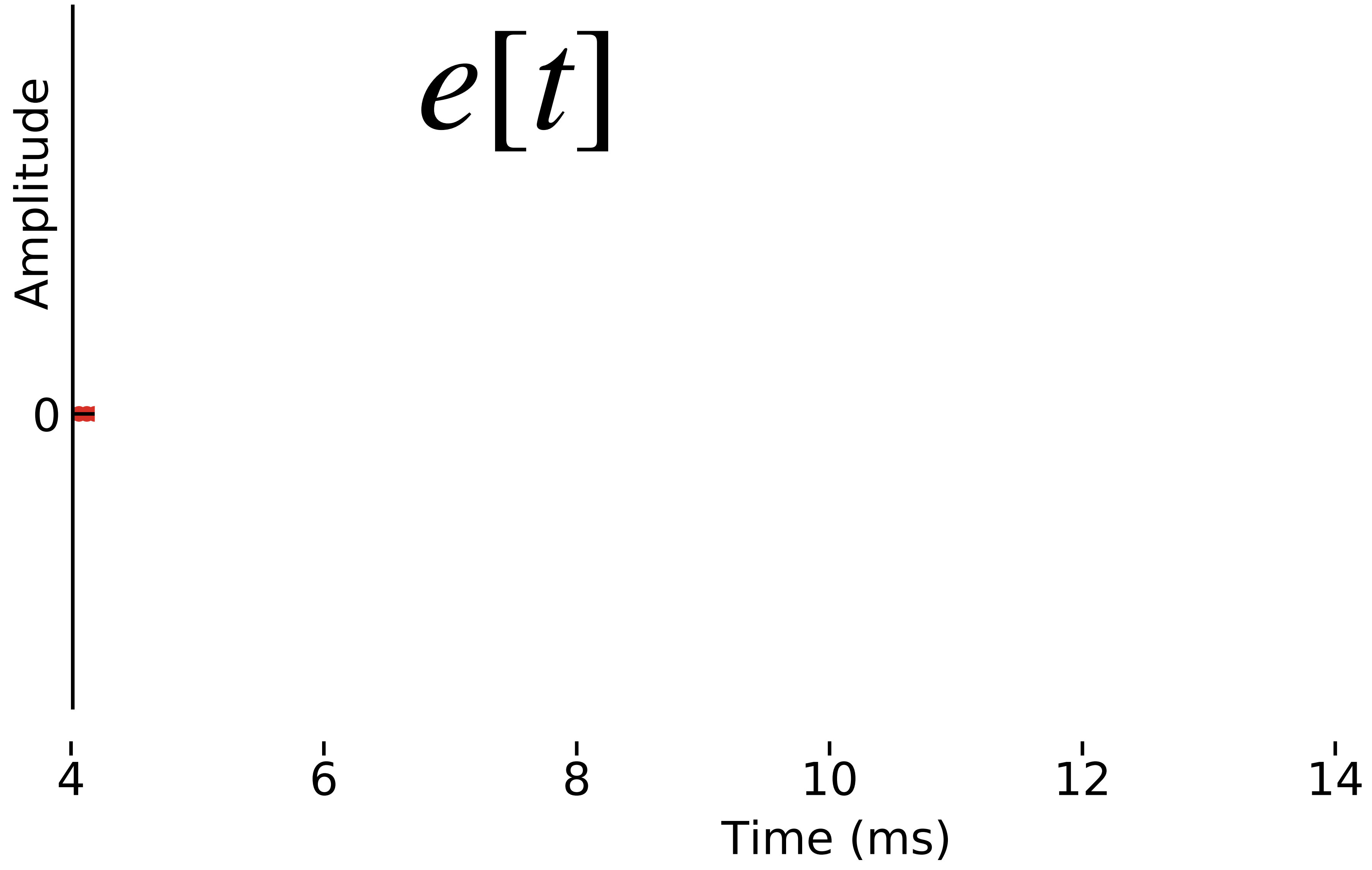
filter, with coefficients $\{a_1, a_2, \ldots a_p\}$

$s[t]$



speech signal

Describing the model in the time domain

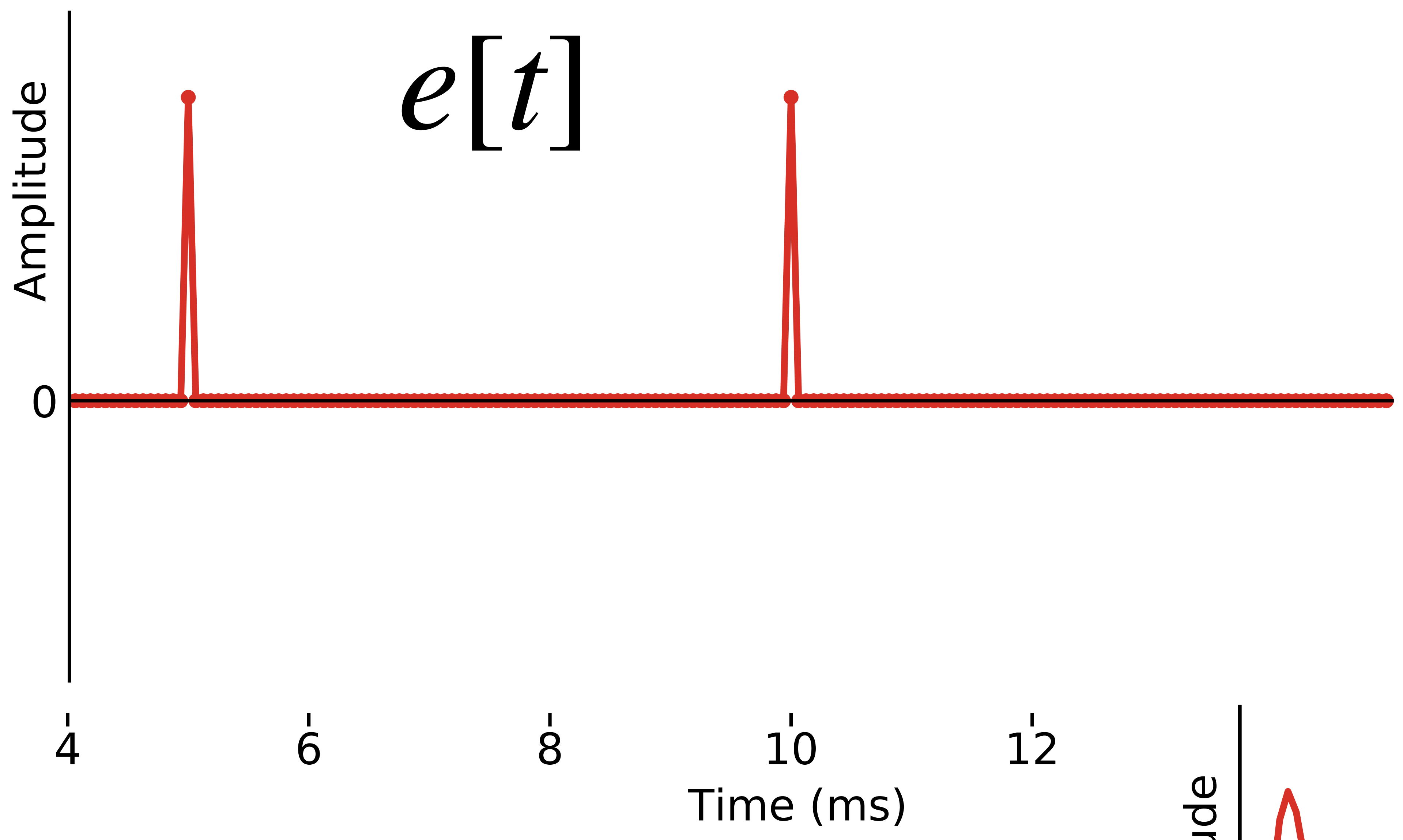


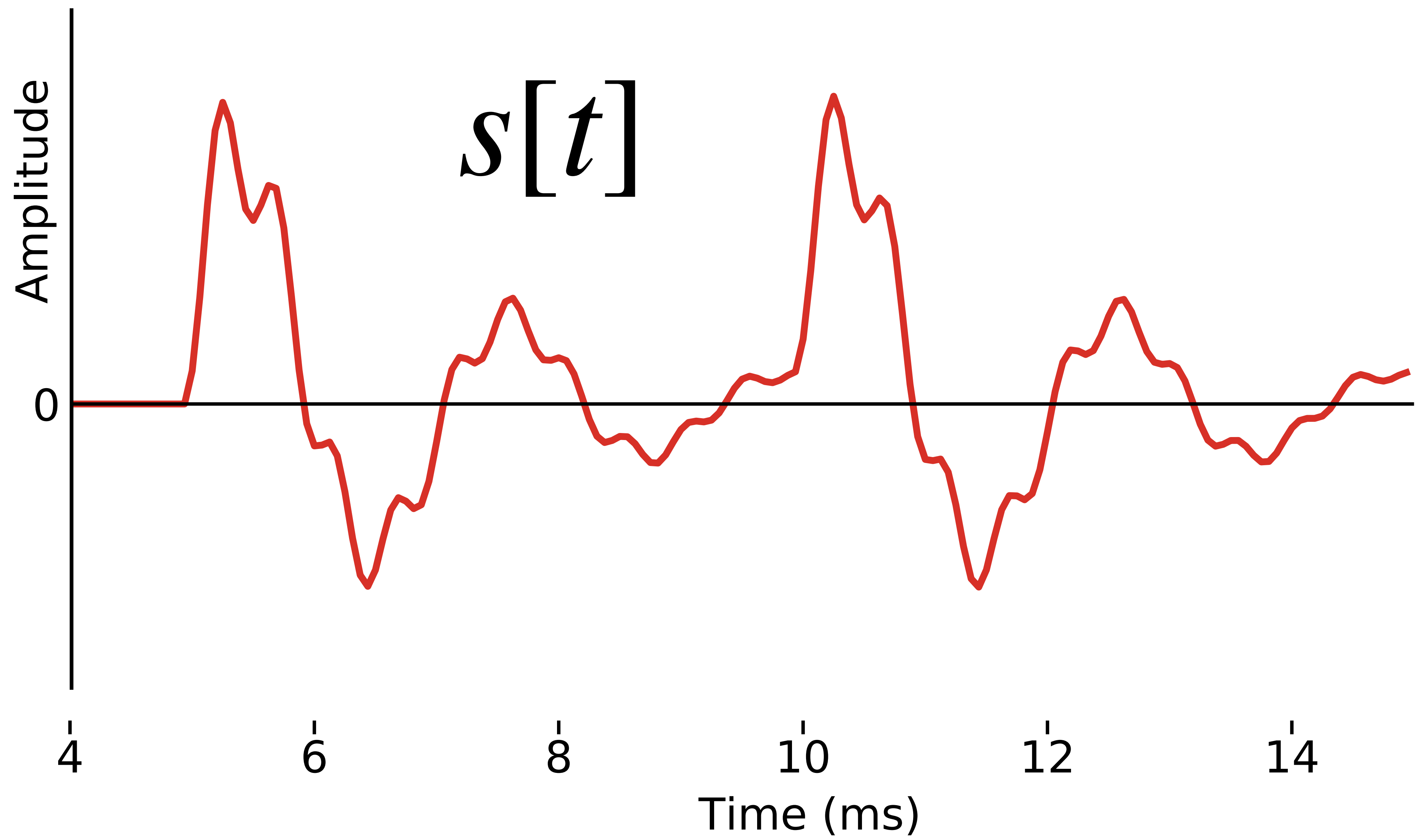
$e[t]$

$s[t]$

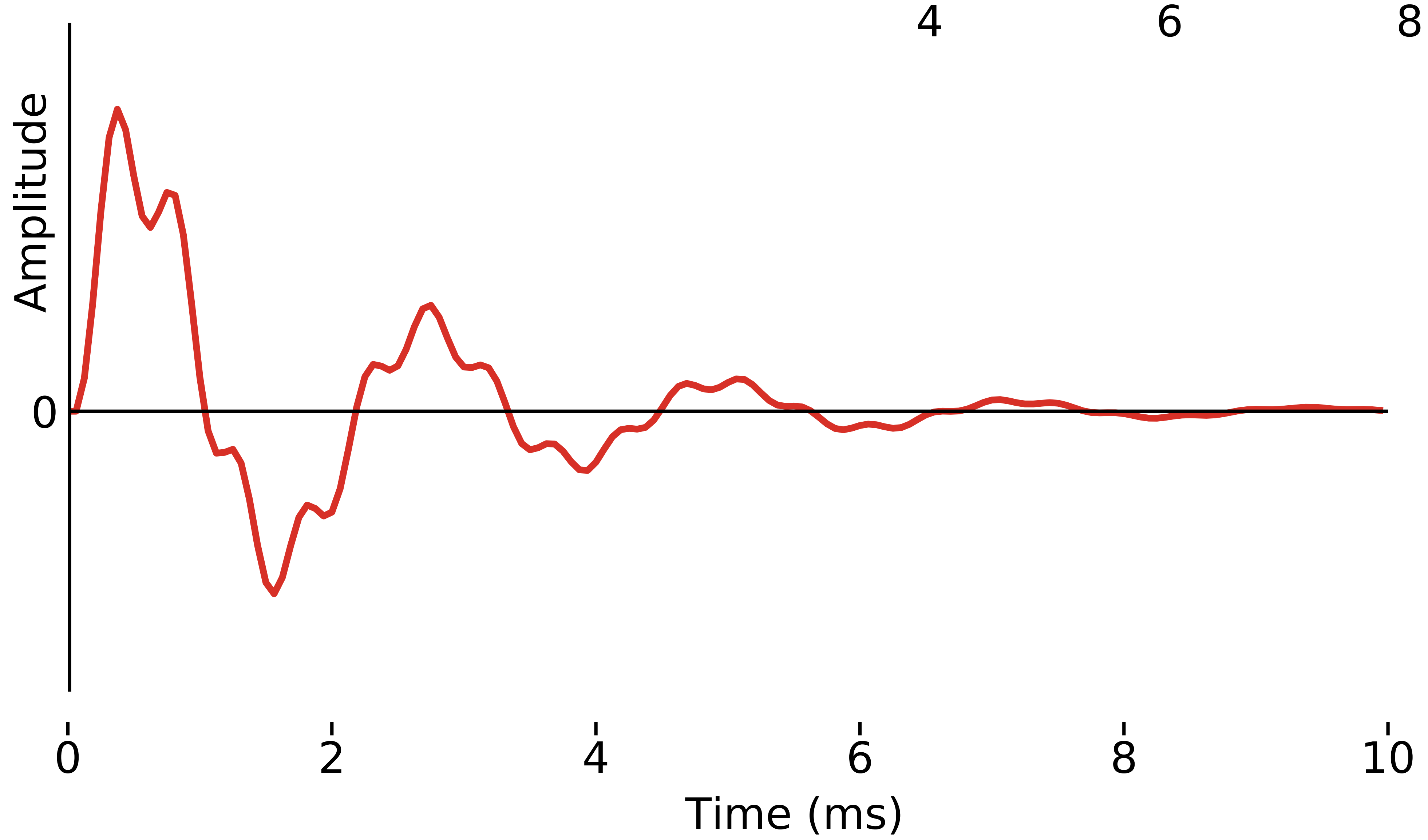
$$s[t] = e[t] + \sum_{k=1}^{p} a_k \, s[t-k]$$

Describing the model in the time domain





filter impulse
response



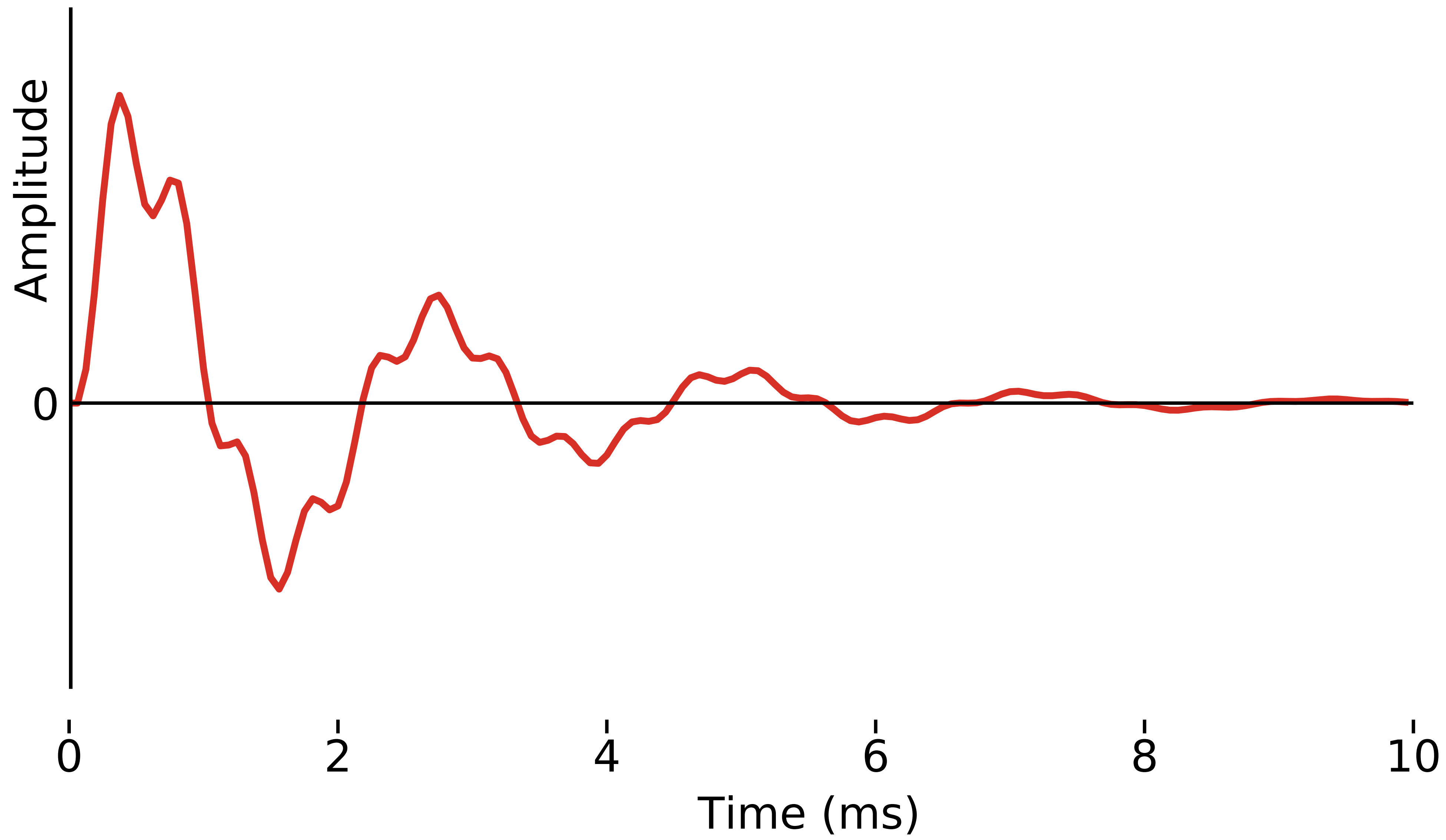
Making speech!

[e]

$$s[t] = e[t] + \sum_{k=1}^{p} a_k \, s[t-k]$$

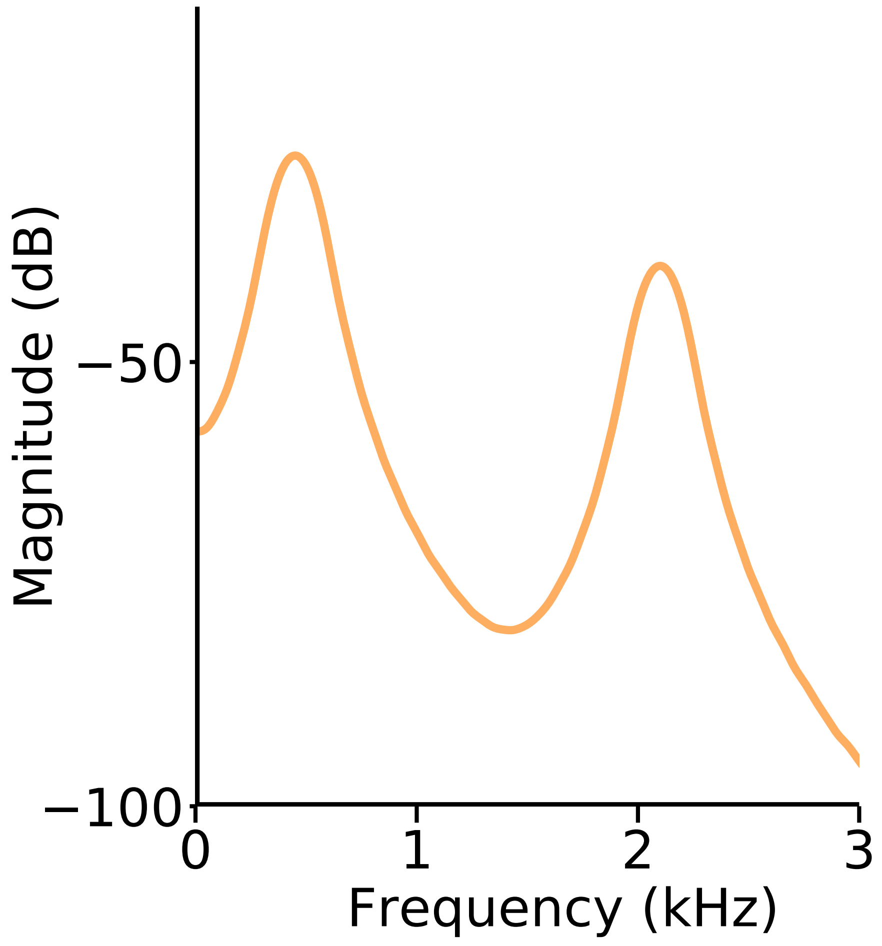
Making speech!

[e]



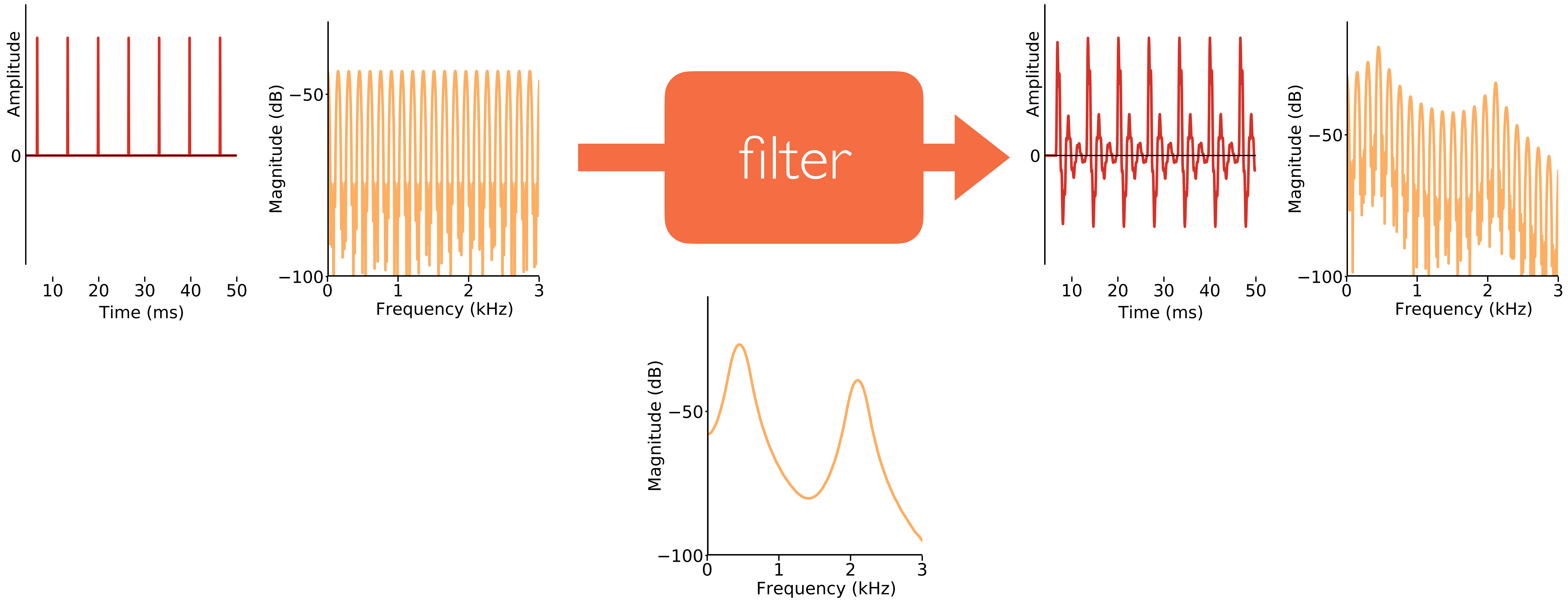
Making speech!

[e]



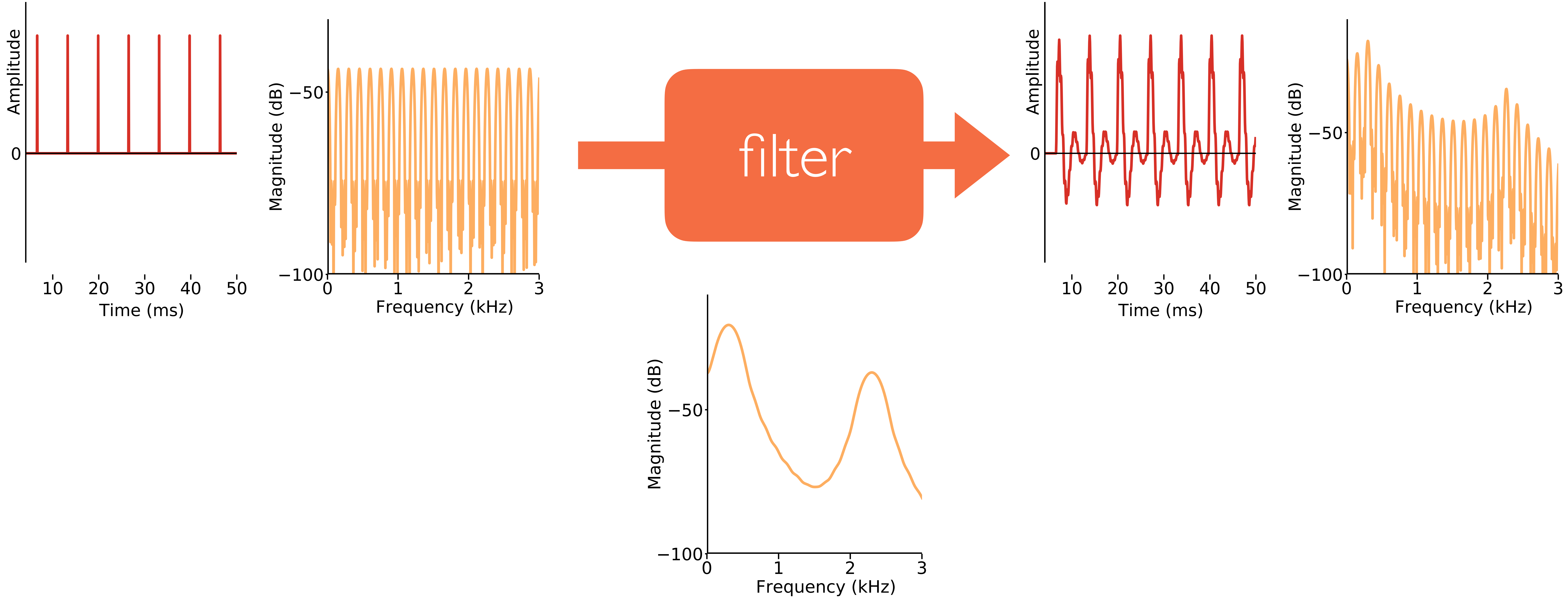
Making speech!

[e]



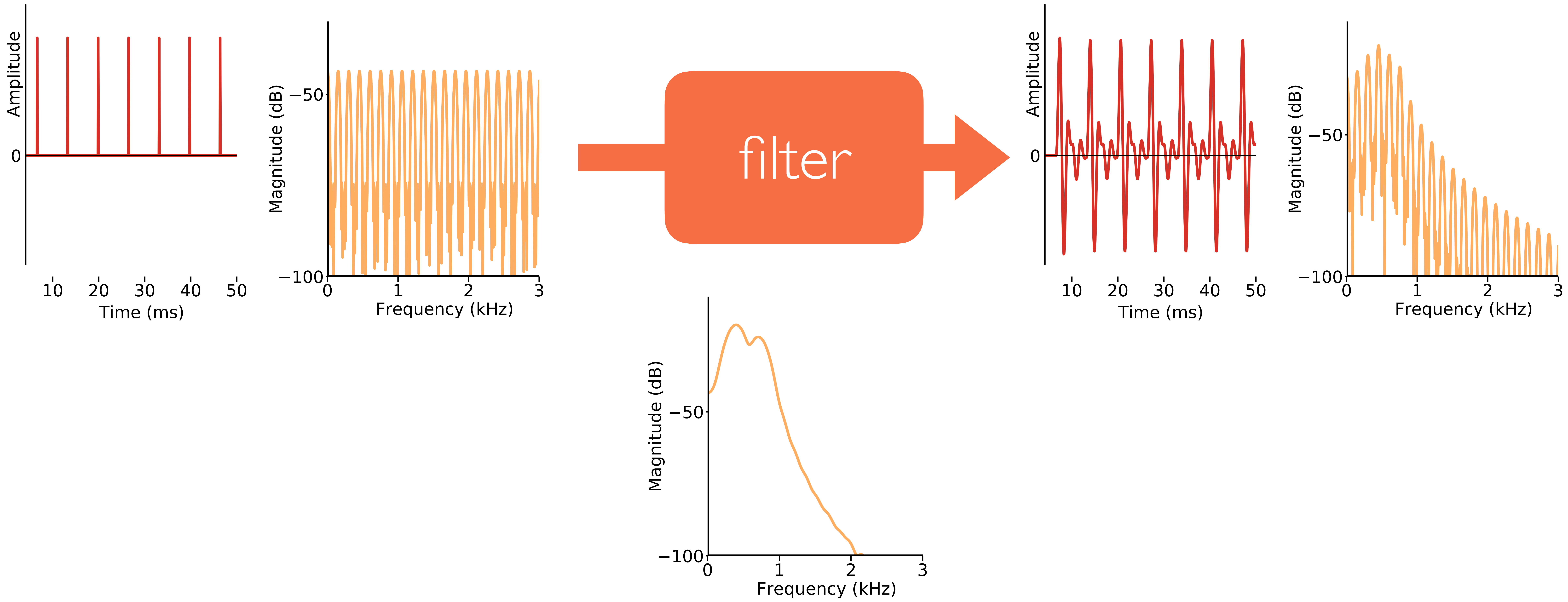
Making speech!

[ɪ]



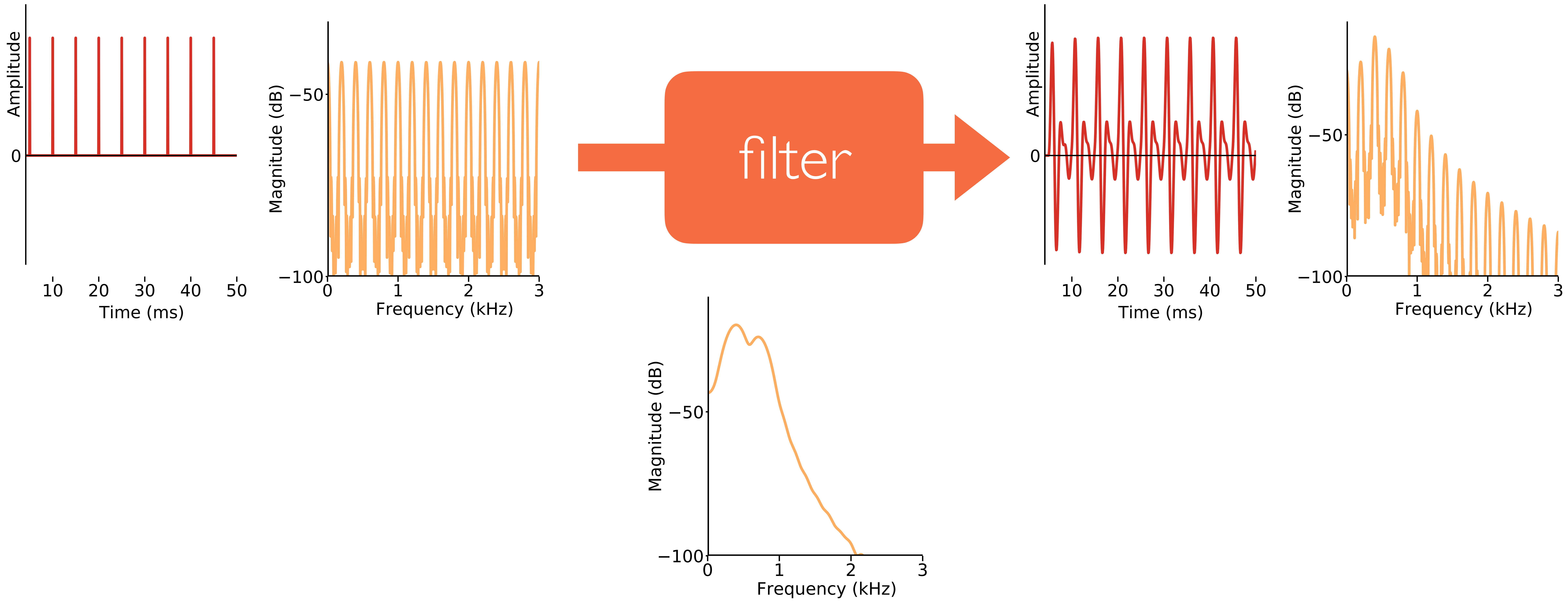
Making speech!

[ɔː]



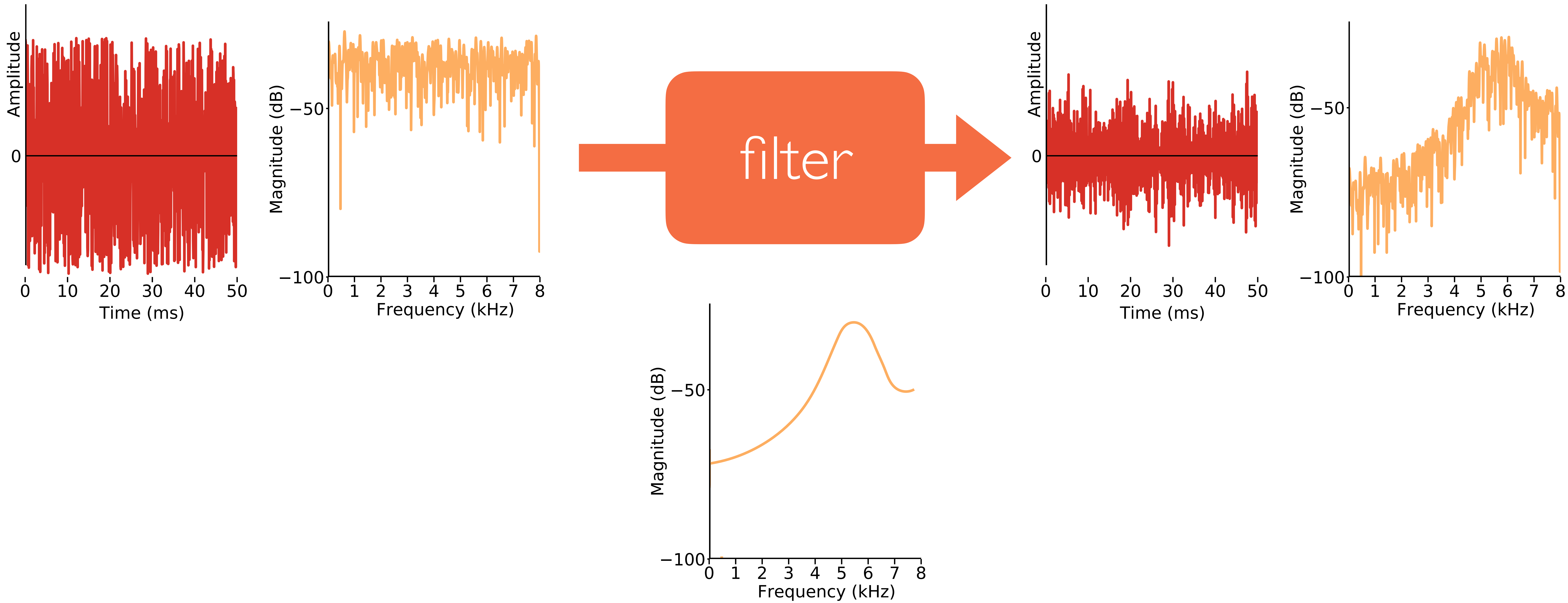
Making speech!

[ɔː]



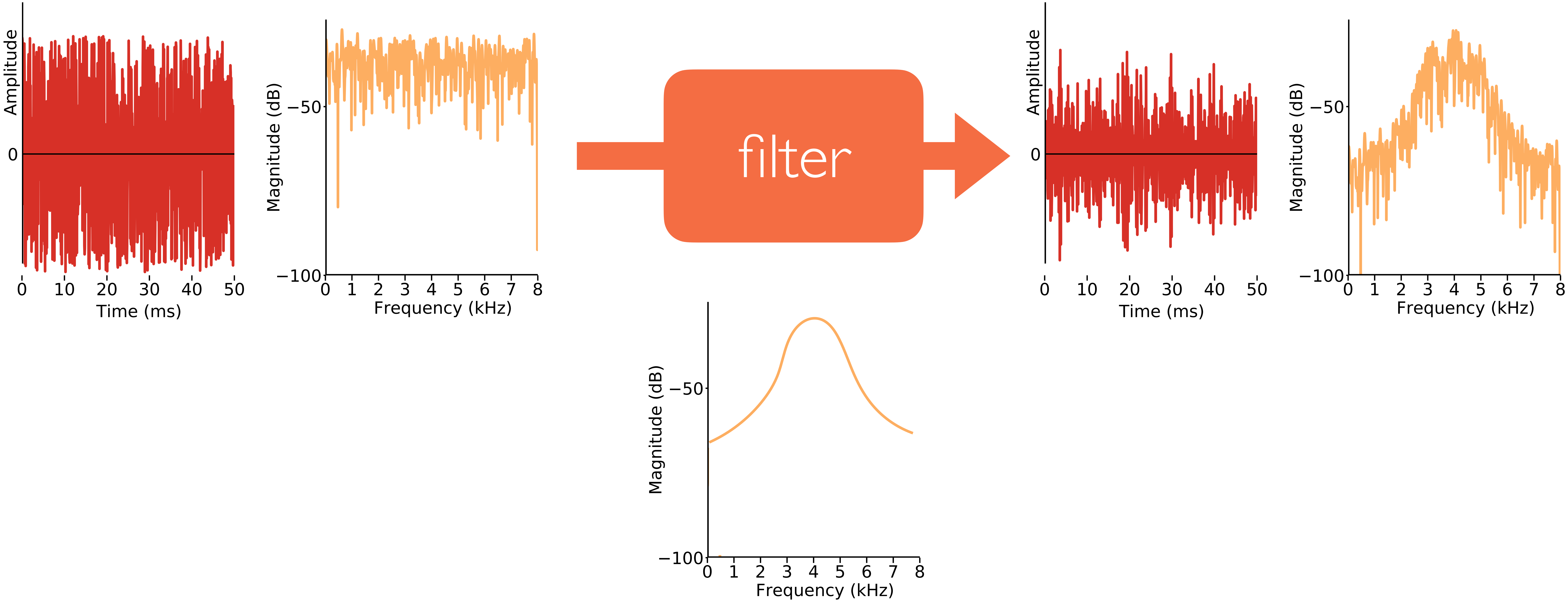
Making speech!

[s]

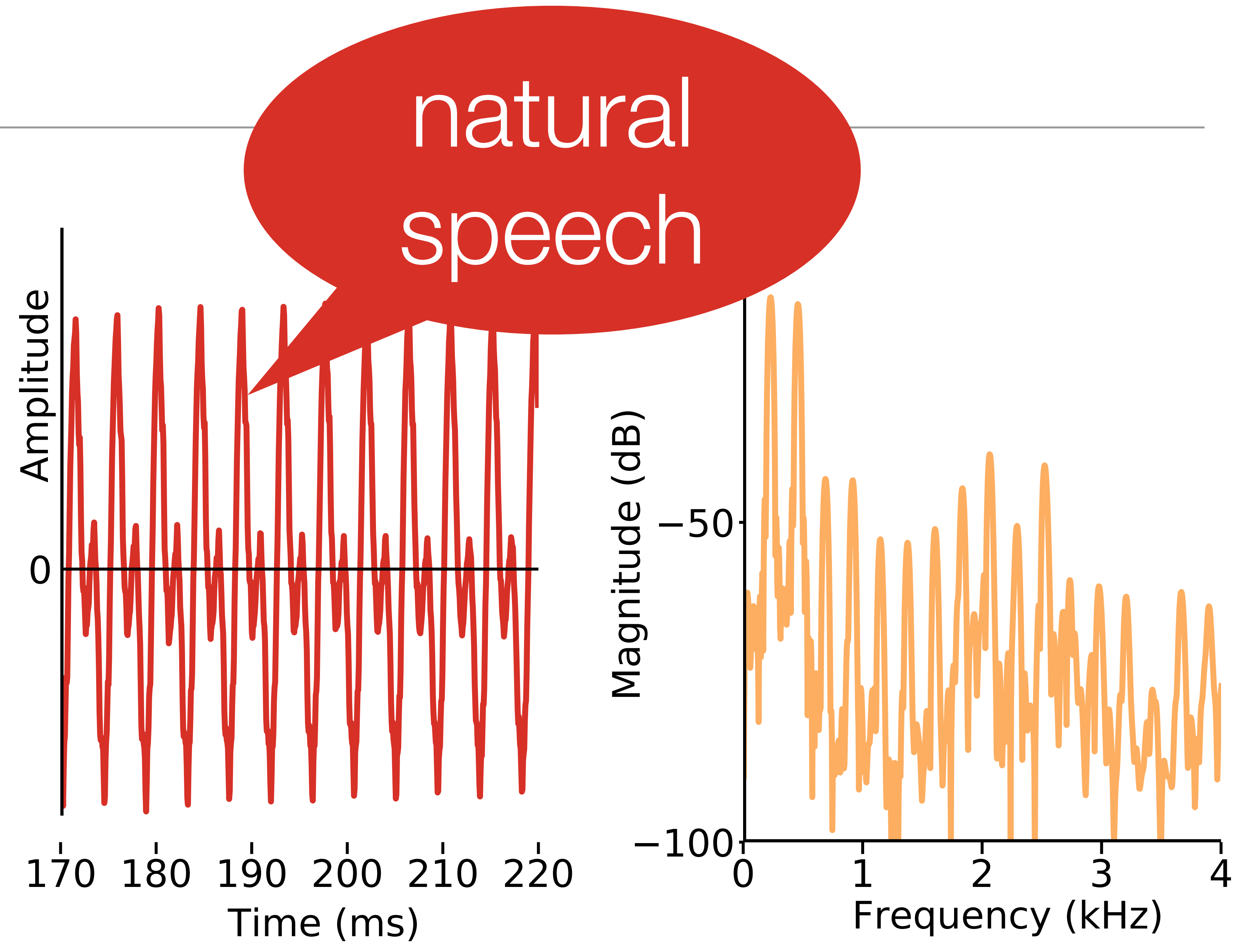


Making speech!

[ʃ]

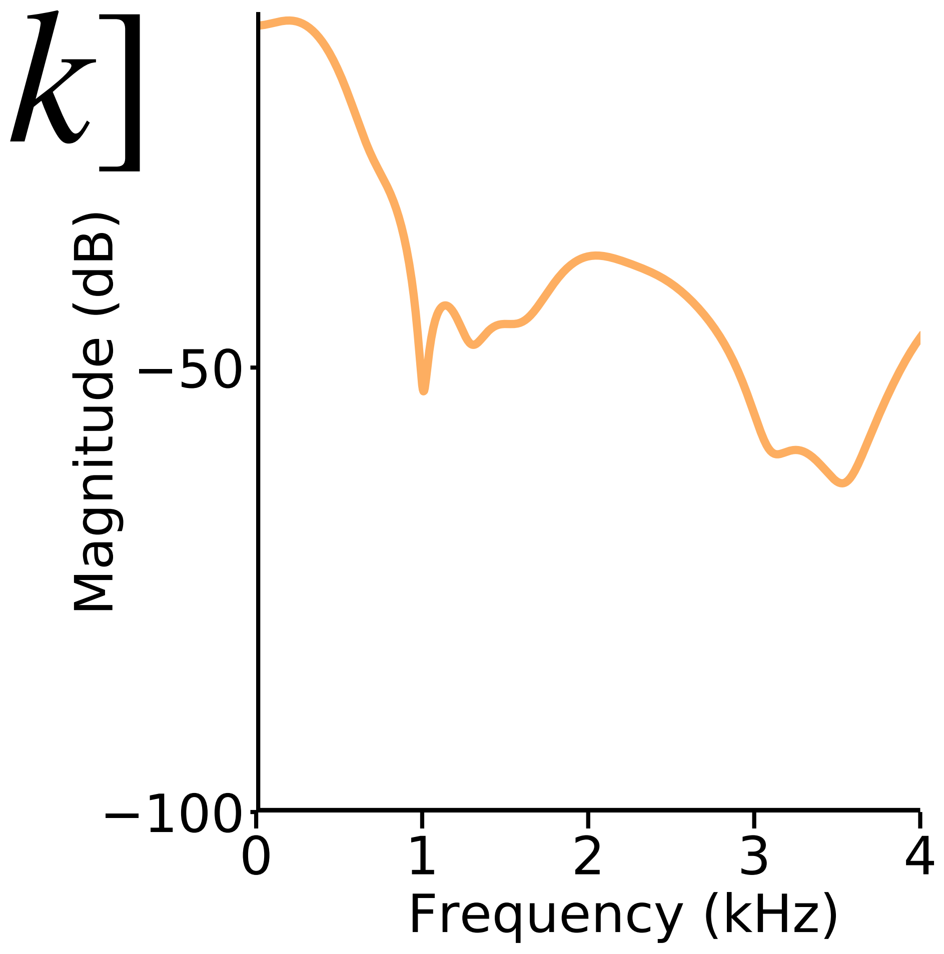


Fitting the model to a natural speech signal

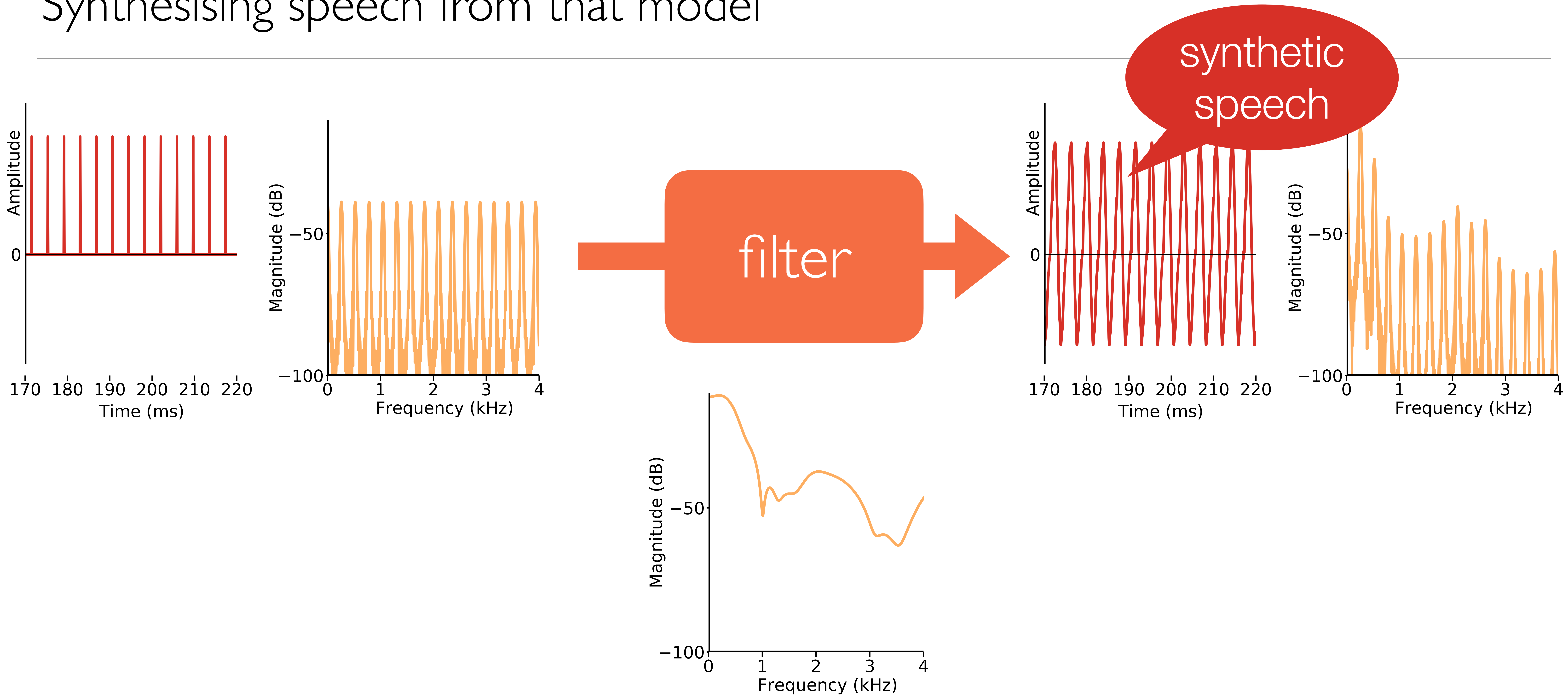


[u:]

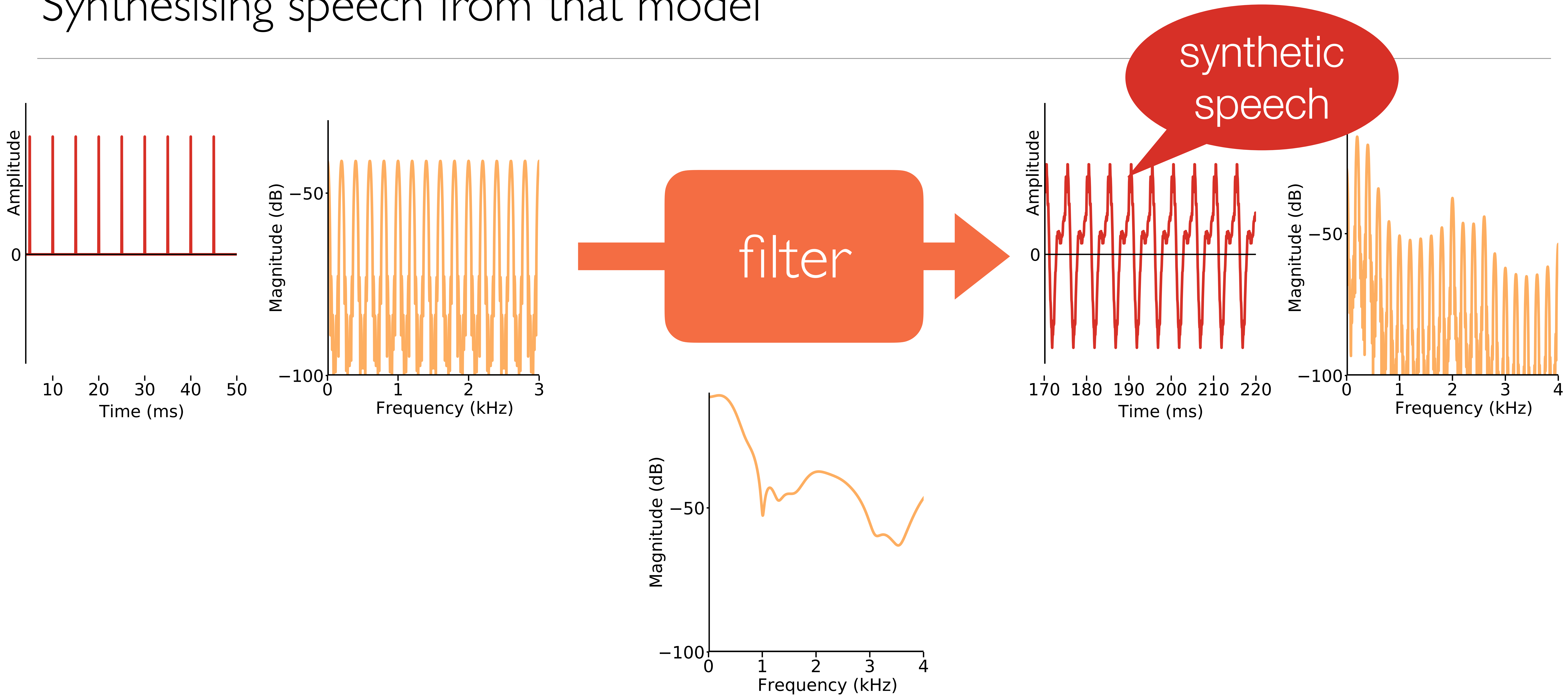
$$s[t] = e[t] + \sum_{k=1}^{p} a_k \, s[t-k]$$



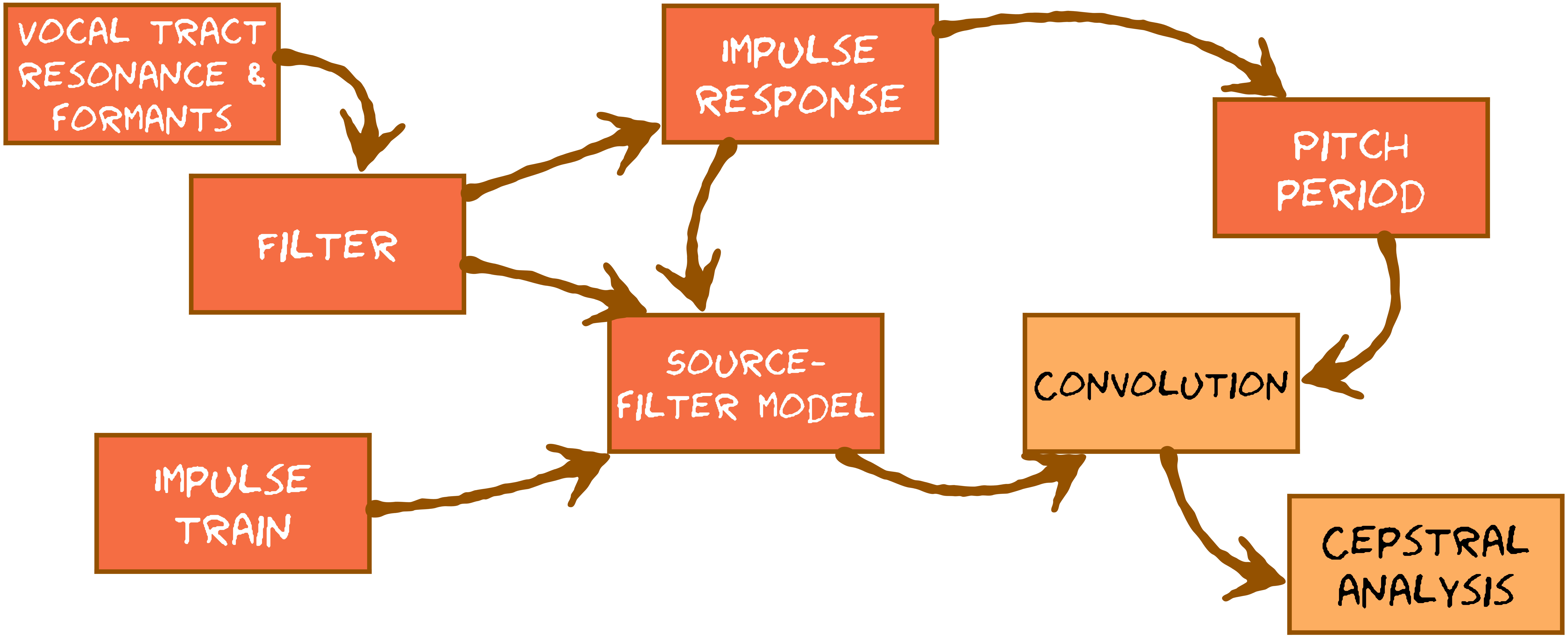
Synthesising speech from that model



Synthesising speech from that model

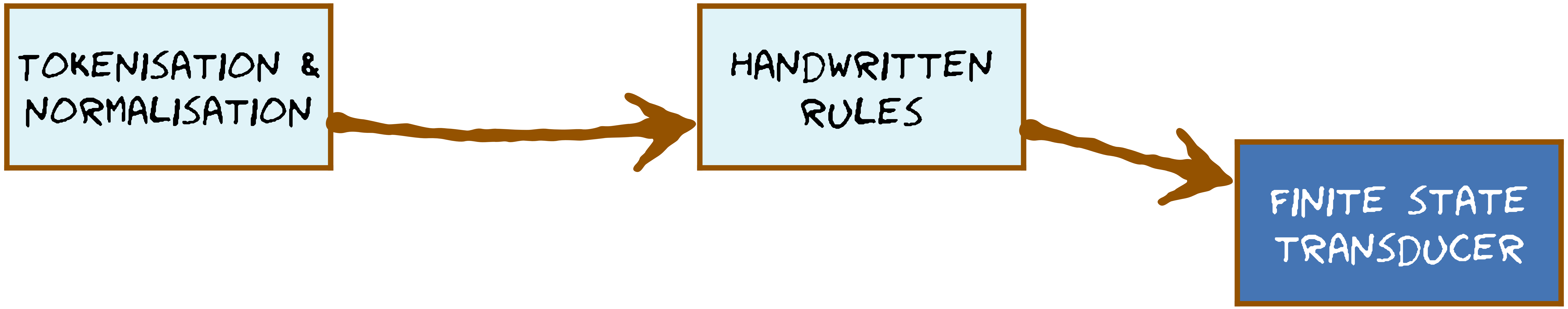


What you can learn next



Module 3

Front end : text processing


TOKENISATION & NORMALISATION
HANDWRITTEN RULES
FINITE STATE TRANSDUCER

TOKENISATION & NORMALISATION

INTERPRETABLE METHODS

What is the problem we are trying to solve?

```
He retired from business about
1790 with £10,000.
```

HE RETIRED FROM BUSINESS ABOUT
SEVENTEEN NINETY WITH TEN THOUSAND POUNDS

What is the problem we are trying to solve?

```
This should be 14 inches long and
3" by 3" inside, made of hard
wood ¾" thick.
```

THIS SHOULD BE FOURTEEN INCHES LONG AND
THREE INCHES BY THREE INCHES INSIDE MADE OF HARD
WOOD THREE QUARTERS OF AN INCH THICK

How hard can it be?

```
It was almost a matter of course that
Dr. Johnson, on arriving in Edinburgh,
August 17, 1773, should have come to
the White Horse, which was then kept
by a person of the name of Boyd.
```

Sentence splitting

```
To keep some command on our direction
required hard and diligent plying of
the paddle. The river was in such a
hurry for the sea! Every drop of water
ran in a panic, like as many people in
a frightened crowd. But what crowd was
ever so numerous, or so single-minded?
```

Sentence splitting

```
Edinburgh was, at the beginning of
George III.’s reign, a picturesque,
odorous, inconvenient, old-fashioned
town, of about seventy thousand
inhabitants.
```

Tokenisation

```
This should be 14 inches long and
3" by 3" inside, made of hard
wood ¾" thick.
```

Text analysis

```
It was almost a matter of course that
Dr. Johnson, on arriving in Edinburgh,
August 17, 1773, should have come to
the White Horse, which was then kept
by a person of the name of Boyd.
```

Text analysis

```
It was almost a matter of course that
Dr. Johnson, on arriving in Edinburgh,
August 17, 1773, should have come to
the White Horse, which was then kept
by a person of the name of Boyd.
```

Ambiguous written form: homographs

abbreviation	`Dr` , `St` , `m`
accidental	`polish` , `does` , `sow`
part-of-speech, or word sense	`record` , `read` , `bass`

Key steps in tokenisation and normalisation

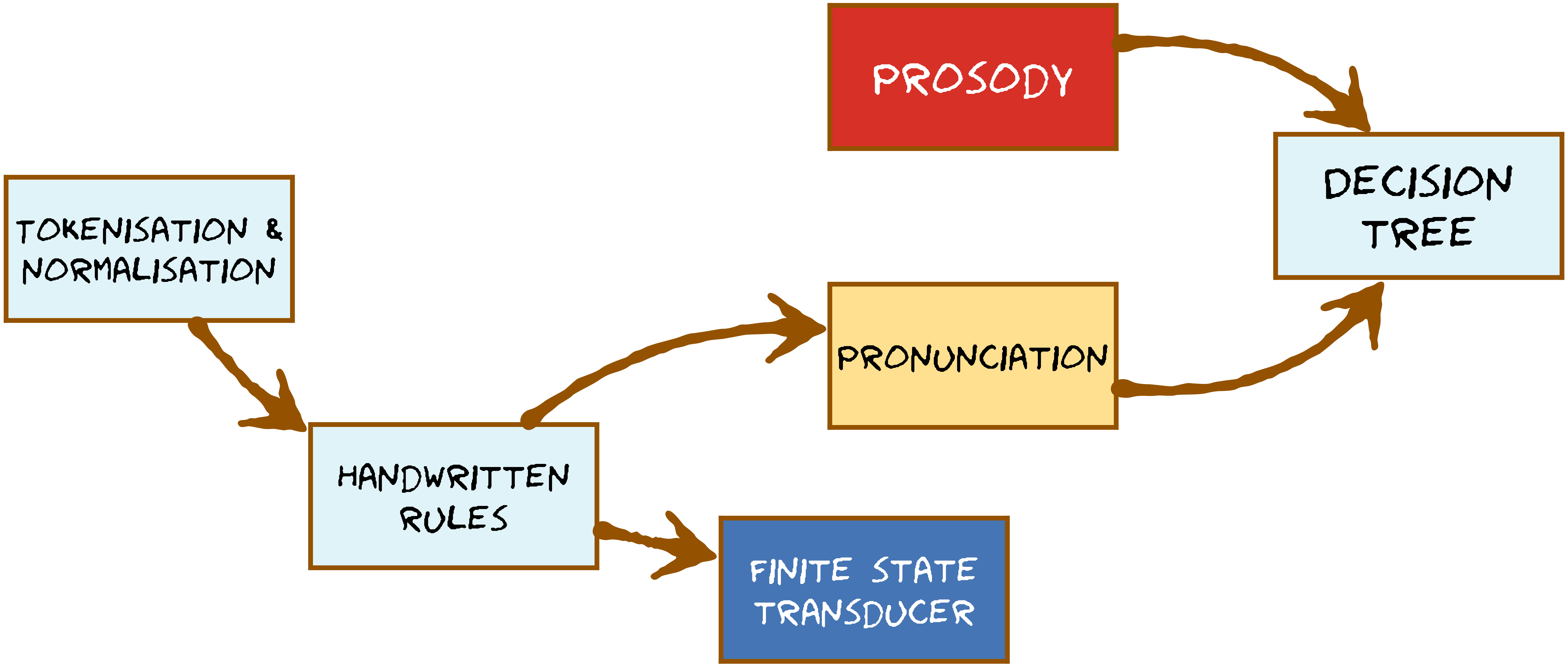
- **Tokenise** the input character sequence, then for each token:

- **Classify** as either
  - natural language
  - Non-Standard Word (NSW) : abbreviation, cardinal number, year, date, money, ...

- **Resolve ambiguity** and find the underlying form

- **Verbalise** NSWs into natural language

What you can learn next



HANDWRITTEN RULES

Example: tokenisation by rule

```
input = “He retired about 1790 with £10,000.”
tokens = []
i = 0
for j in range(len(input)):
    if input[j] == " ":
        tokens.append(input[ i : j ])
        i=j
tokens.append(input[ i : ])
```

Example: disambiguating `Dr.` using context-sensitive rewrite rules

`...that Dr. Johnson, on arriving...`

`...turn into Burns Dr. then...`

Dr. → [Capitalised word] / **Drive** / [anything]

Dr. → [anything] / **Doctor** / [Capitalised word]

Example: word-sense disambiguation using a collocation rule

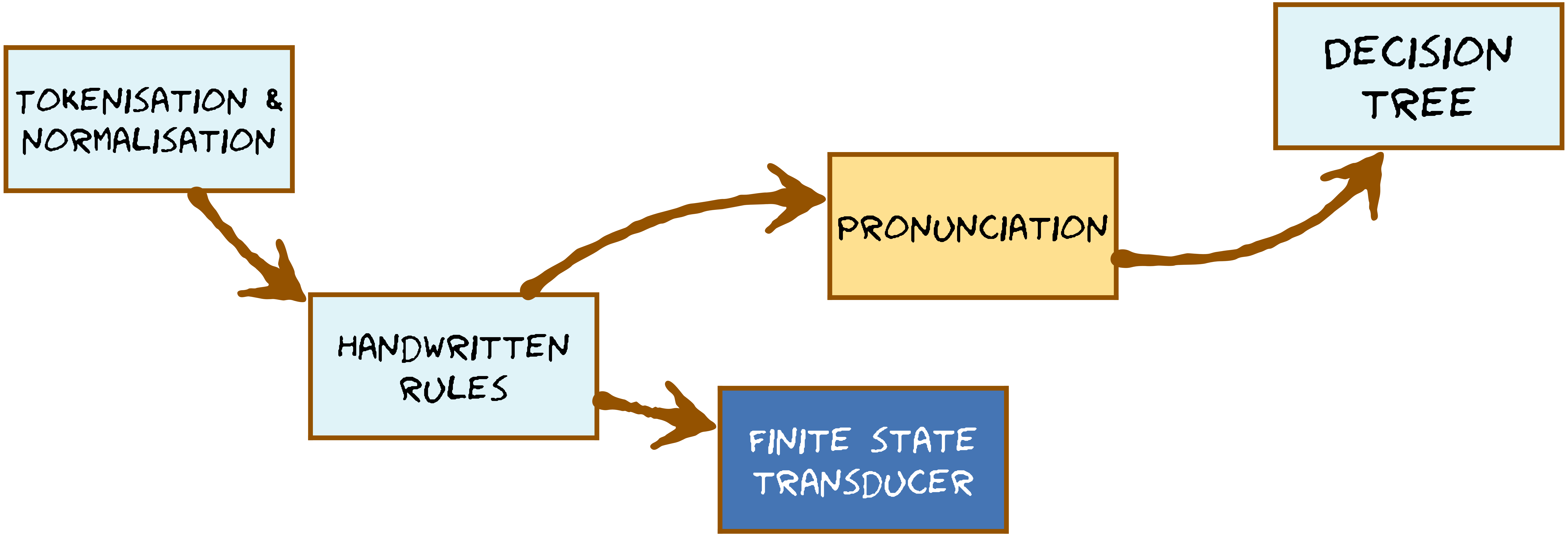
```
...I caught a large bass yesterday...

...the bass player is...
```

$\texttt{bass} \rightarrow \underset{\text{BASS-FISH}}{\texttt{bass}} \mid \texttt{[caught,river,fish,..]}$

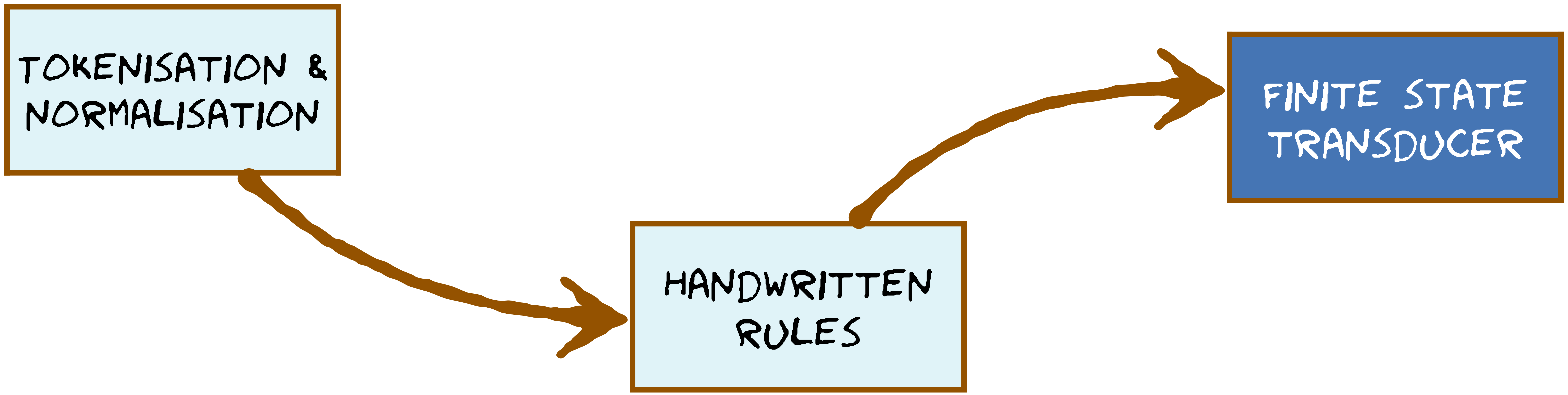
$\texttt{bass} \rightarrow \underset{\text{BASS-MUSIC}}{\texttt{bass}} \mid \texttt{[player,band,guitar,..]}$

What you can learn next



FINITE STATE TRANSDUCER

What you need to know already

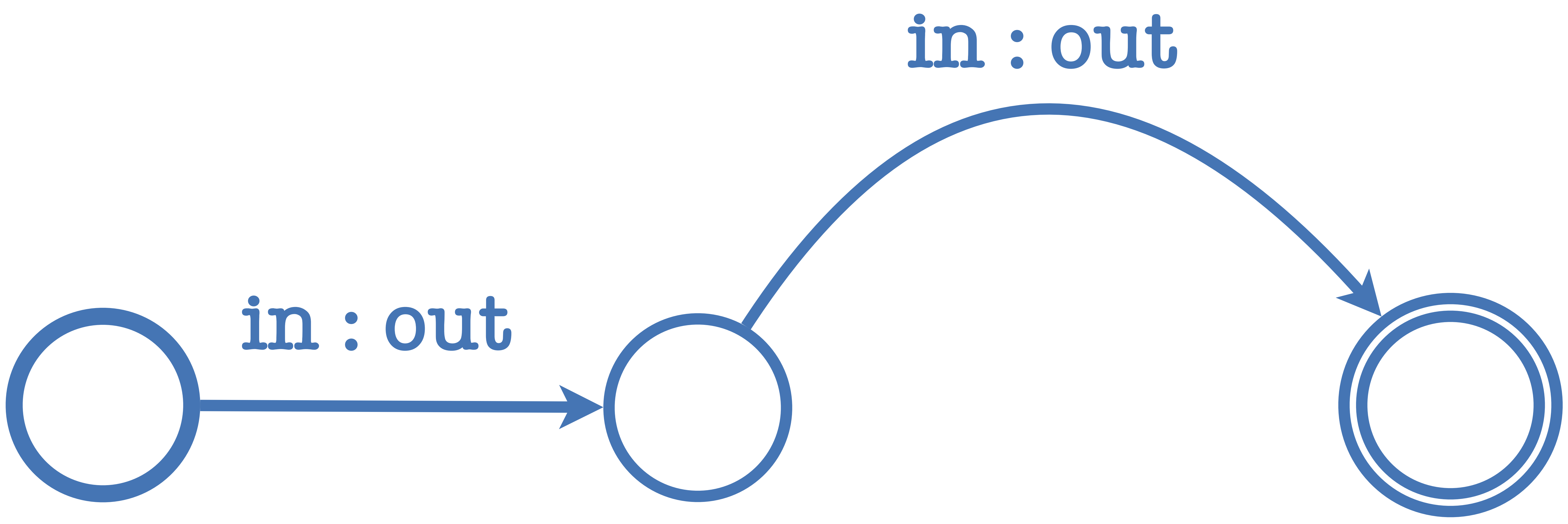


What is the task we are performing?

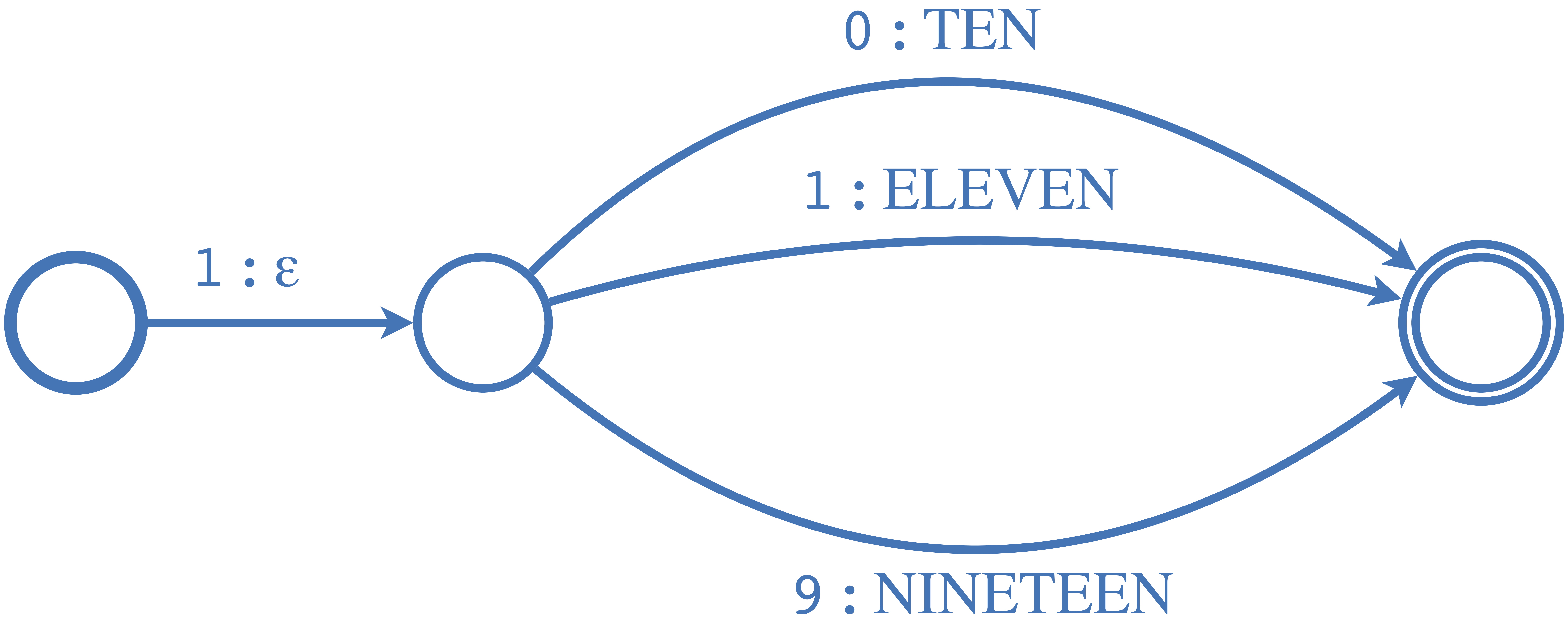
```
He retired from business about
1790 with £10,000.
```

HE RETIRED FROM BUSINESS ABOUT
SEVENTEEN NINETY WITH TEN THOUSAND POUNDS

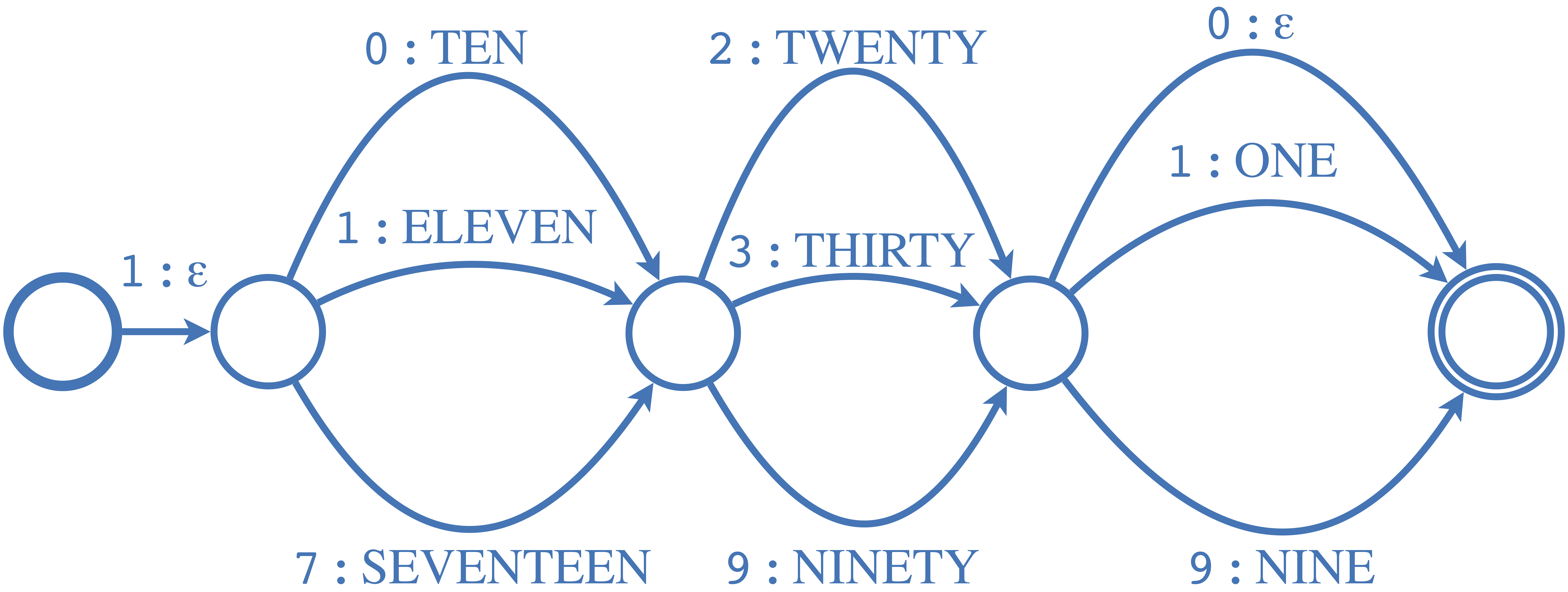
Finite State Transducer (FST)



Verbalising the numbers `10` to `19`

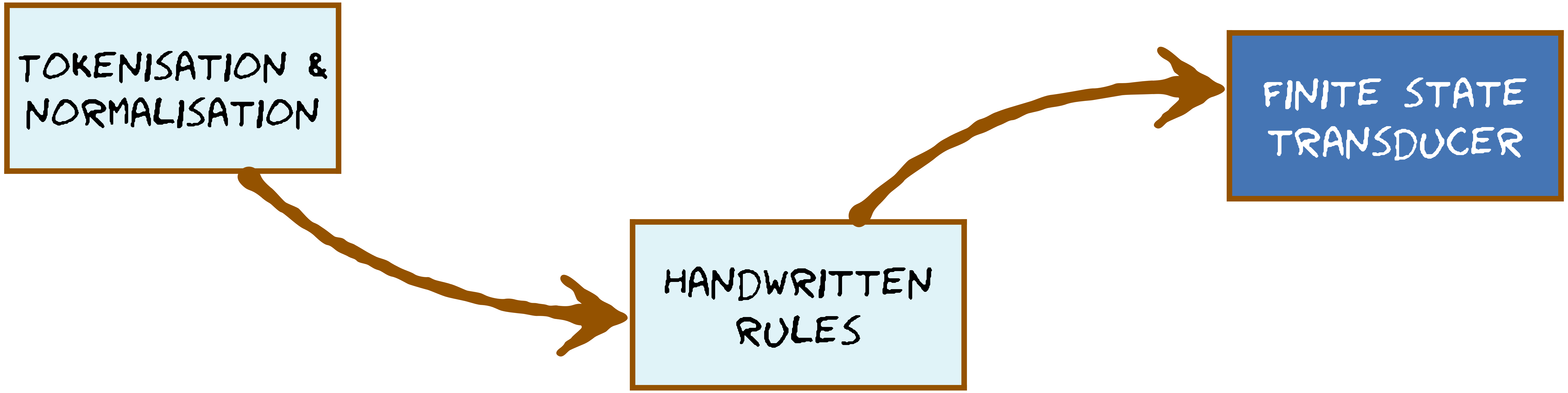


Verbalising years with 4 digits, such as `1790`



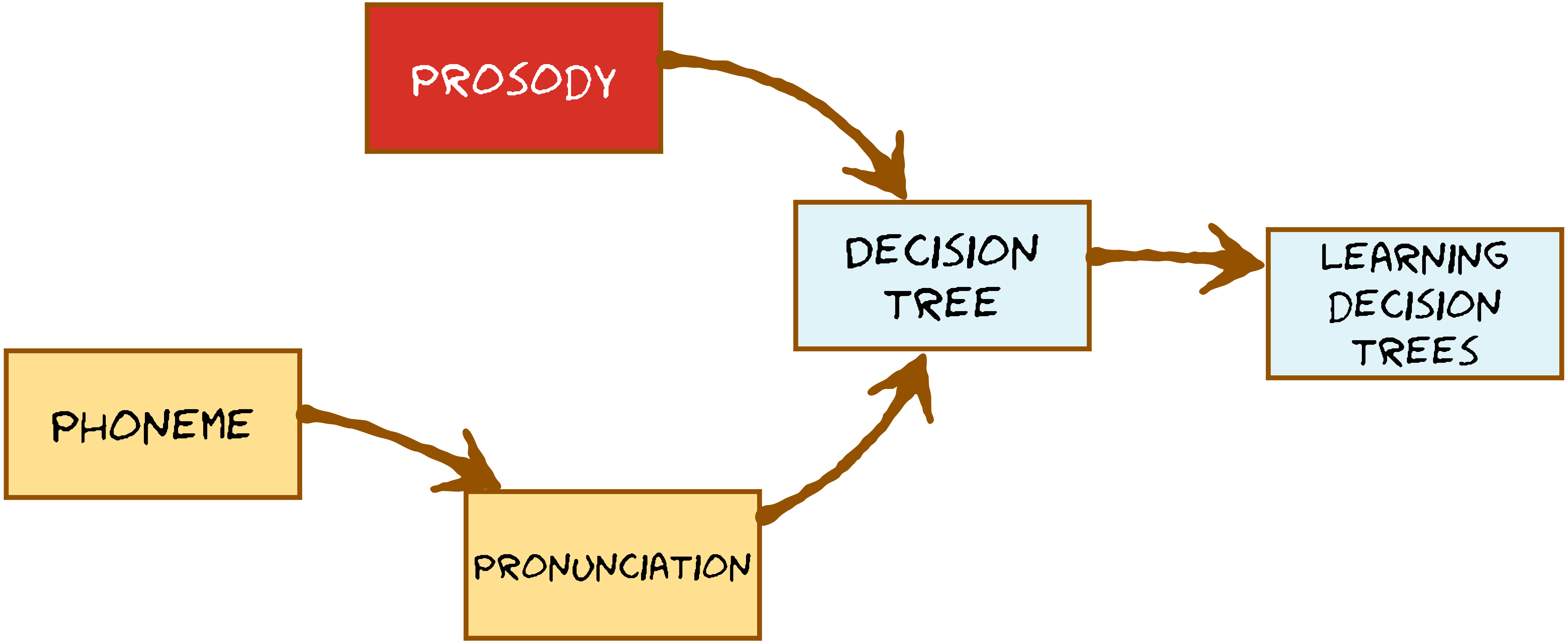
Verbalising money amounts, such as `£10,000`

What you can learn next



Module 4

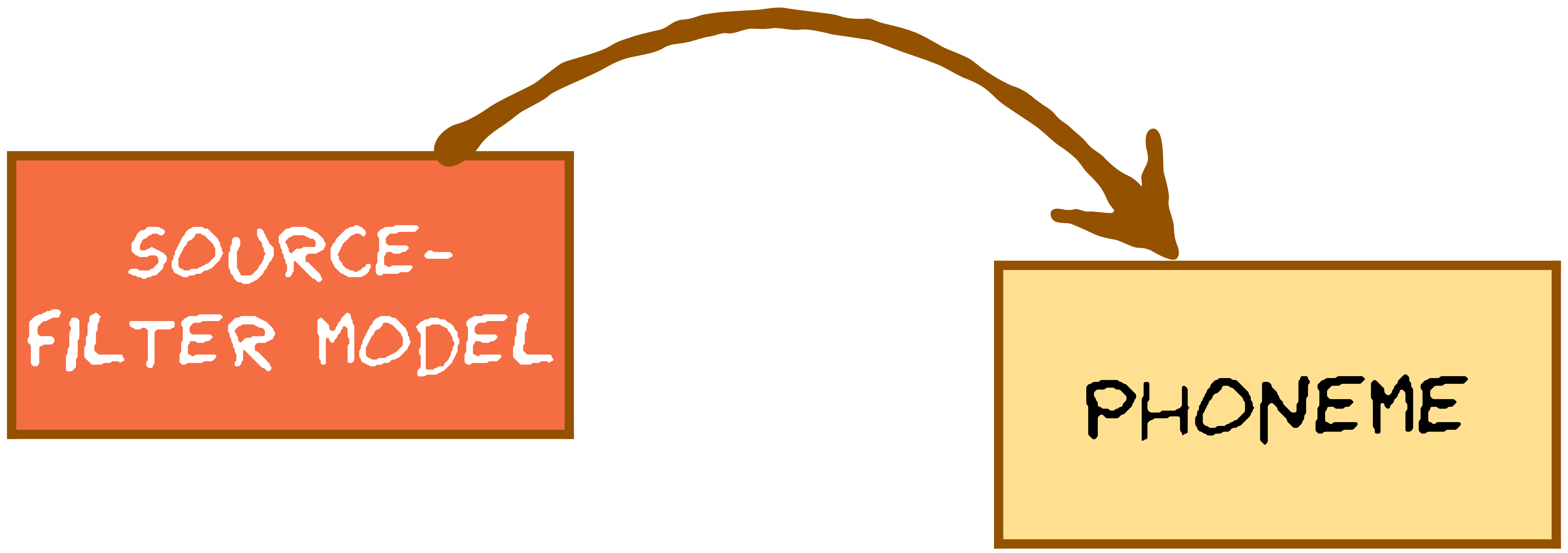
Front end : pronunciation & prosody


PROSODY
DECISION TREE
LEARNING DECISION TREES
PHONEME
PRONUNCIATION

PHONEME

SOUND CATEGORIES

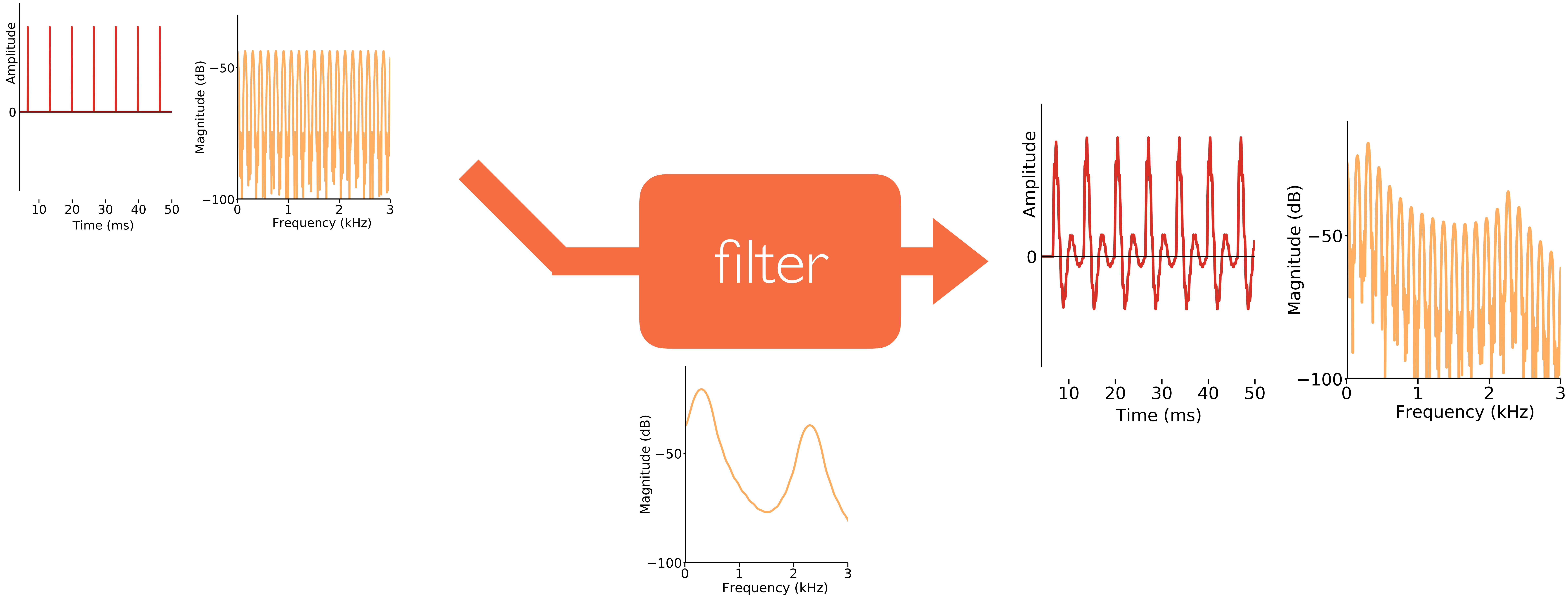
What you need to know already

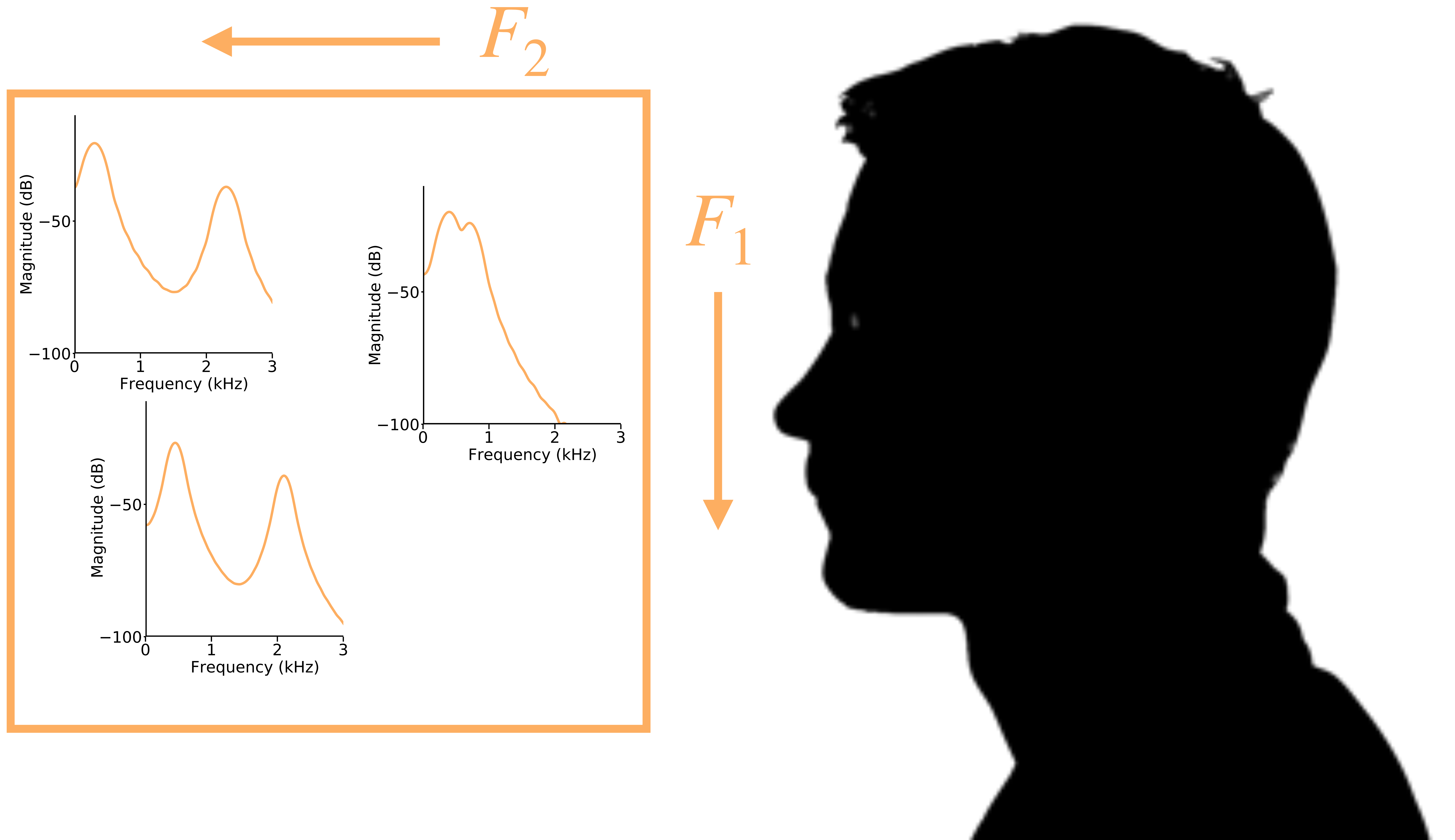


Acknowledgement for the IPA chart

What can a speaker control,
to encode the message to the listener?

Using the source-filter model to explain vowels

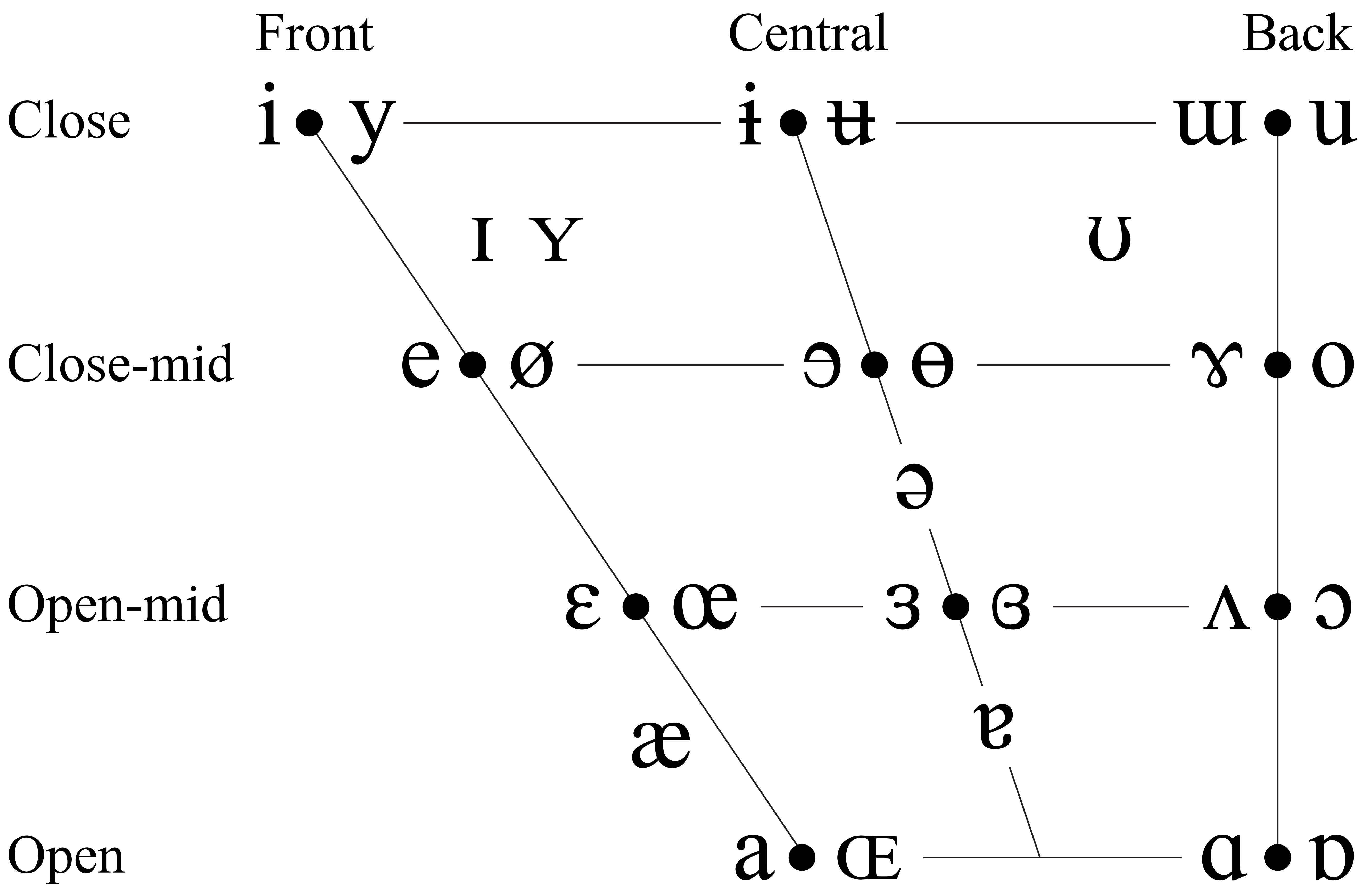




F_2
F_1
Magnitude (dB)
−50
−100
0
1
2
3
Frequency (kHz)
Magnitude (dB)
−50
−100
0
1
2
3
Frequency (kHz)
Magnitude (dB)
−50
−100
0
1
2
3
Frequency (kHz)

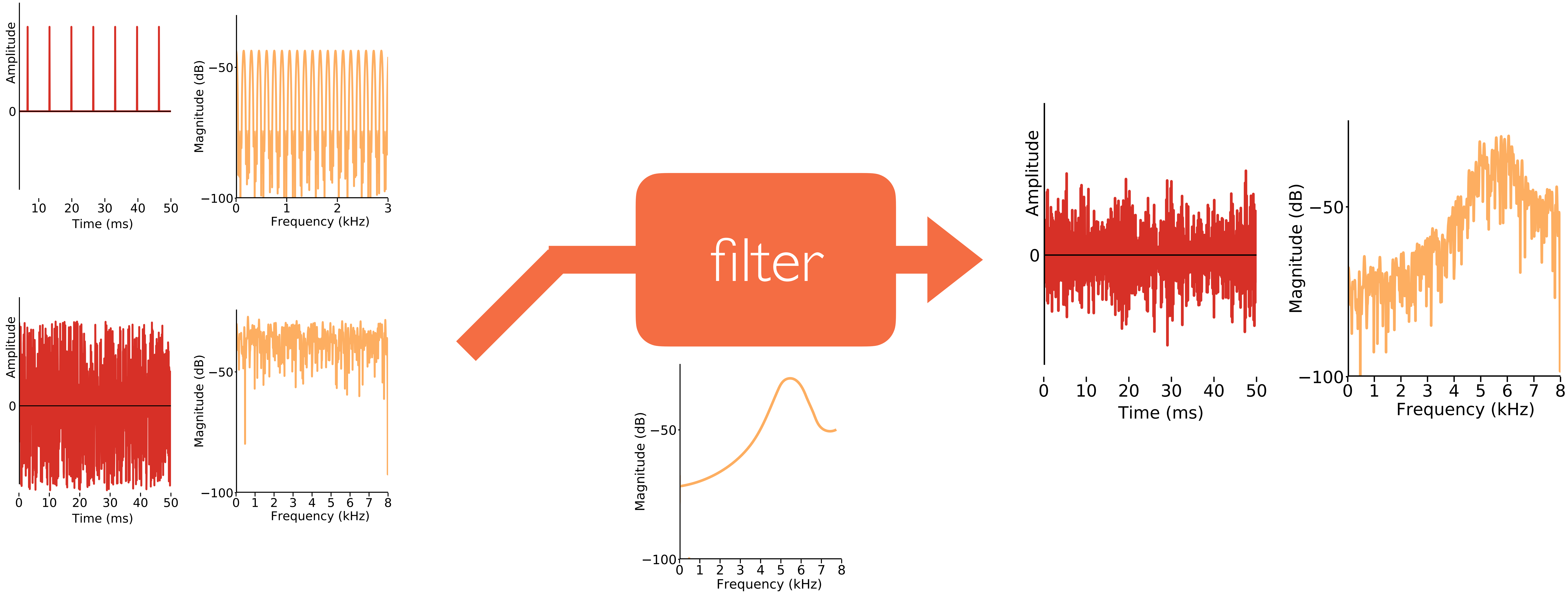

Front
Central
Back
Close
i y
ɨ ʉ
ɯ u
ɪ ʏ
ʊ
Close-mid
e ø
ɘ ɵ
ɤ o
ə
Open-mid
ɛ œ
ɜ ɞ
ʌ ɔ
æ
ɐ
Open
a ɶ
ɑ ɒ
Where symbols appear in pairs, the one
to the right represents a rounded vowel.

IPA vowel chart



Where symbols appear in pairs, the one
to the right represents a rounded vowel.

Using the source-filter model to explain fricatives

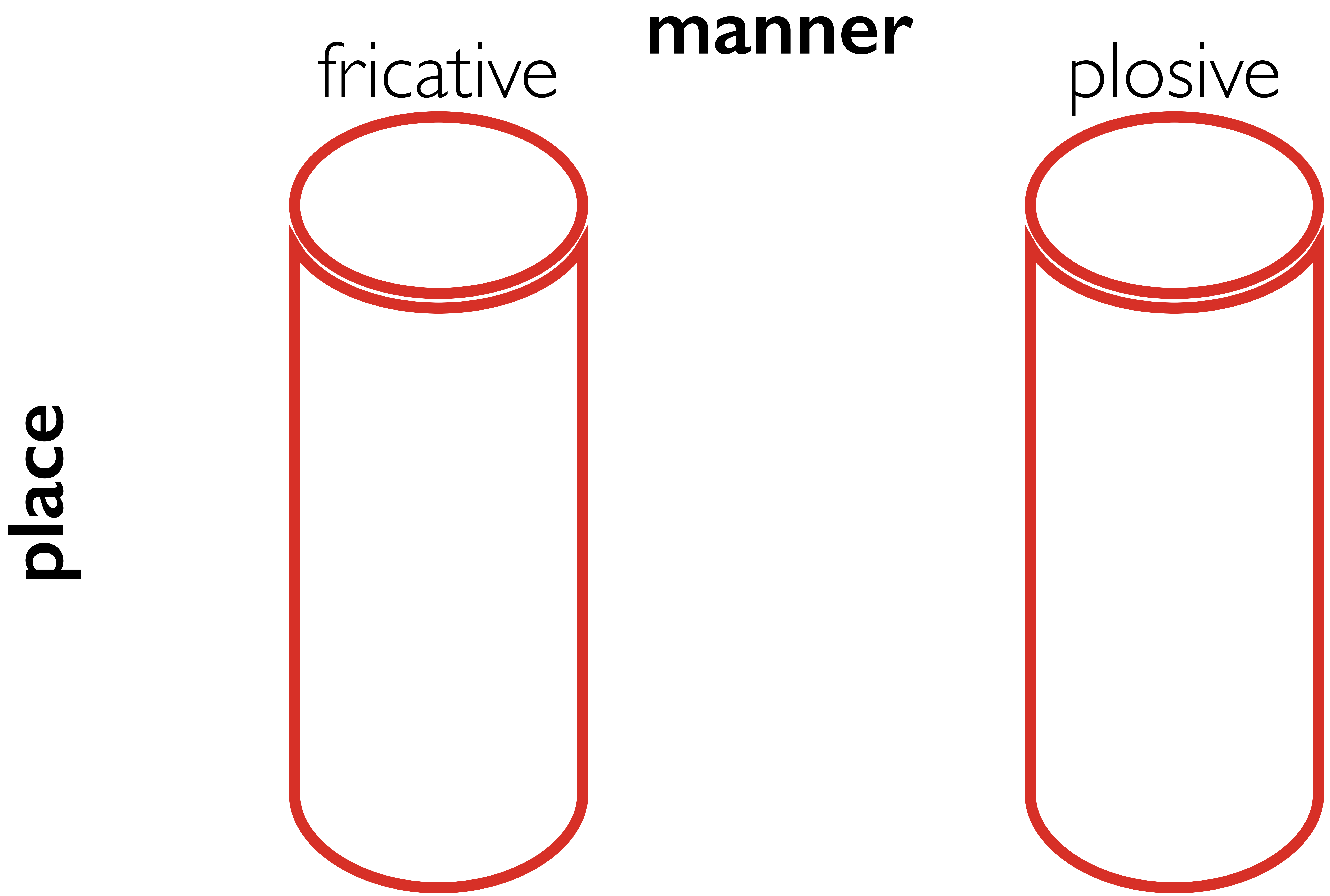


Lots of fricatives

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

Place and manner



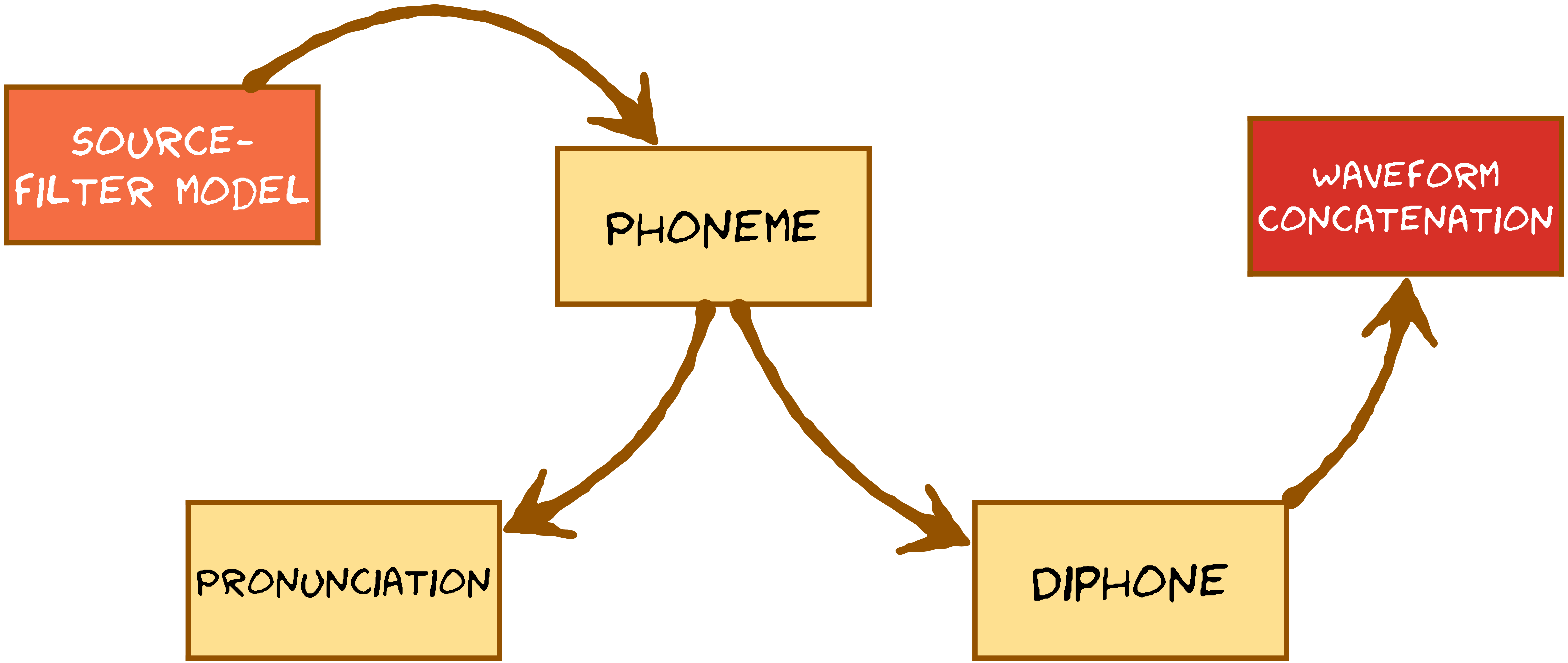
IPA consonant chart

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

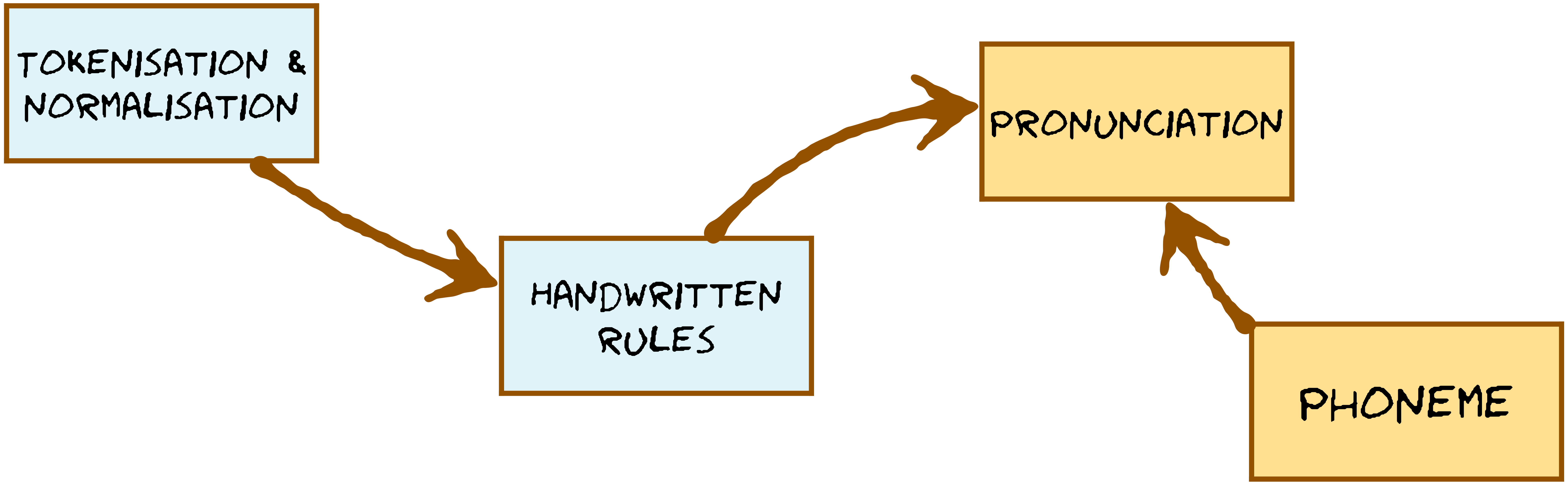

THIS VIDEO
A COURSE ON PHONETICS

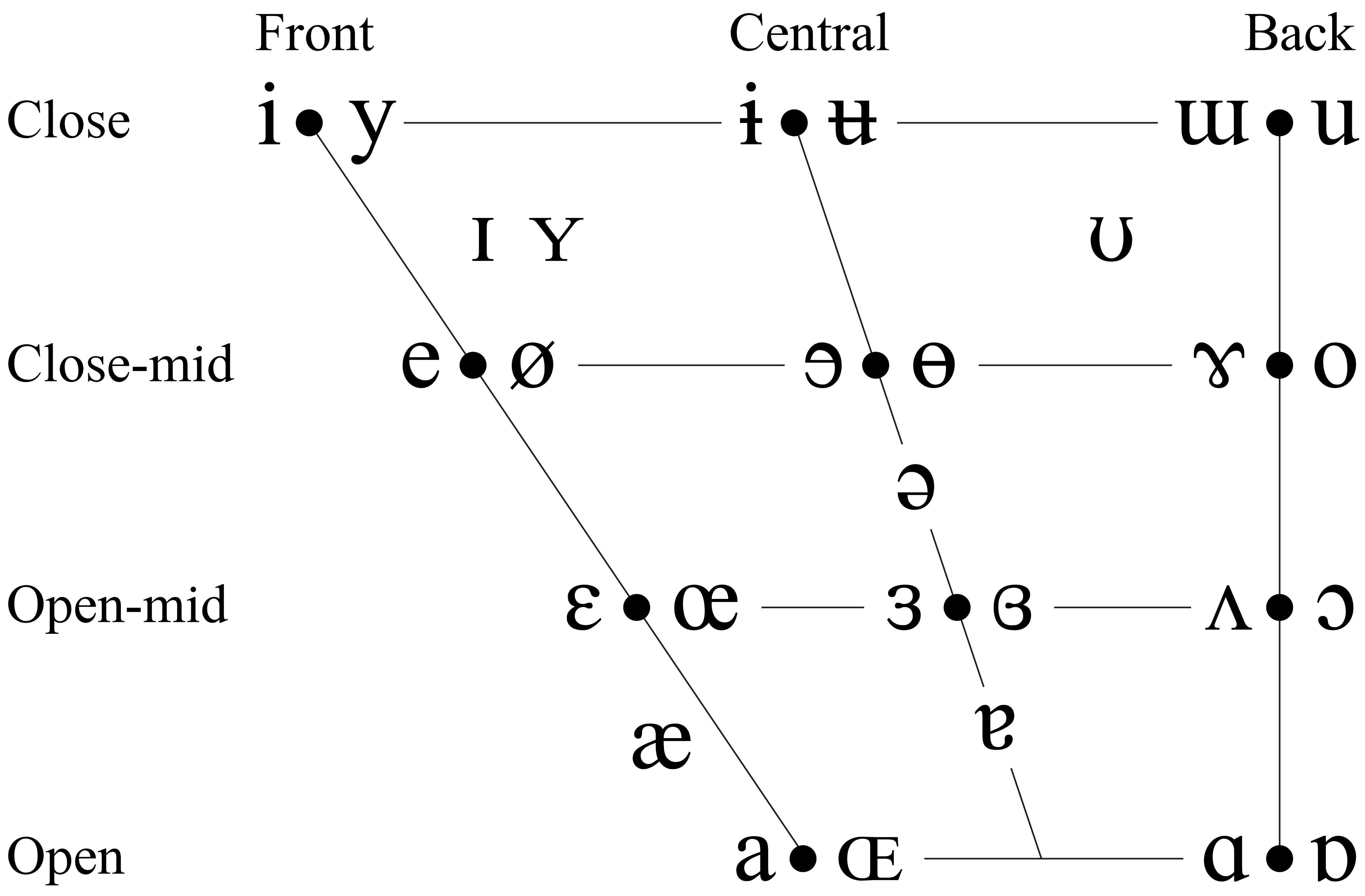
What you can learn next



PRONUNCIATION

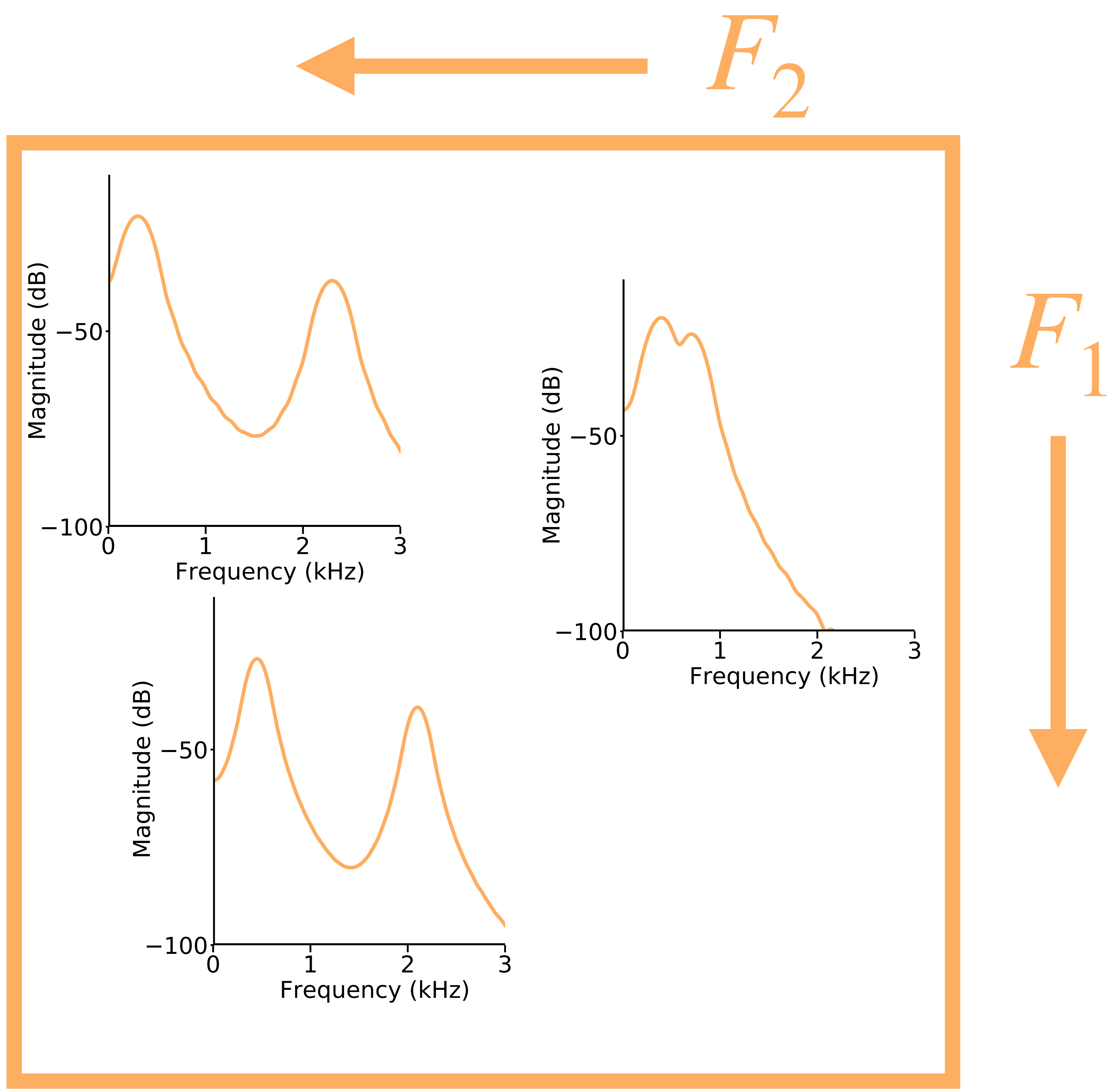
What you need to know already

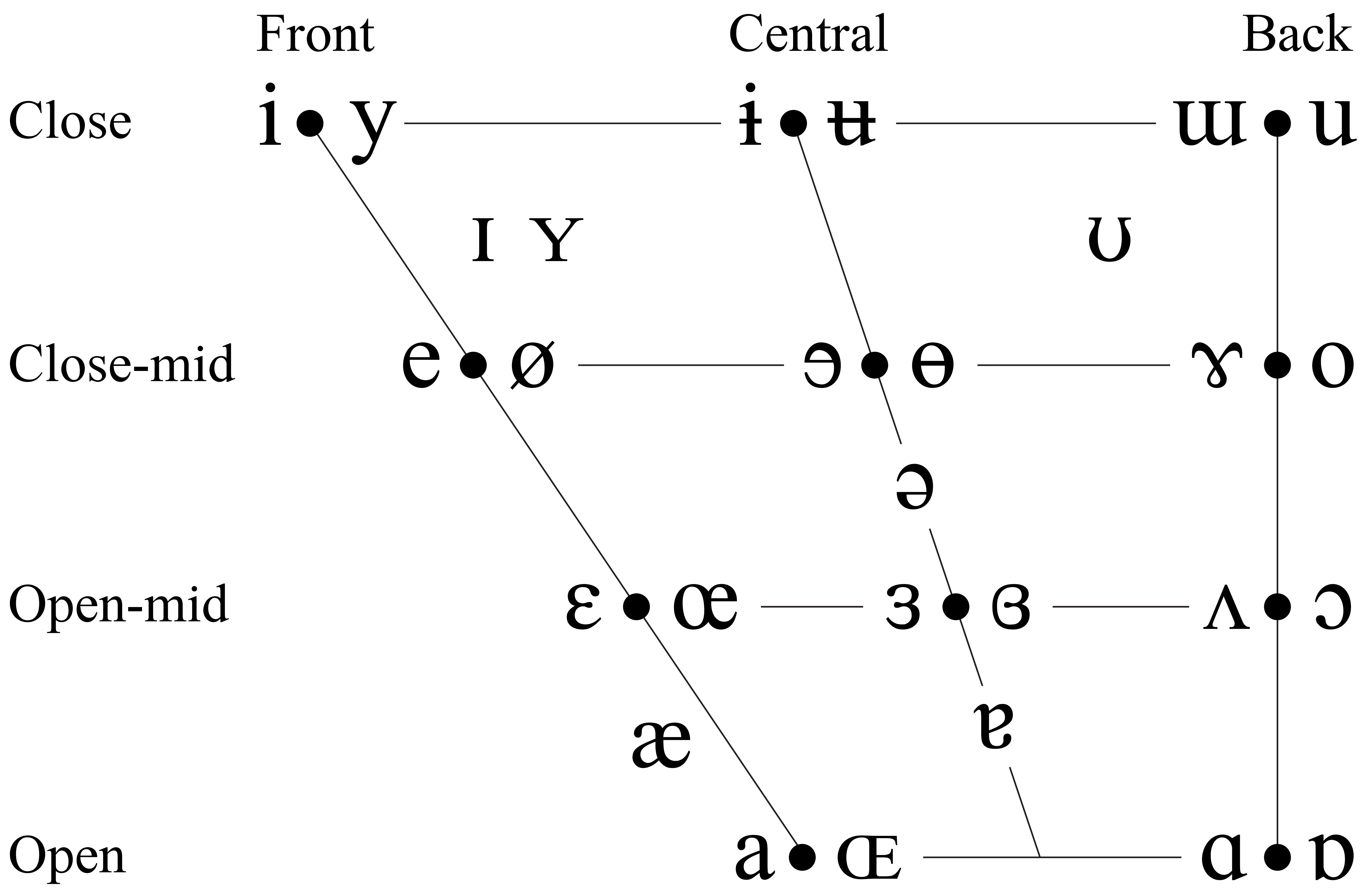




Where symbols appear in pairs, the one
to the right represents a rounded vowel.

What exactly *is* a phoneme?



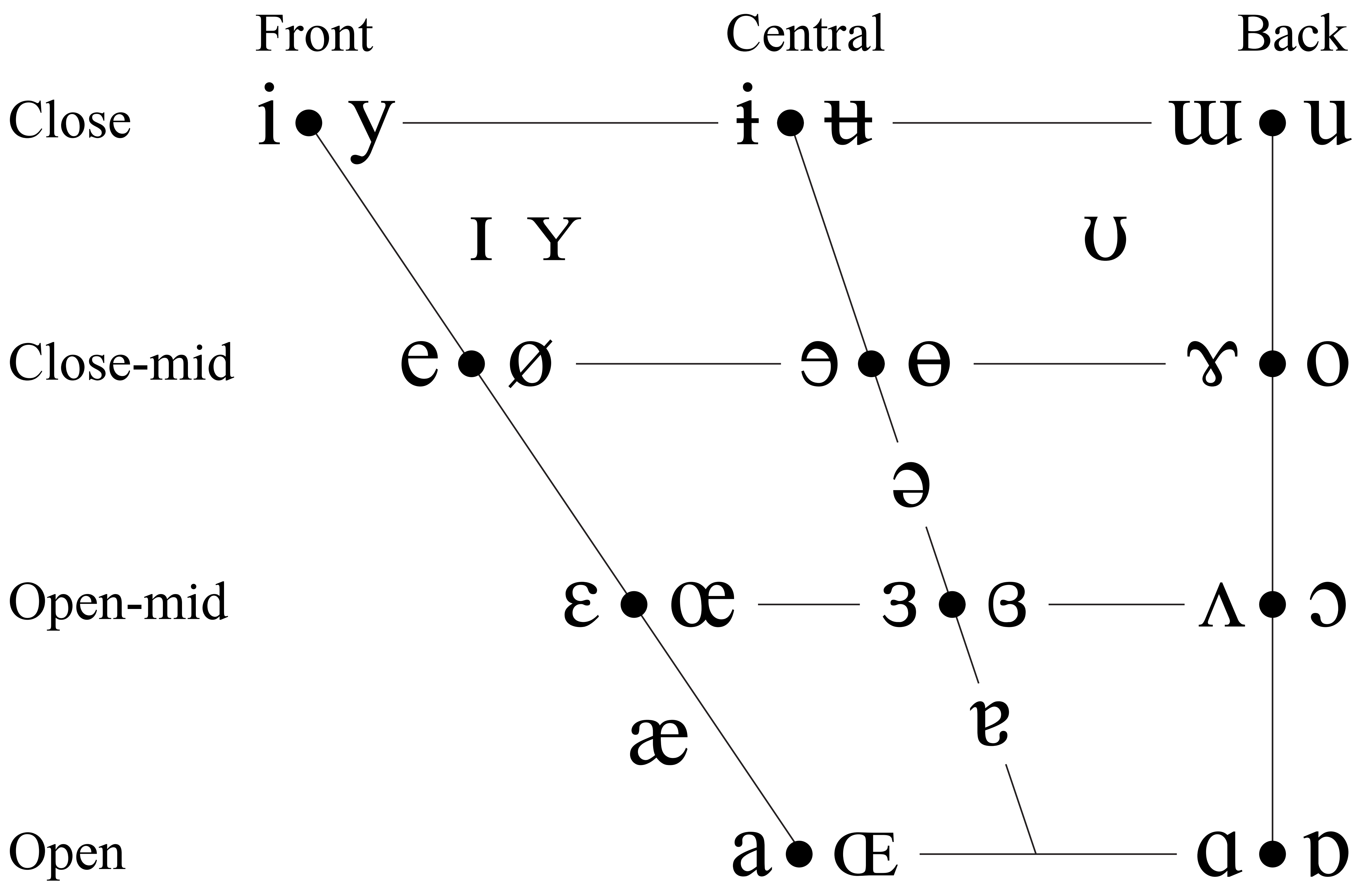


Where symbols appear in pairs, the one to the right represents a rounded vowel.

Minimal pairs

/bɪt/ - /bet/

/bɪt/ - /biːt/



Where symbols appear in pairs, the one
to the right represents a rounded vowel.

Minimal pairs

/bɪt/ - /pɪt/

/bɪt/ - /dɪt/

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar
Plosive	p b			t d	
Nasal	m	ɱ		n	
Trill	ʙ			r	
Tap or Flap		ⱱ		ɾ	
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ
Lateral fricative				ɬ ɮ	
Approximant		ʋ		ɹ	
Lateral approximant				l	

Symbols to the right in a cell are voiced, to the left are voic

Allophones

long

pull

letter

lull

/lʊl/ /lʊɫ/


THIS VIDEO
A COURSE ON PHONOLOGY

Finding the pronunciation of a word : grapheme-to-phoneme (G2P) rules

Castillian Spanish

```
[ a ] = / a /
[ e ] = / e /
[ i ] = / i /
[ c ] i = / θ /
[ b ] = / b /
[ v ] = / b /
```

Southern British English

```
[ o ] r e = / ɔː /
[ e ] e = / iː /
C [ i ] C = / ɪ /
[ c ] i = / s /
[ c ] l = / k /
[ c ] h = / tʃ /
```

Finding the pronunciation of a word : dictionary + G2P

[o] r e = / ɔː /

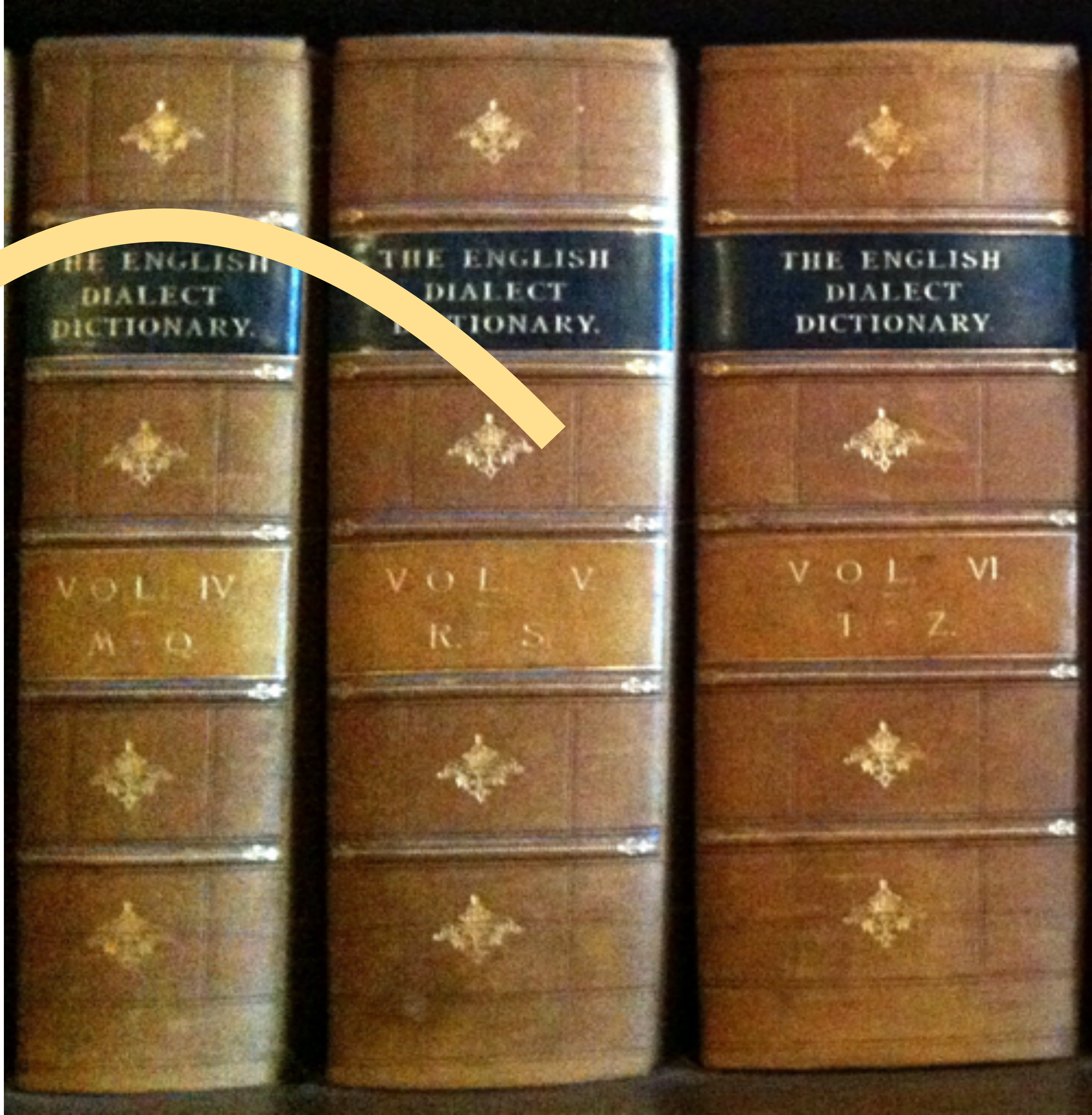
[e] e = / iː /

C [i] C = / ɪ /

[c] i = / s /

[c] l = / k /

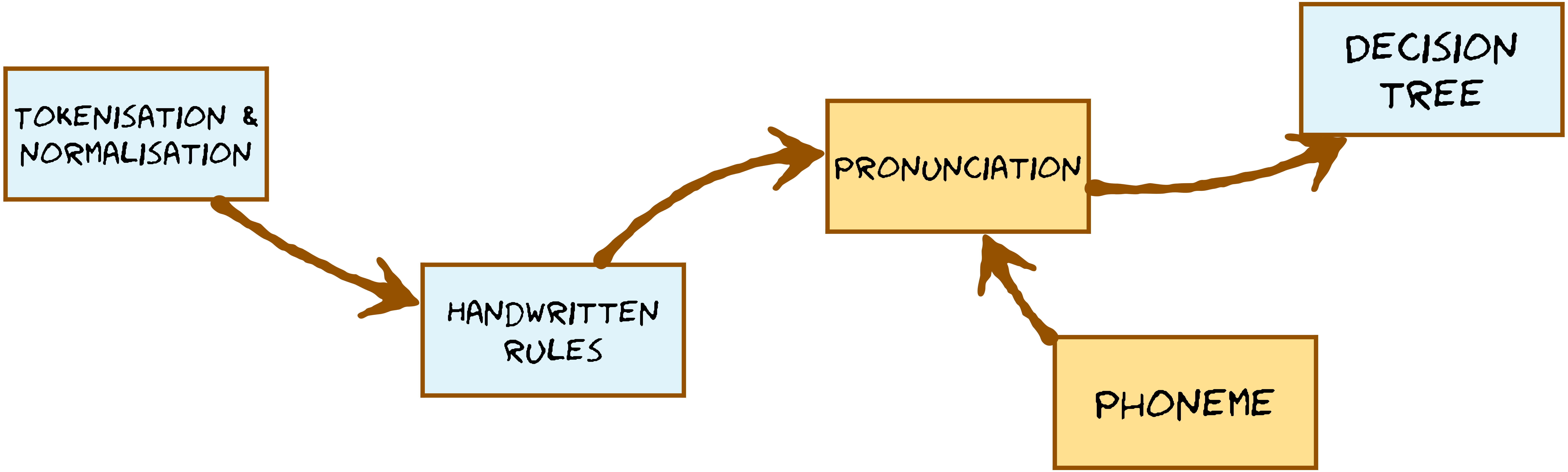
[c] h = / tʃ /



Pronunciation dictionary

`impossible`	[ɪ m p ɒ s ə b ə l]
`impossible`	ih m p oh s ax b l
`impossible`	jj ((i m) 0) ((p o) 1) ((s i) 0) ((b l!) 0)

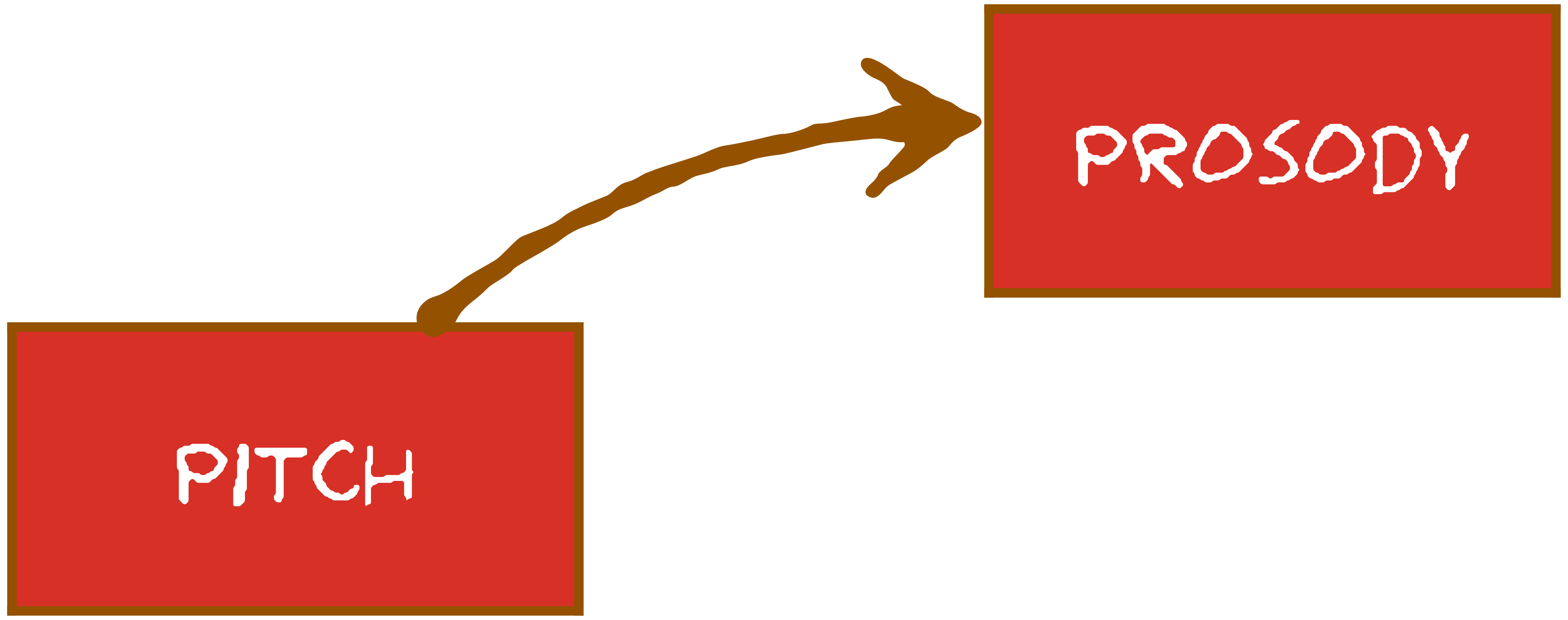
What you can learn next



PROSODY

PERIODIC SIGNALS IN THE TIME DOMAIN

What you need to know already



Nothing’s impossible.

n ʌ θ ɪ ŋ z ɪ m p ɒ s ə b ə l

‘Defining’ prosody

<u>Linguistic functions</u>

phrasing

rhythm

emphasis

intonation (‘tune’)

<u>Para-linguistic functions</u>

attitude

emotion

<u>Acoustic correlates</u>

F_0

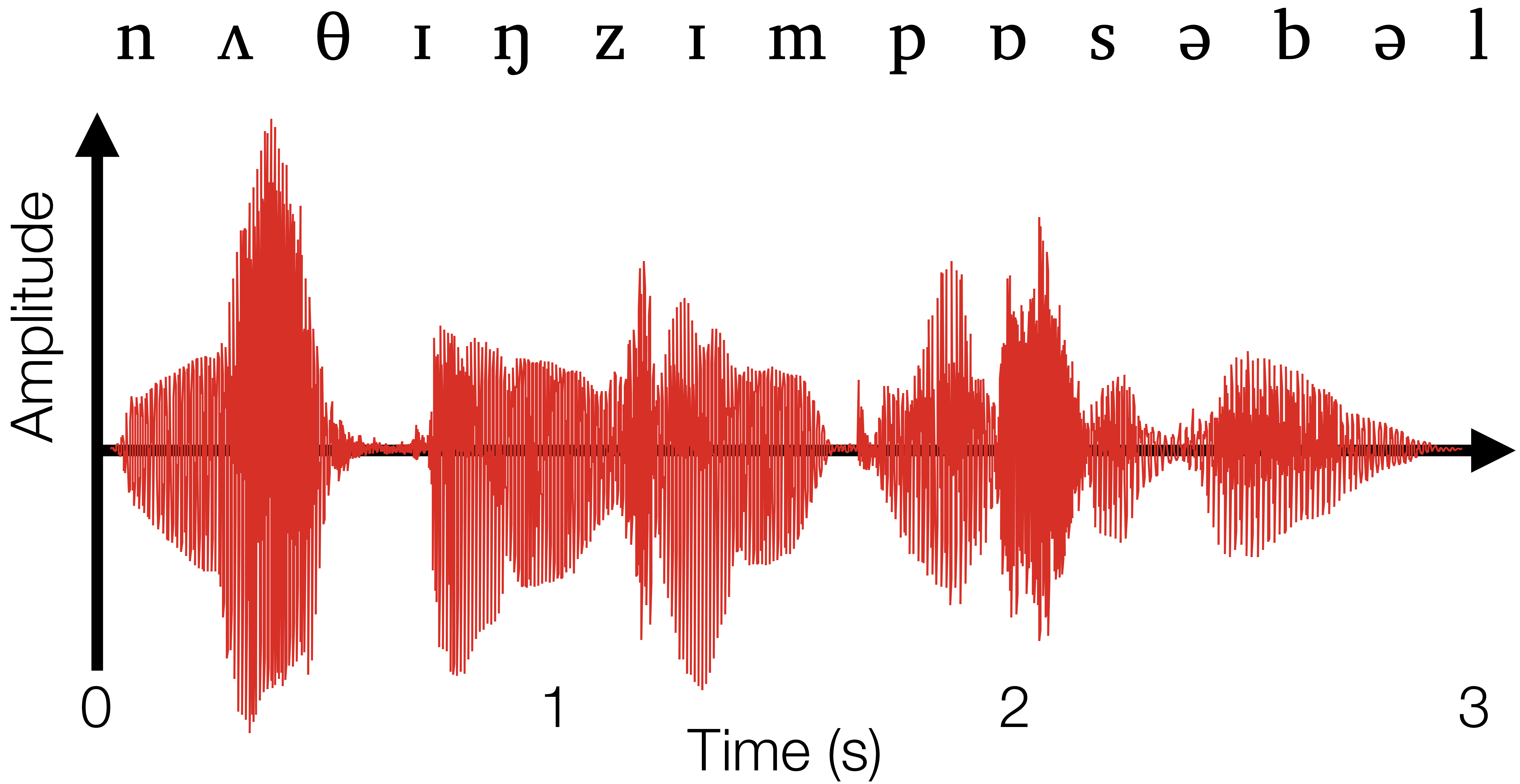
duration

voice quality

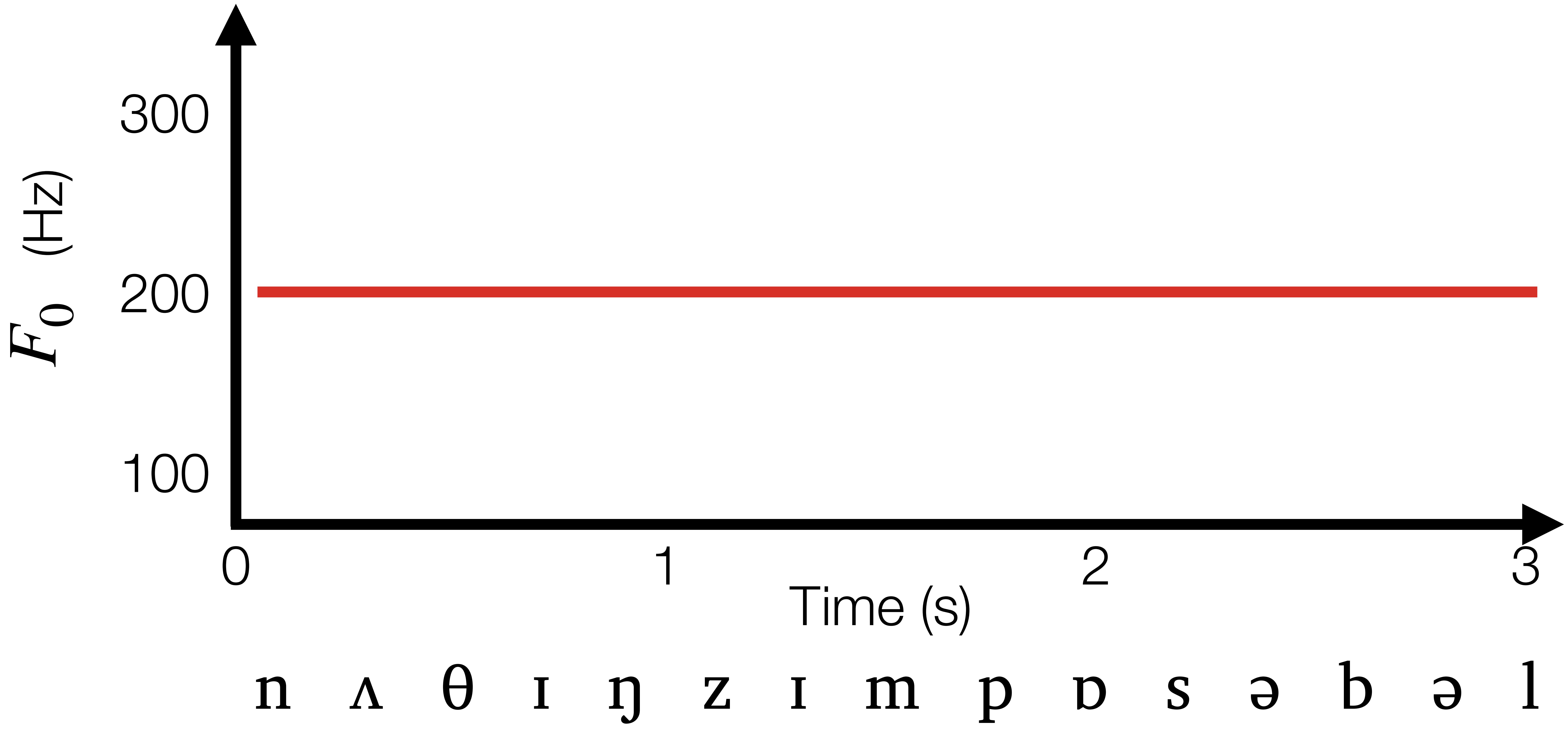
Phrasing

Presently Wilbur raised his head and began speaking in that strange, resonant fashion which hinted at sound-producing organs unlike the run of mankind's.

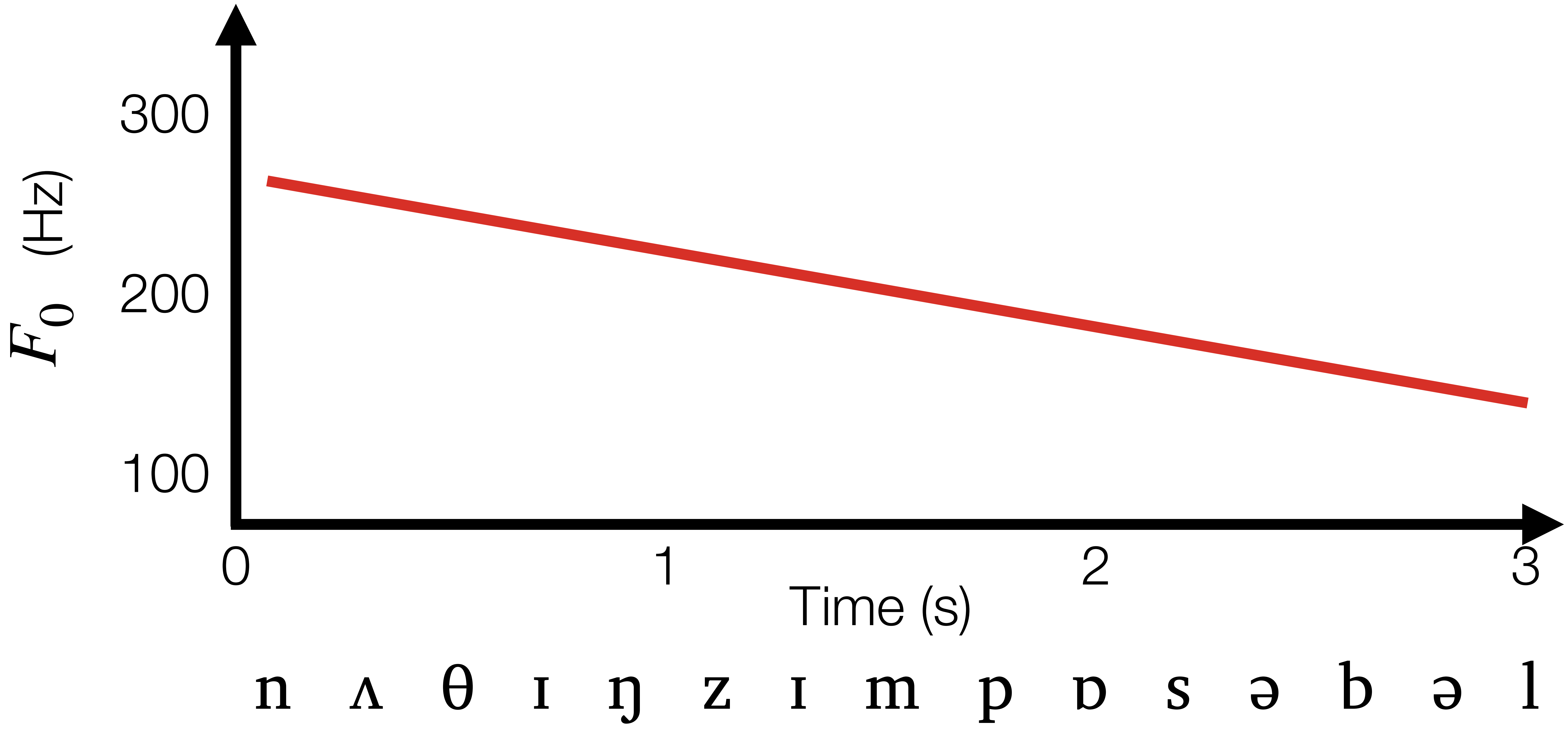
Duration



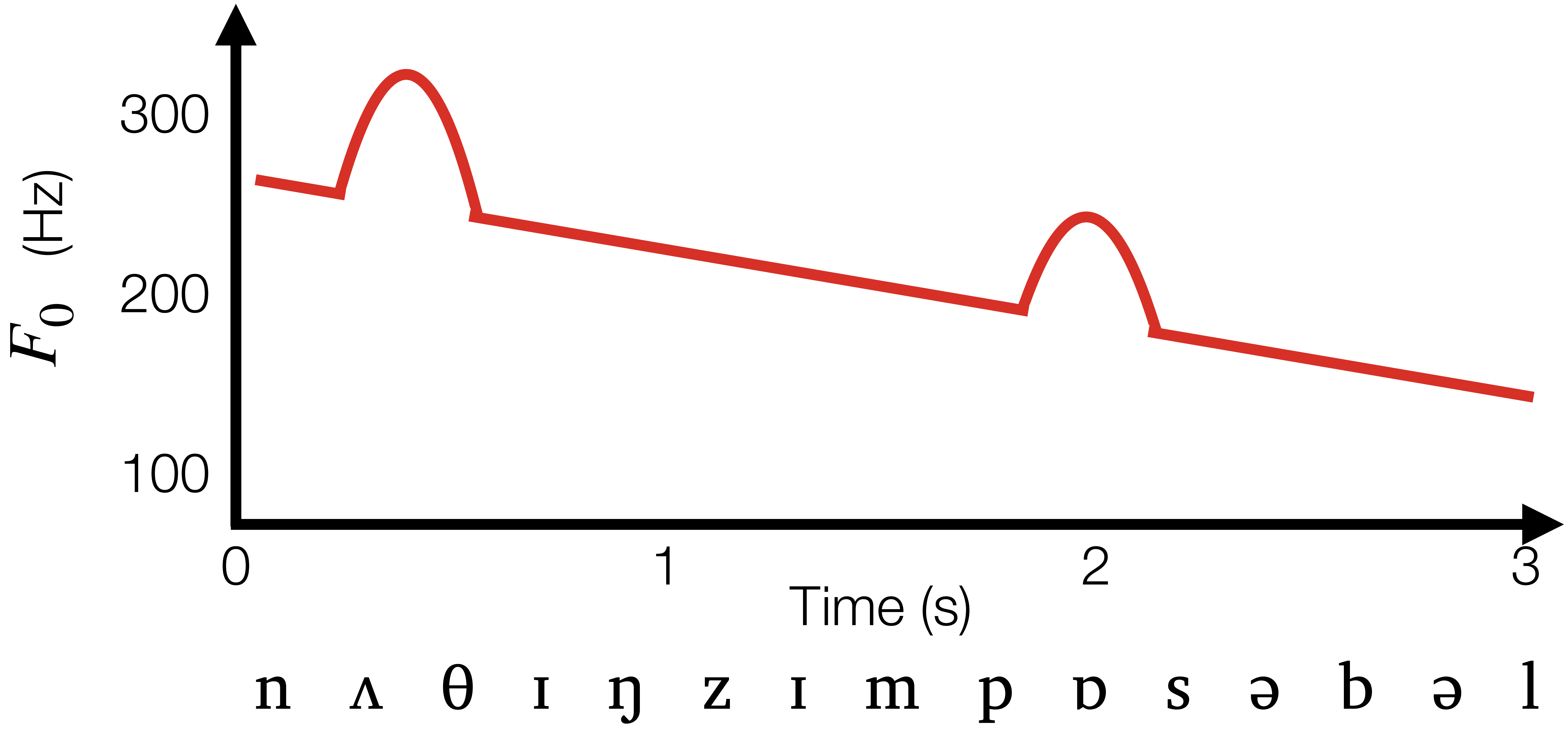
F_0



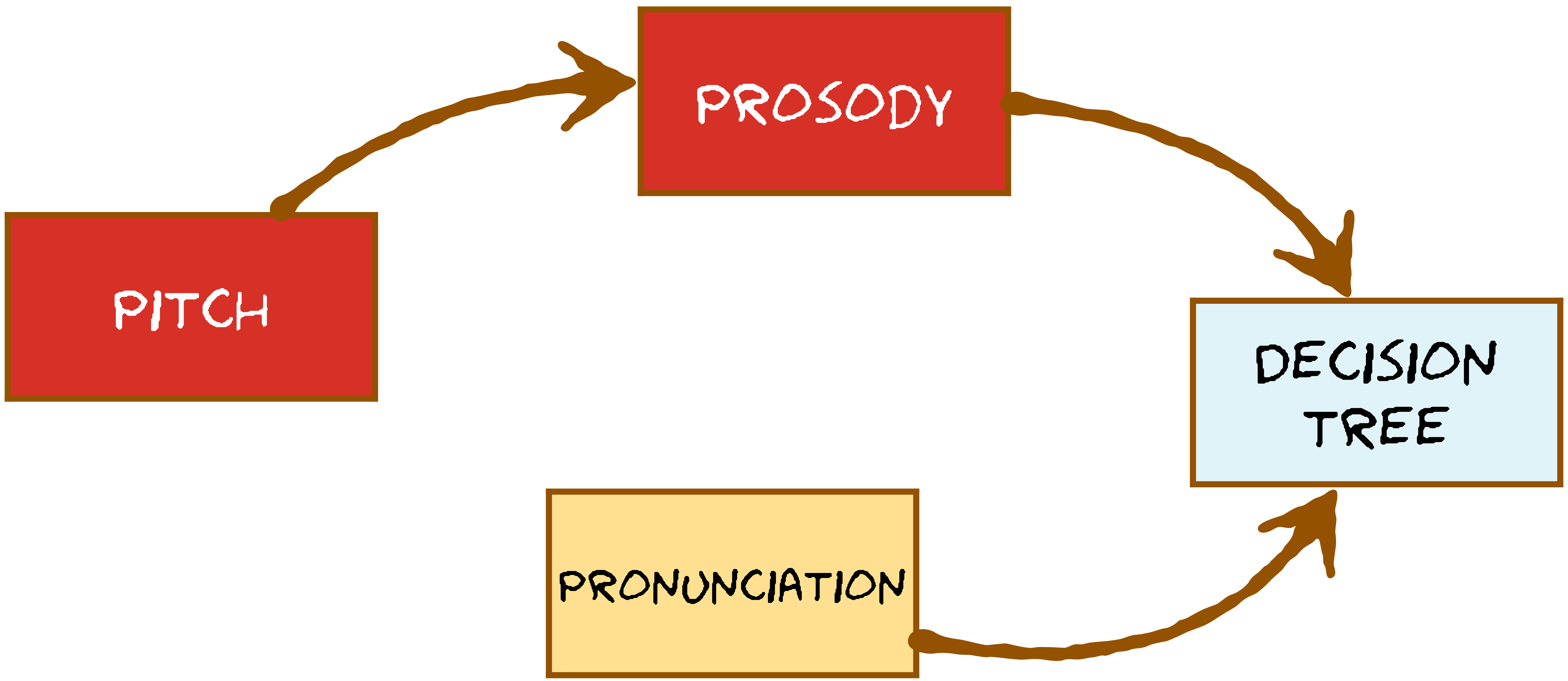
F_0



F_0

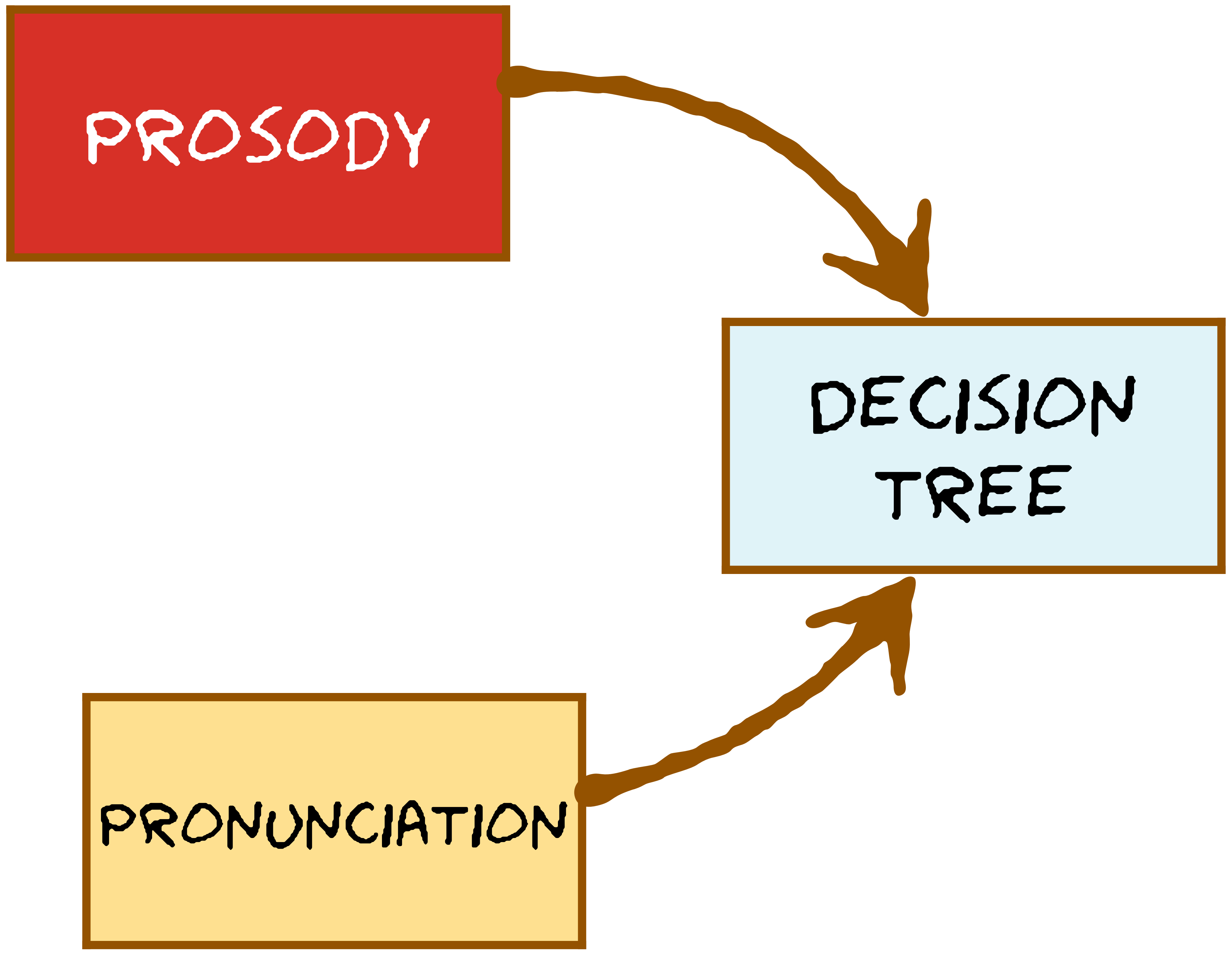


What you can learn next



DECISION TREE

What you need to know already

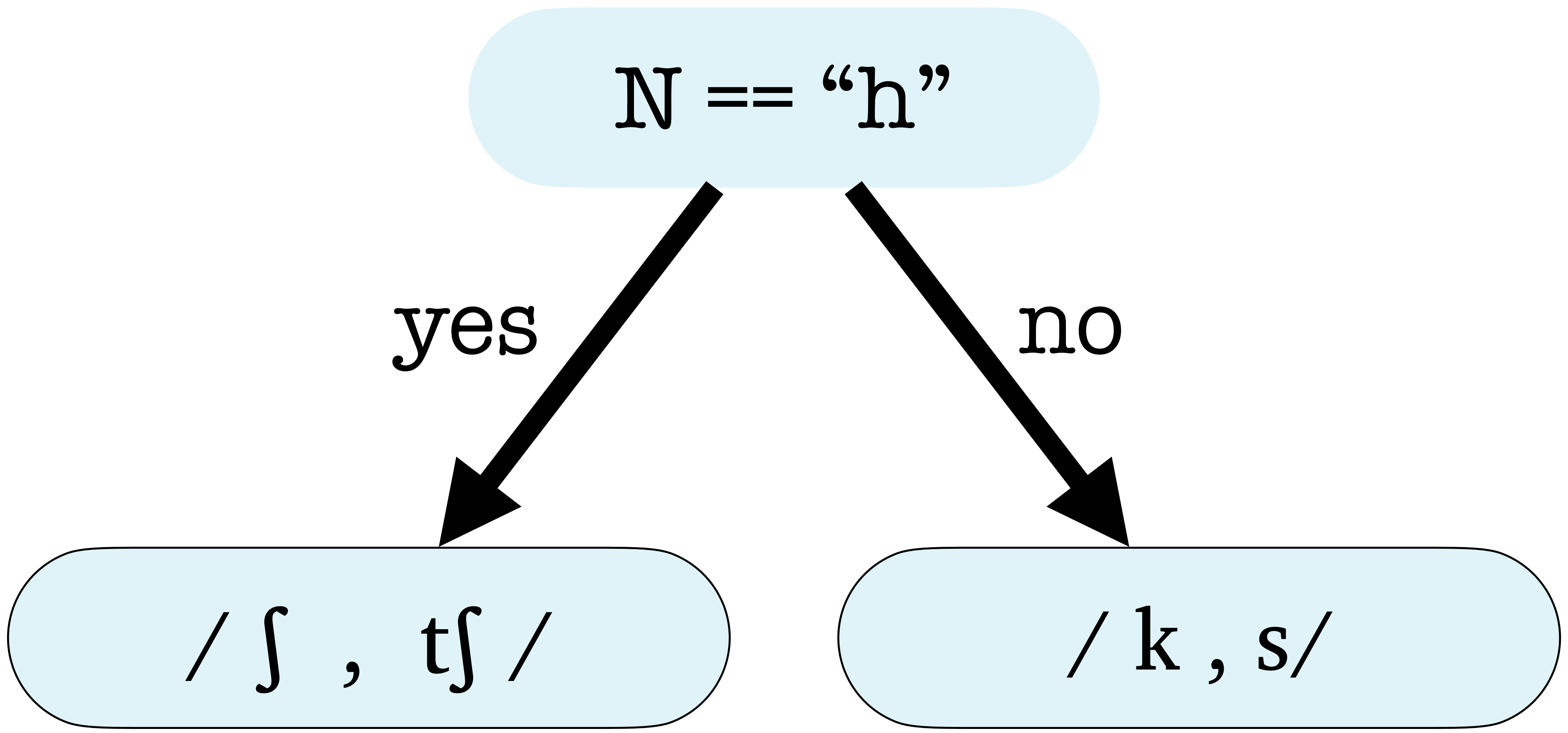


Predictors and predictee

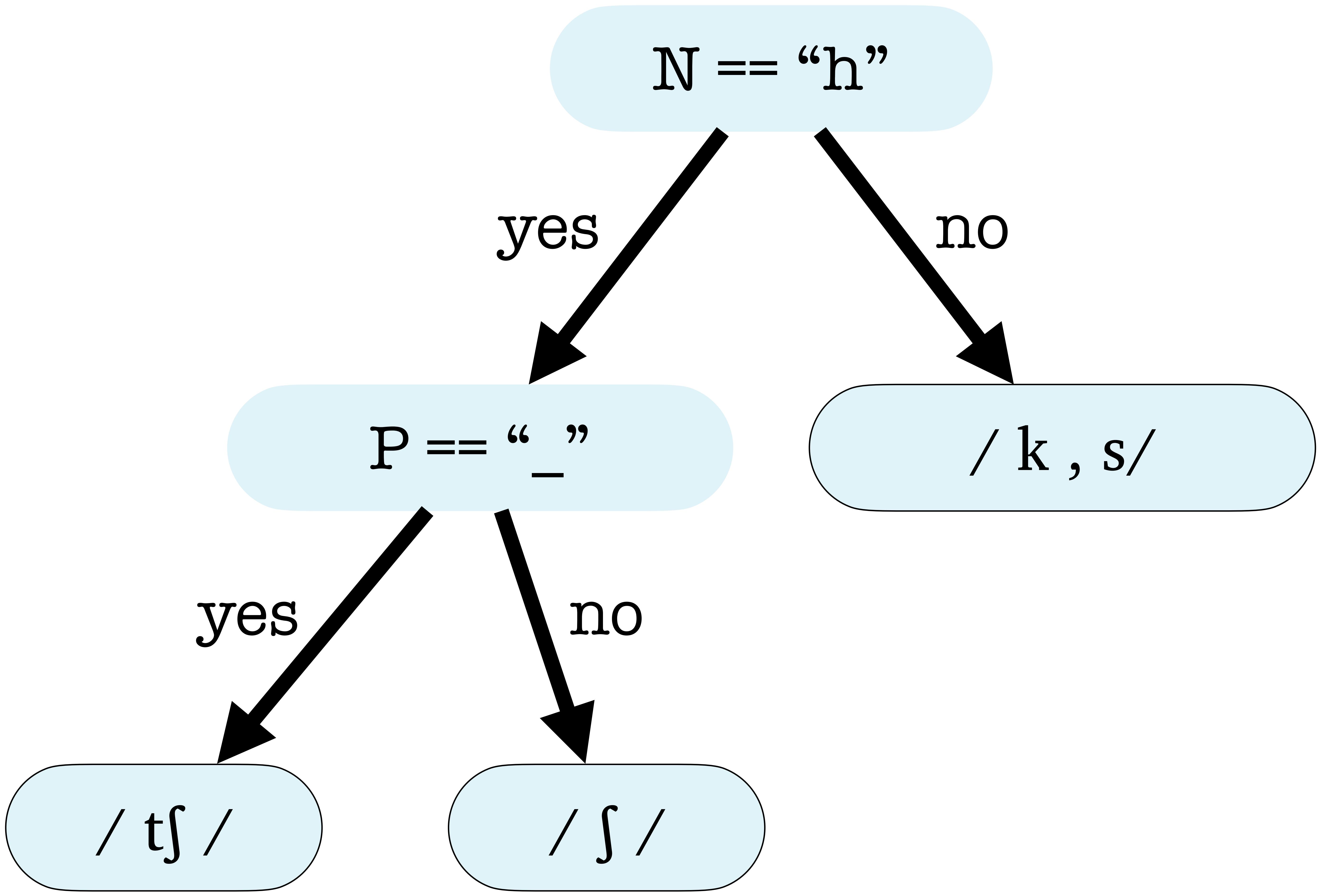
P	C	N		
_	c	i	=	/ s /
_	c	o	=	/ k /
_	c	h	=	/ tʃ /
i	c	h	=	/ ʃ /
i	c	e	=	/ s /
o	c	o	=	/ k /
e	c	o	=	/ k /

N == “h”

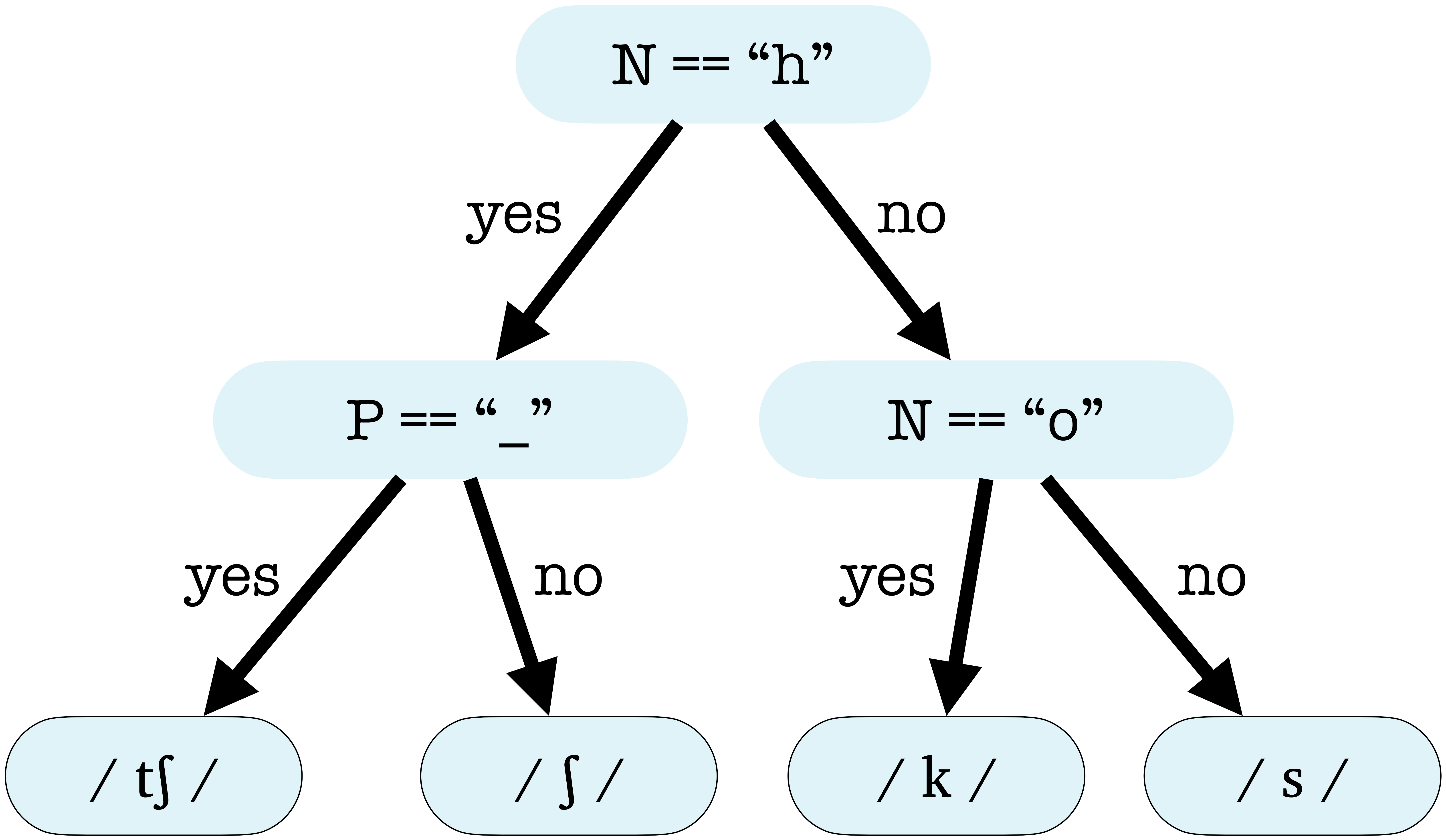
P	C	N		
_	c	i	=	/ s /
_	c	o	=	/ k /
_	c	h	=	/ tʃ /
i	c	h	=	/ ʃ /
i	c	e	=	/ s /
o	c	o	=	/ k /
e	c	o	=	/ k /



P	C	N		
_	c	i	=	/ s /
_	c	o	=	/ k /
_	c	h	=	/ tʃ /
i	c	h	=	/ ʃ /
i	c	e	=	/ s /
o	c	o	=	/ k /
e	c	o	=	/ k /

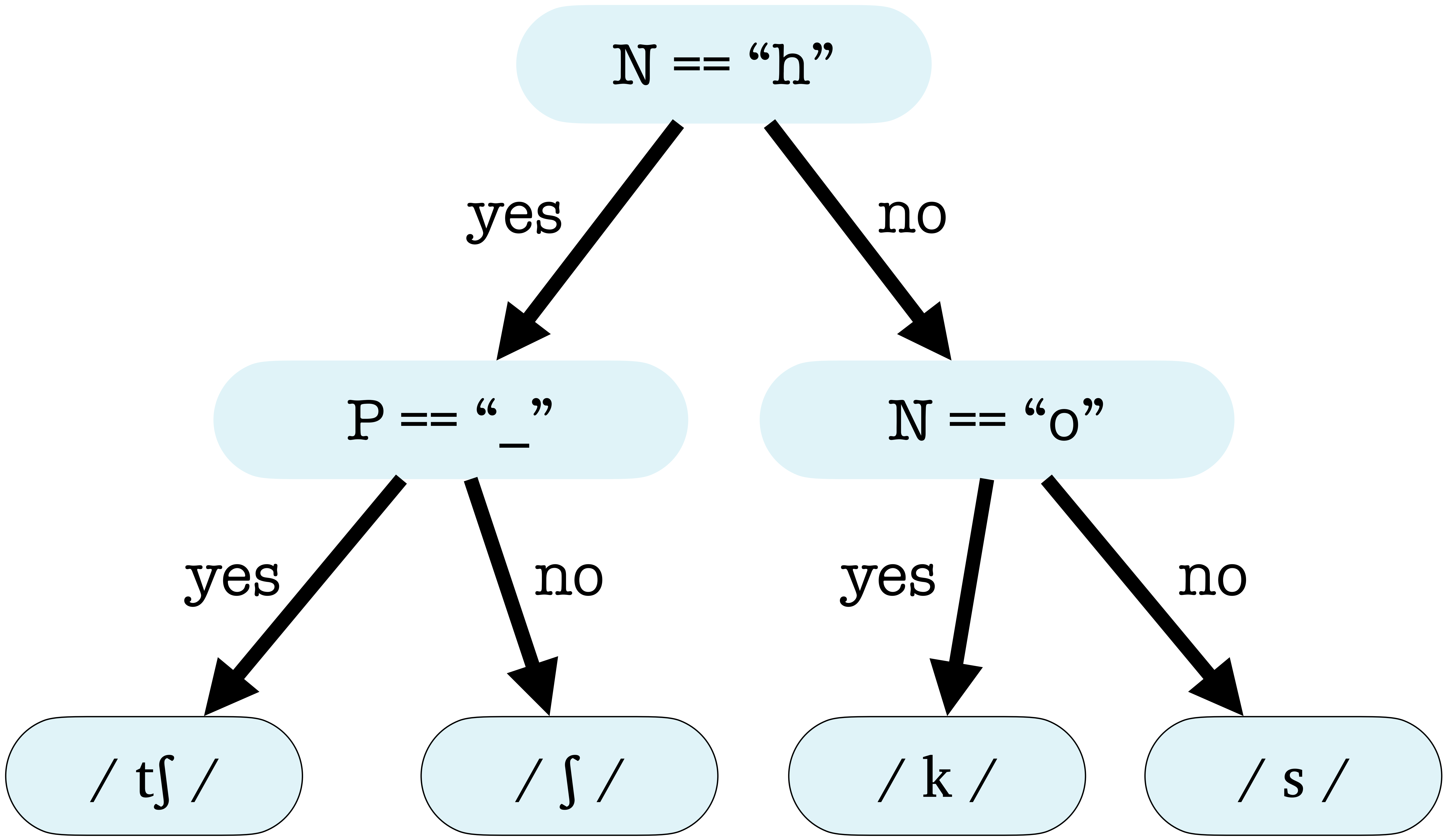


P	C	N		
_	c	i	=	/ s /
_	c	o	=	/ k /
_	c	h	=	/ tʃ /
i	c	h	=	/ ʃ /
i	c	e	=	/ s /
o	c	o	=	/ k /
e	c	o	=	/ k /

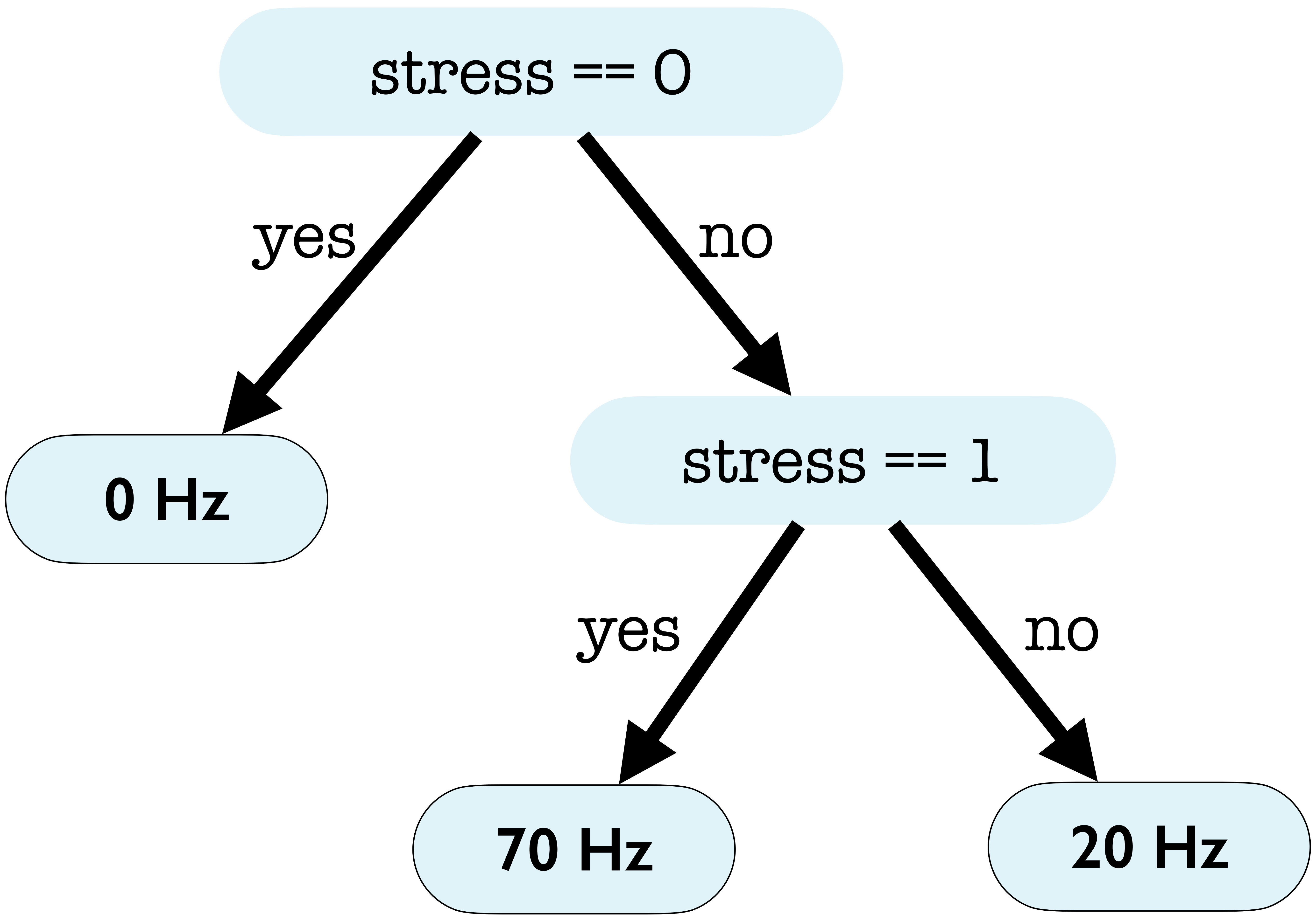


P	C	N		
_	c	i	=	/ s /
_	c	o	=	/ k /
_	c	h	=	/ tʃ /
i	c	h	=	/ ʃ /
i	c	e	=	/ s /
o	c	o	=	/ k /
e	c	o	=	/ k /

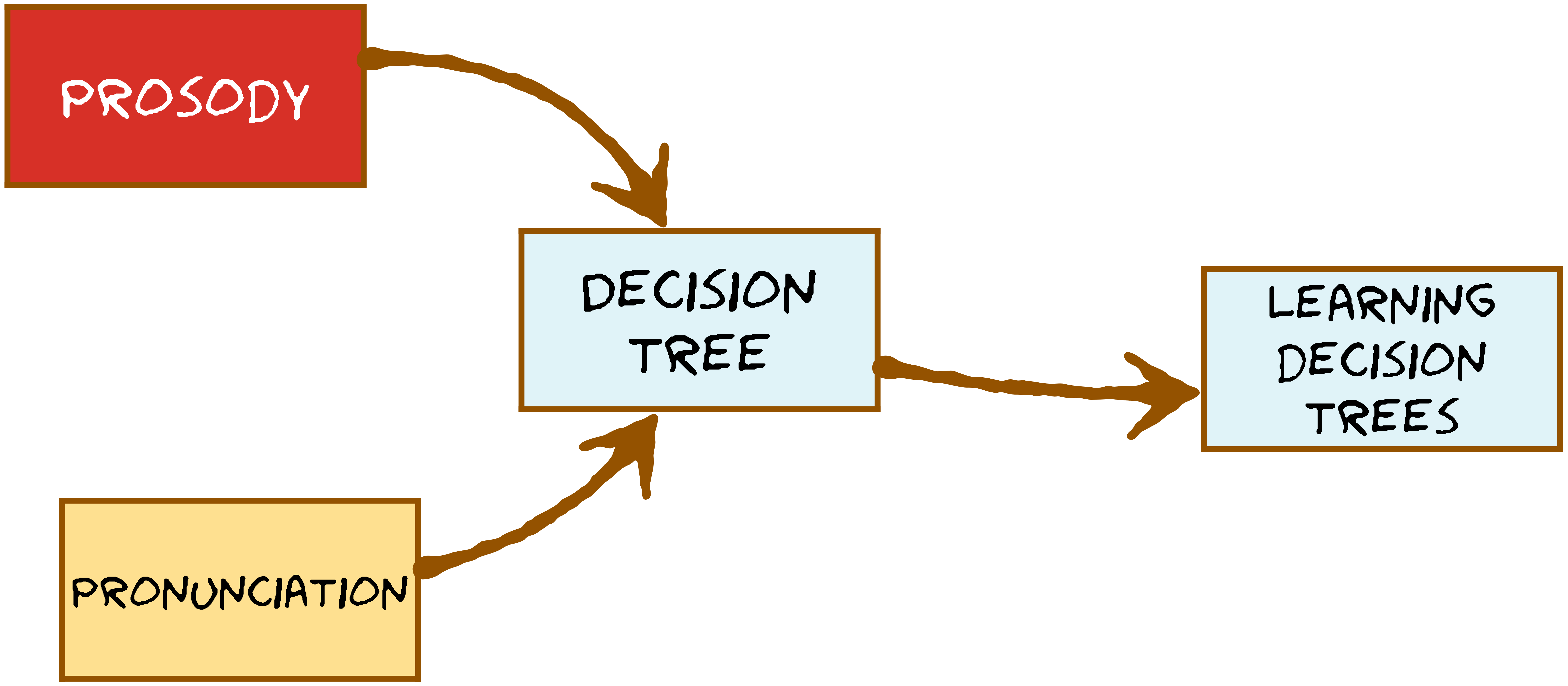
Classification Tree



Regression Tree

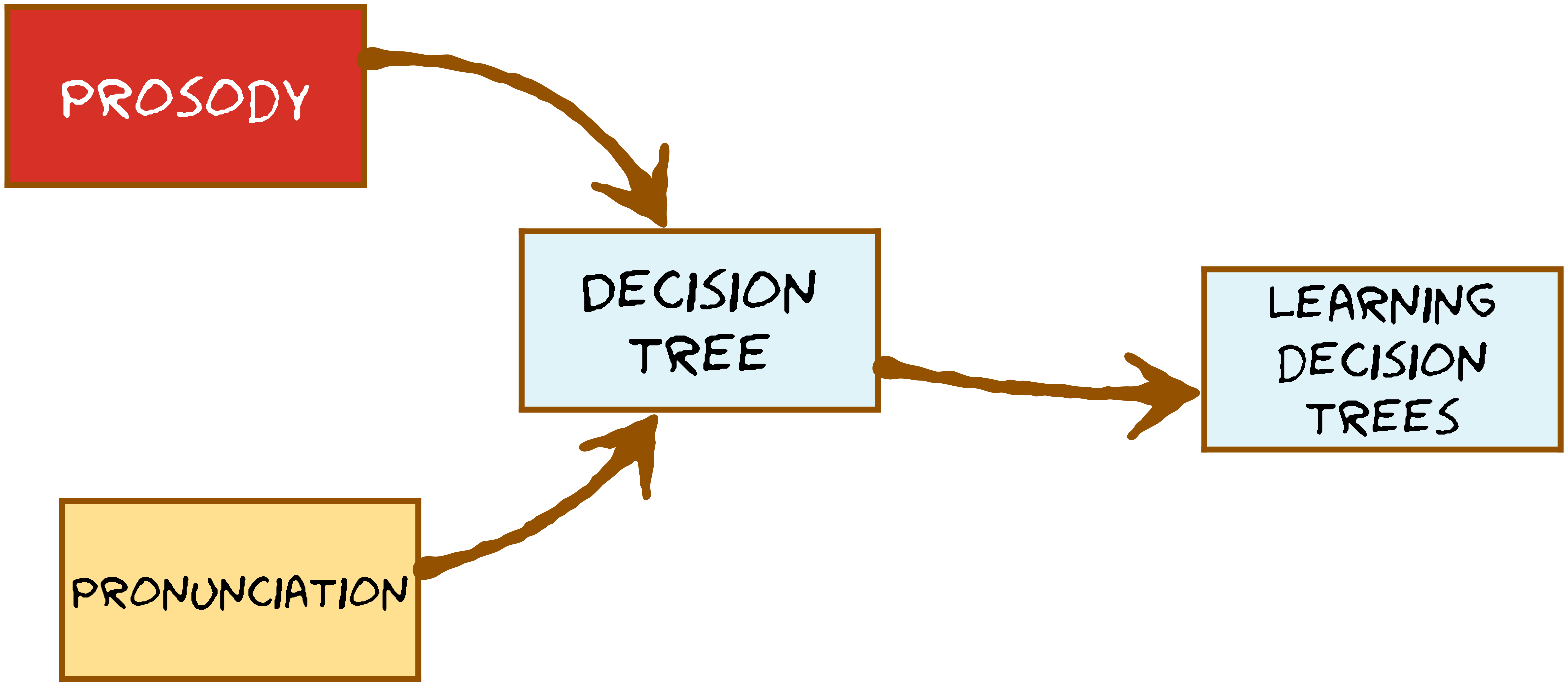


What you can learn next



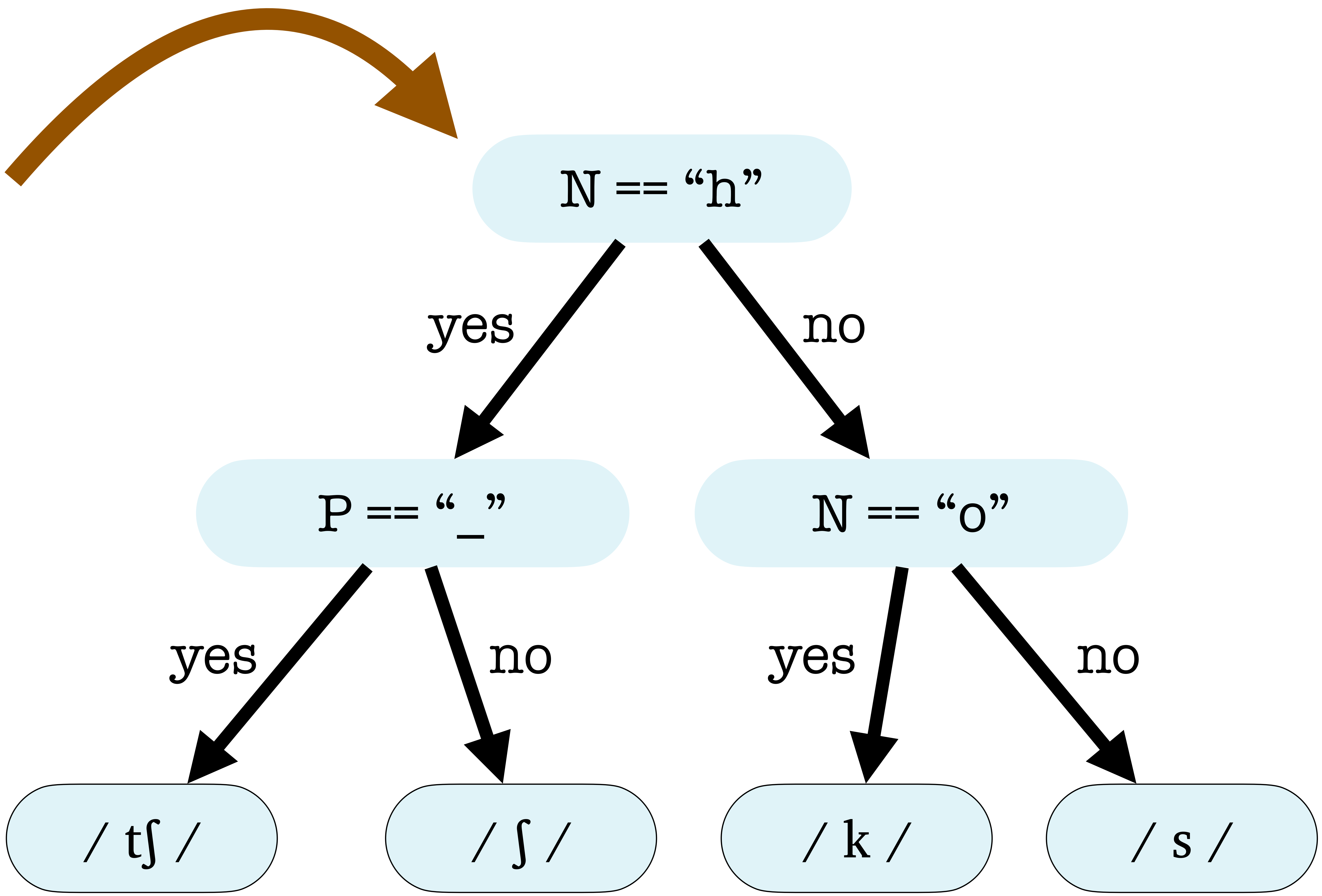
LEARNING DECISION TREES

What you need to know already



_	c	i	=	/s/
_	c	o	=	/k/
_	c	h	=	/tʃ/
i	c	h	=	/ʃ/
i	c	e	=	/s/
o	c	o	=	/k/
e	c	o	=	/k/

_	c	i	=	/ s /
_	c	o	=	/ k /
_	c	h	=	/ tʃ /
i	c	h	=	/ ʃ /
i	c	e	=	/ s /
o	c	o	=	/ k /
e	c	o	=	/ k /



chit tʃɪt

chloride klɔraɪd

chrome kroʊm

colony kɑːləni

deflect dɪflɛkt

freelancer friːlænsɝ

lance læns

locked lɑːkt

marched mɑːrtʃt

...

watchdog wɑːtʃdɔg

yachts jɑːts

zinc zɪŋk

Get the data ready for machine learning

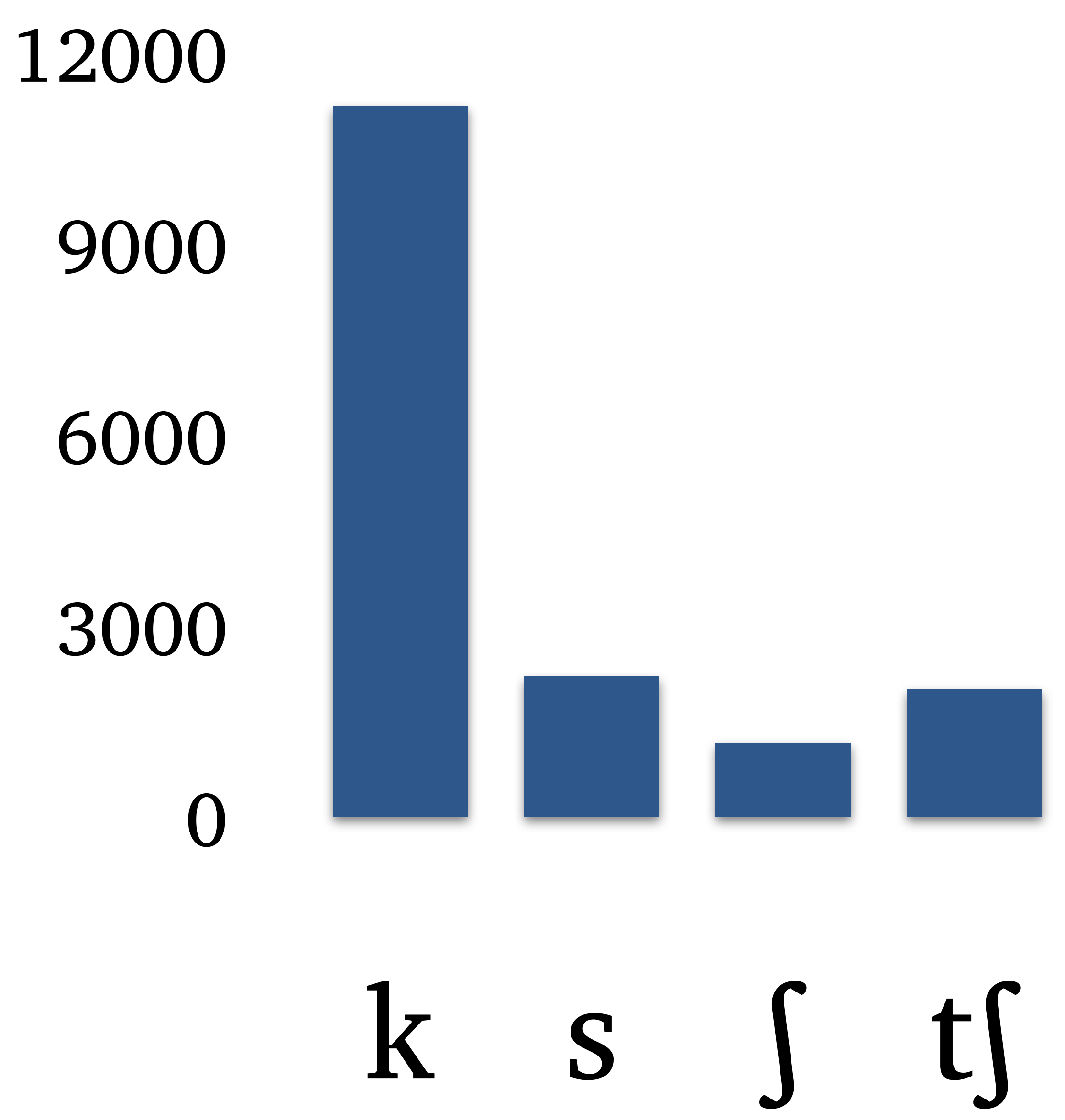
`aback əbæk`	`back_ k`
`achieve ətʃiːv`	`_achi tʃ`
`acord əkɔrd`	`_acor k`
`alice ælɪs`	`lice_ s`
`ambiance æmbiːəns`	`ance_ s`
`bench bɛntʃ`	`ench_ tʃ`
`borsch bɔrʃ`	`rsch_ ʃ`
`branch bræntʃ`	`anch_ tʃ`
`call kɔl`	`__cal k`
`cardboard kɑːrdbɔrd`	`__car k`

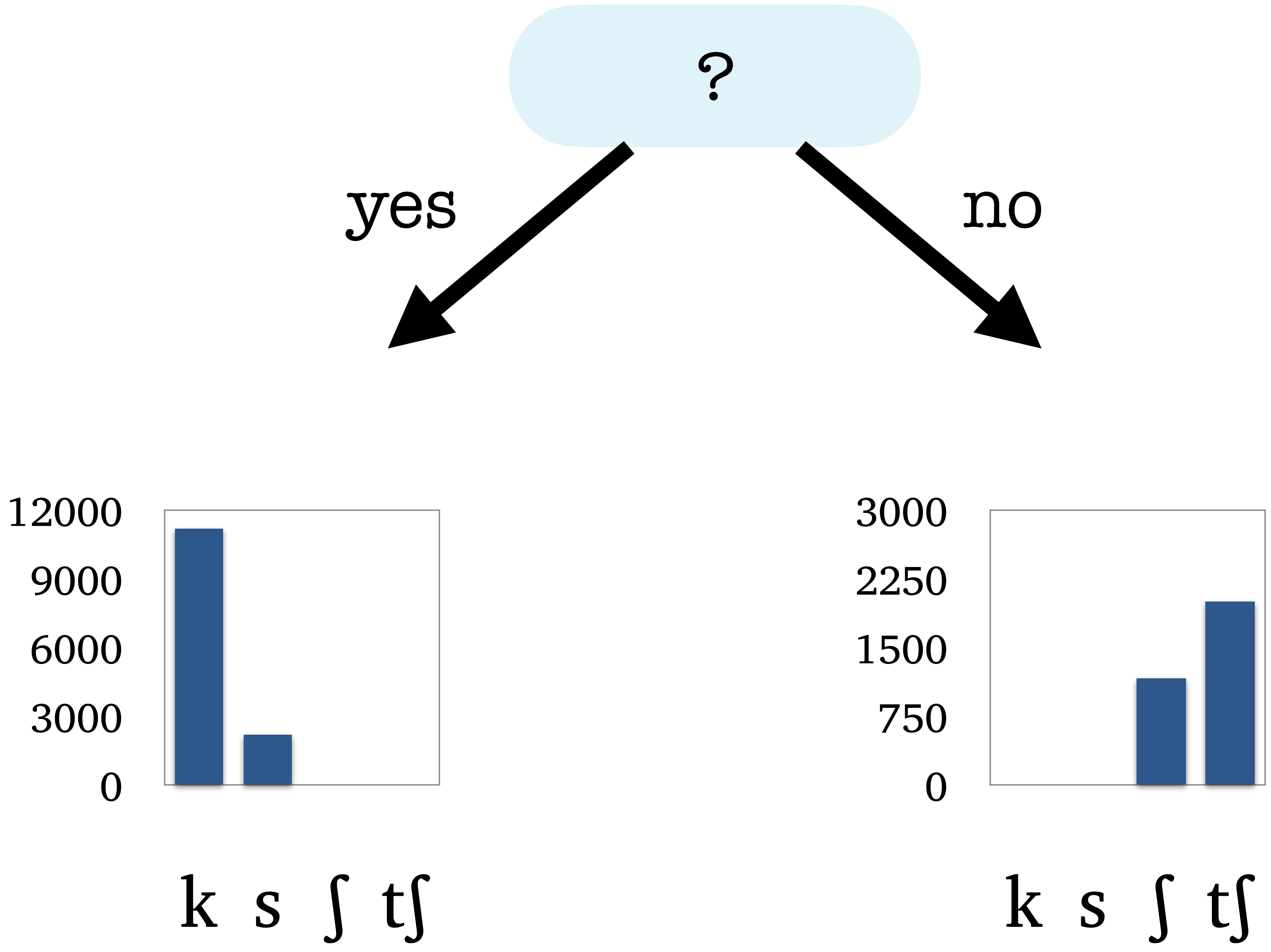
ieced	s	__car	k	licat	k	recei	s
_scho	ʃ	recor	k	rich_	k	nic__	k
__cal	k	focht	k	pecor	k	_acqu	k
gic__	k	__cha	tʃ	dacit	s	anca_	k
arcos	k	__cha	tʃ	rick_	k	ouch_	tʃ
__che	tʃ	ercei	s	decad	k	__car	k
orca_	k	racto	k	__cot	k	isch_	ʃ
duca_	k	uechn	k	__cha	tʃ	__cla	k
__cir	s	recia	tʃ	__cen	s	__chi	tʃ
__cha	tʃ	decli	k	mac__	k	_mcle	k

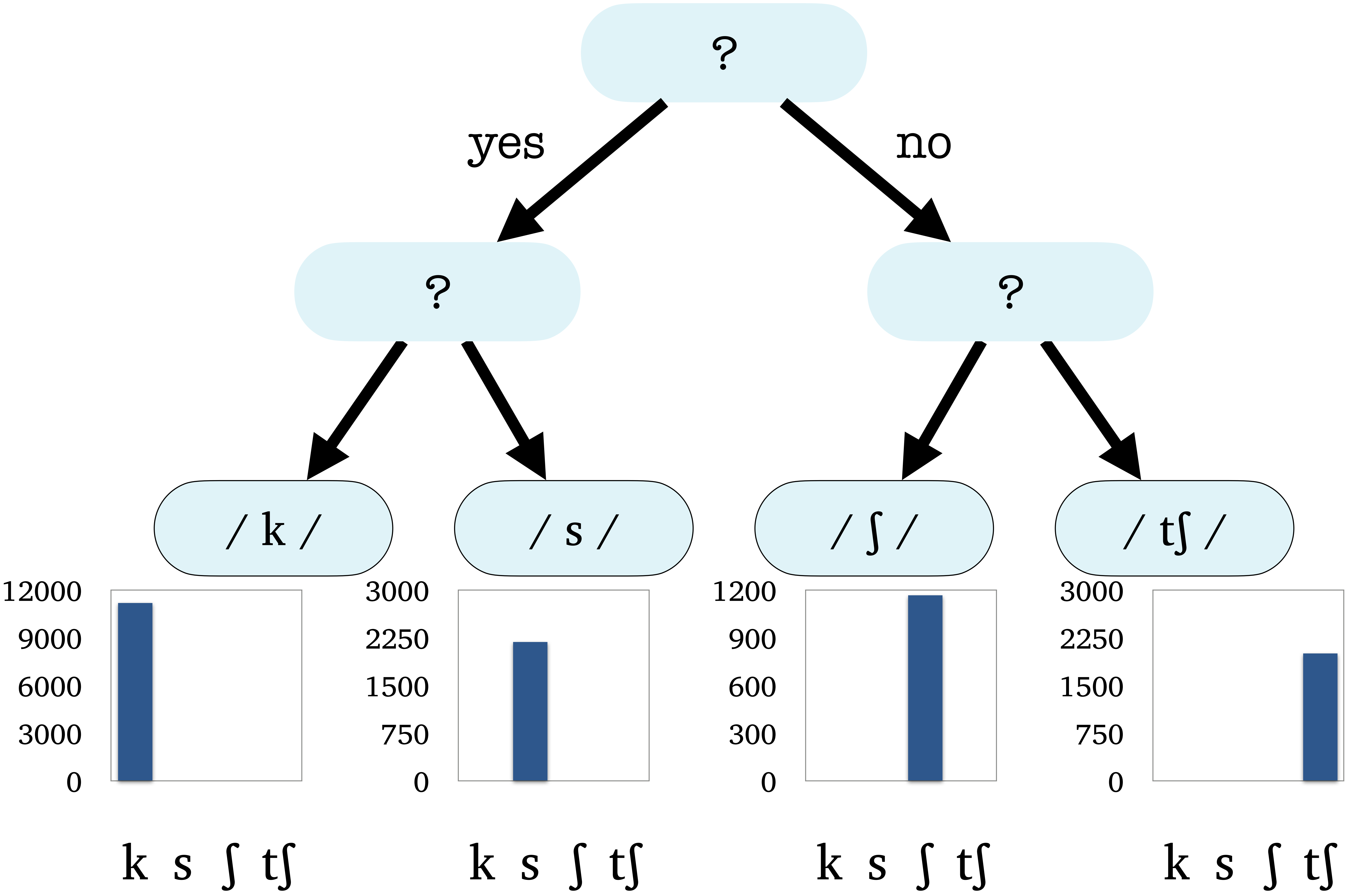
What exactly *is* machine learning?

`archi ?`

ieced	s	__car	k	licat	k	recei	s
_scho	ʃ	recor	k	rich_	k	nic__	k
__cal	k	focht	k	pecor	k	_acqu	k
gic__	k	__cha	tʃ	dacit	s	anca_	k
arcos	k	__cha	tʃ	rick_	k	ouch_	tʃ
__che	tʃ	ercei	s	decad	k	__car	k
orca_	k	racto	k	__cot	k	isch_	ʃ
duca_	k	uechn	k	__cha	tʃ	__cla	k
__cir	s	recia	tʃ	__cen	s	__chi	tʃ
__cha	tʃ	decli	k	mac__	k	_mcle	k

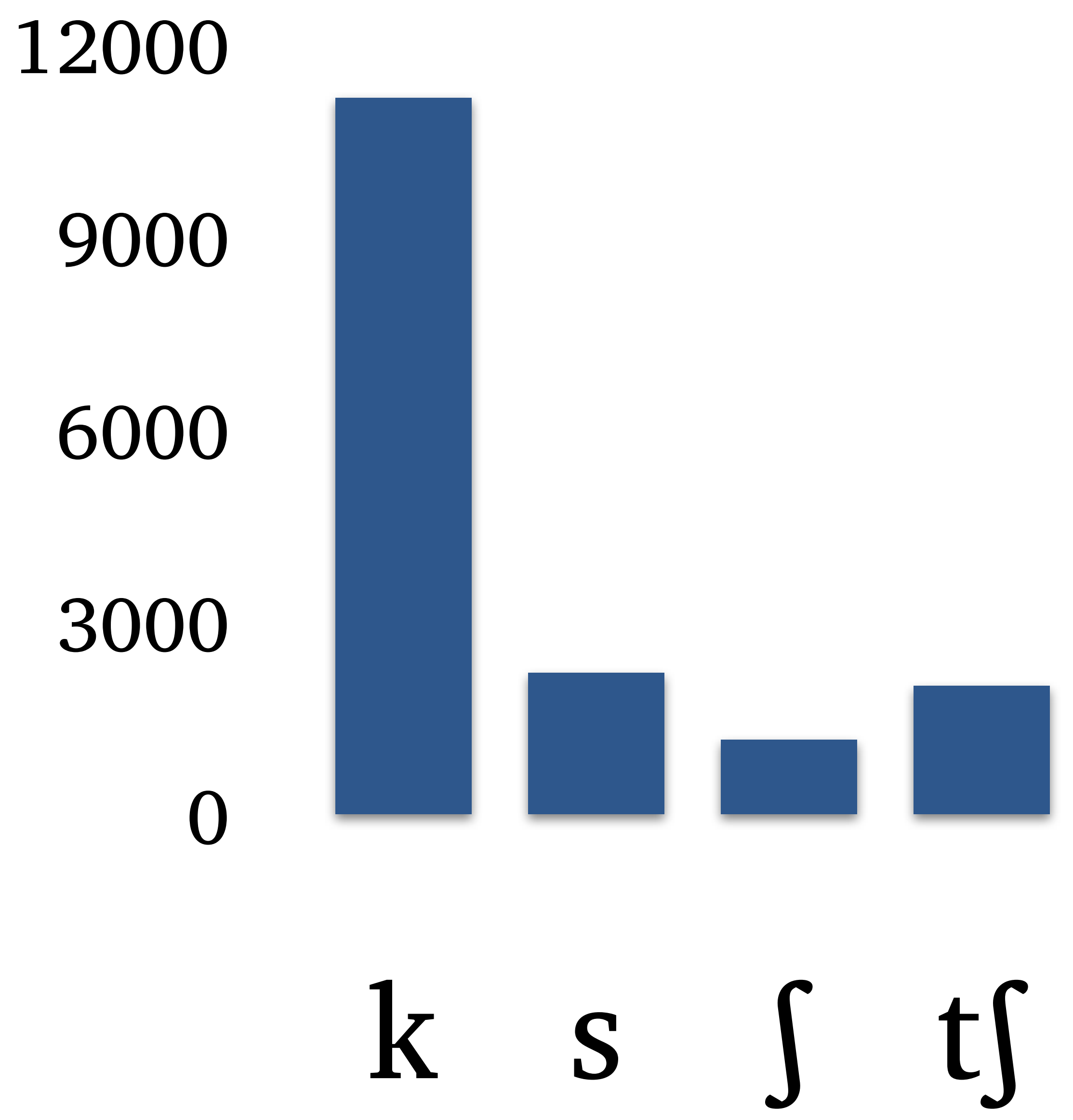



?
yes
no
12000
9000
6000
3000
0
k s ʃ tʃ
3000
2250
1500
750
0
k s ʃ tʃ

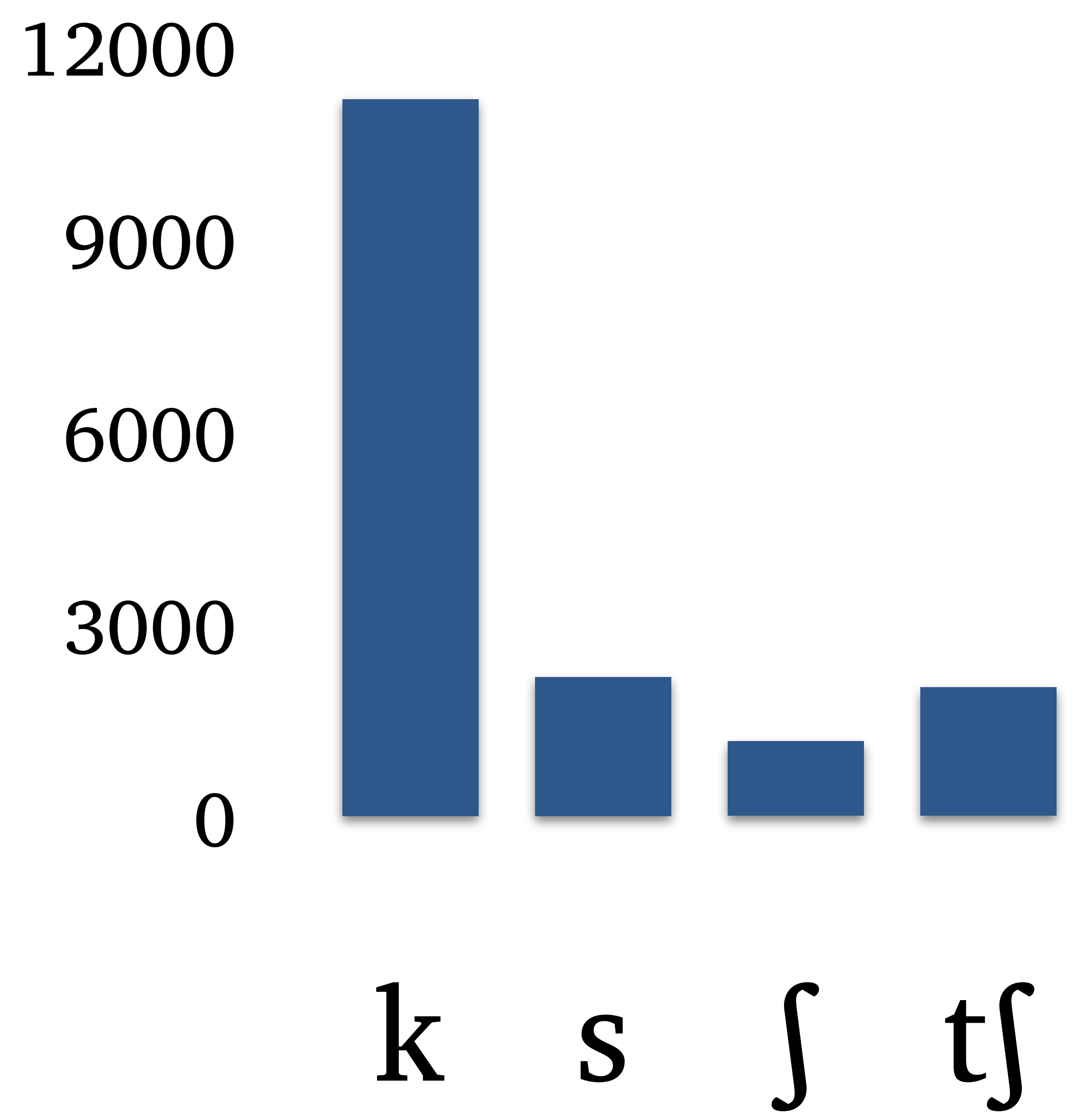

?
yes
no
?
?
/ k /
/ s /
/ ʃ /
/ tʃ /
12000
9000
6000
3000
0
3000
2250
1500
750
0
1200
900
600
300
0
3000
2250
1500
750
0
k s ʃ tʃ
k s ʃ tʃ
k s ʃ tʃ
k s ʃ tʃ

ieced	s	__car	k	licat	k	recei	s
_scho	ʃ	recor	k	rich_	k	nic__	k
__cal	k	focht	k	pecor	k	_acqu	k
gic__	k	__cha	tʃ	dacit	s	anca_	k
arcos	k	__cha	tʃ	rick_	k	ouch_	tʃ
__che	tʃ	ercei	s	decad	k	__car	k
orca_	k	racto	k	__cot	k	isch_	ʃ
duca_	k	uechn	k	__cha	tʃ	__cla	k
__cir	s	recia	tʃ	__cen	s	__chi	tʃ
__cha	tʃ	decli	k	mac__	k	_mcle	k

/ k /

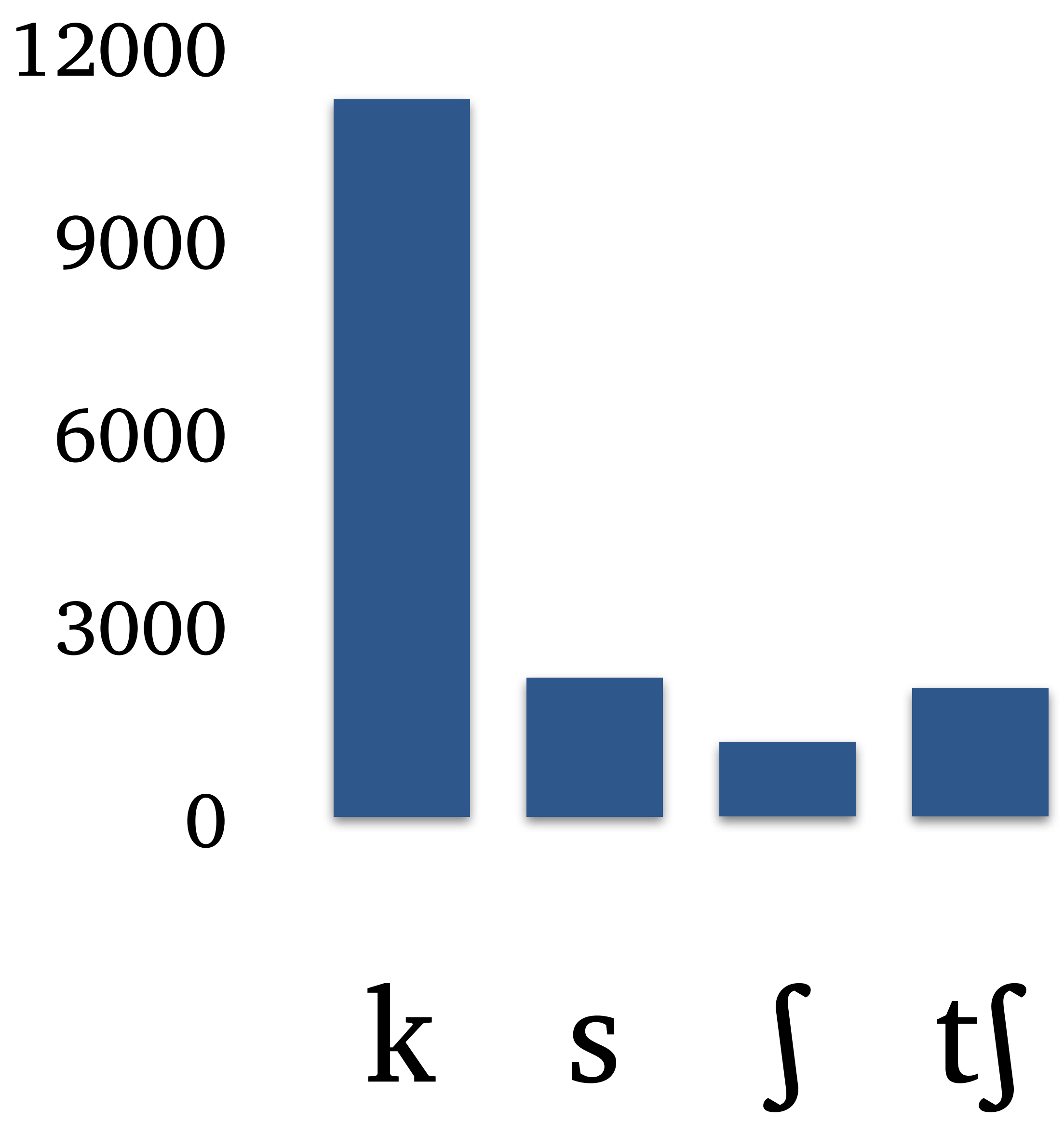


/ k /

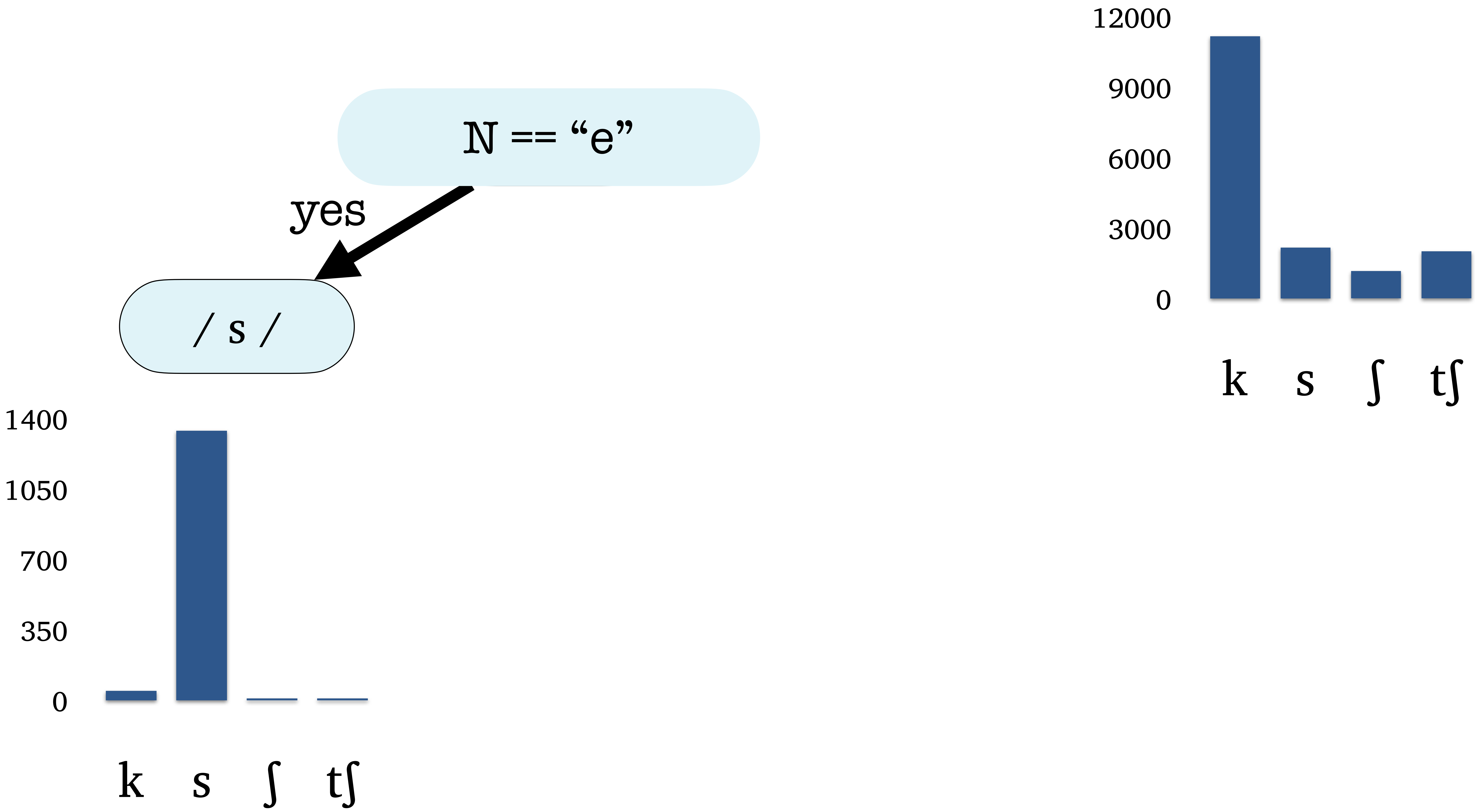


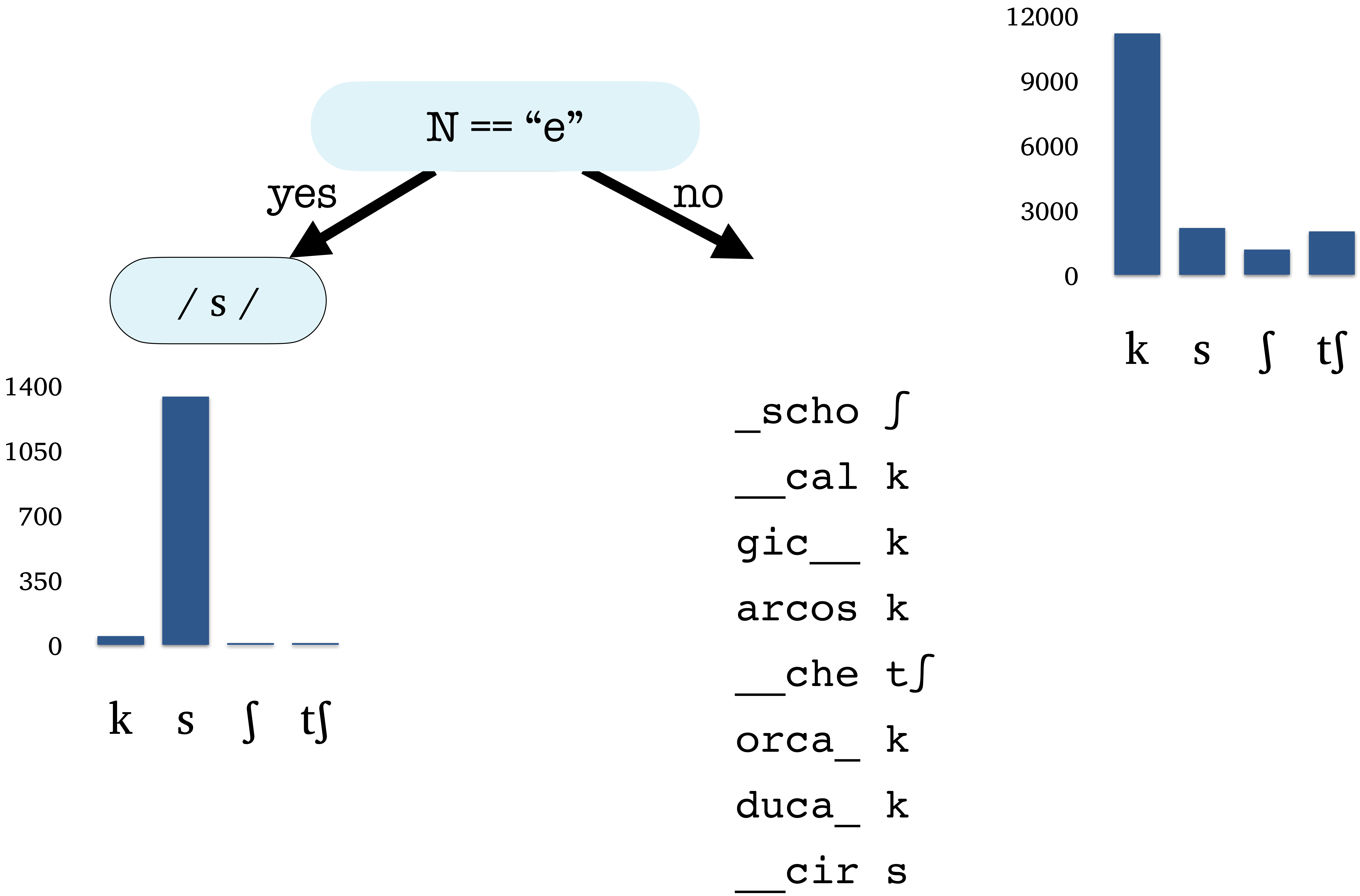
N == “e”

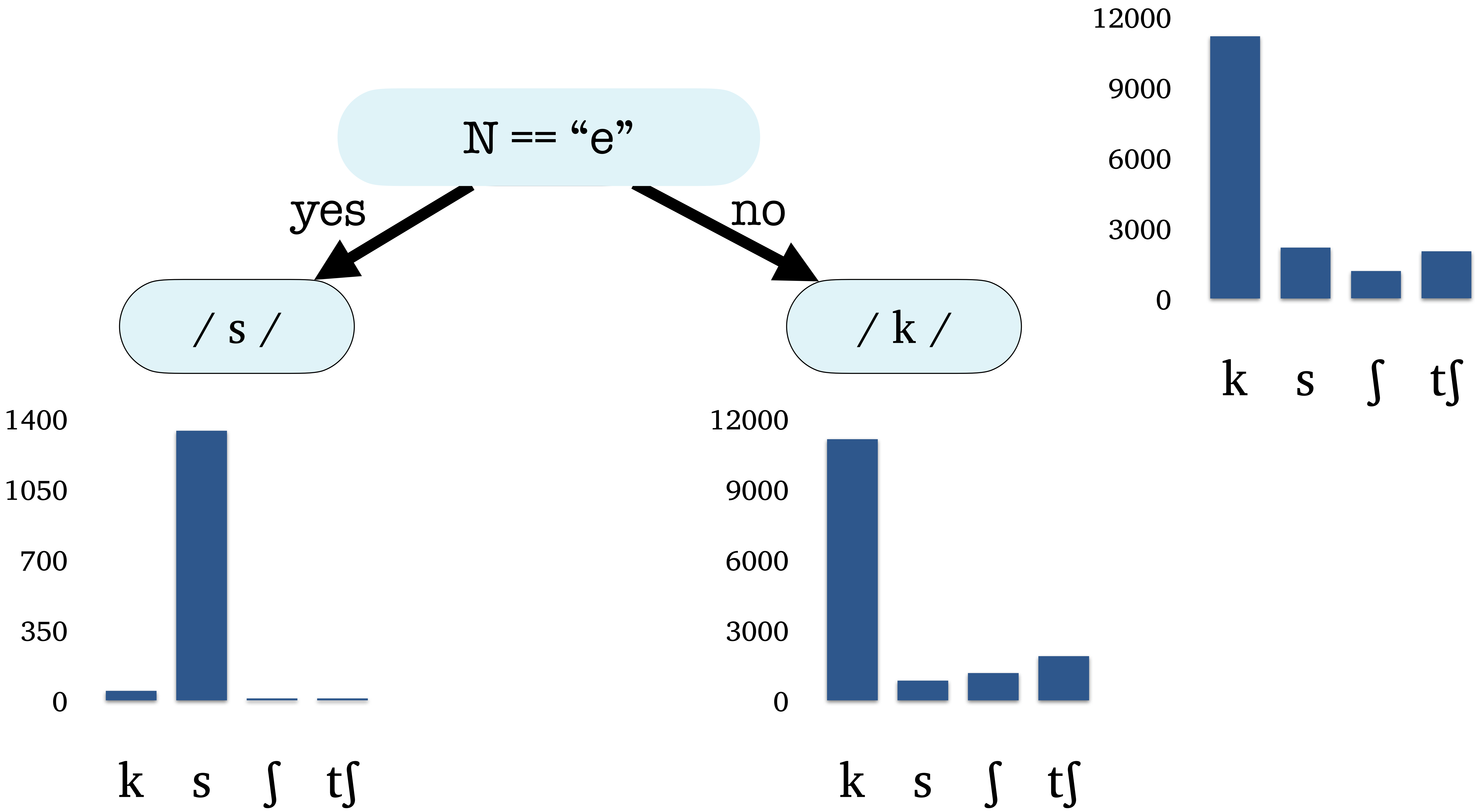
yes



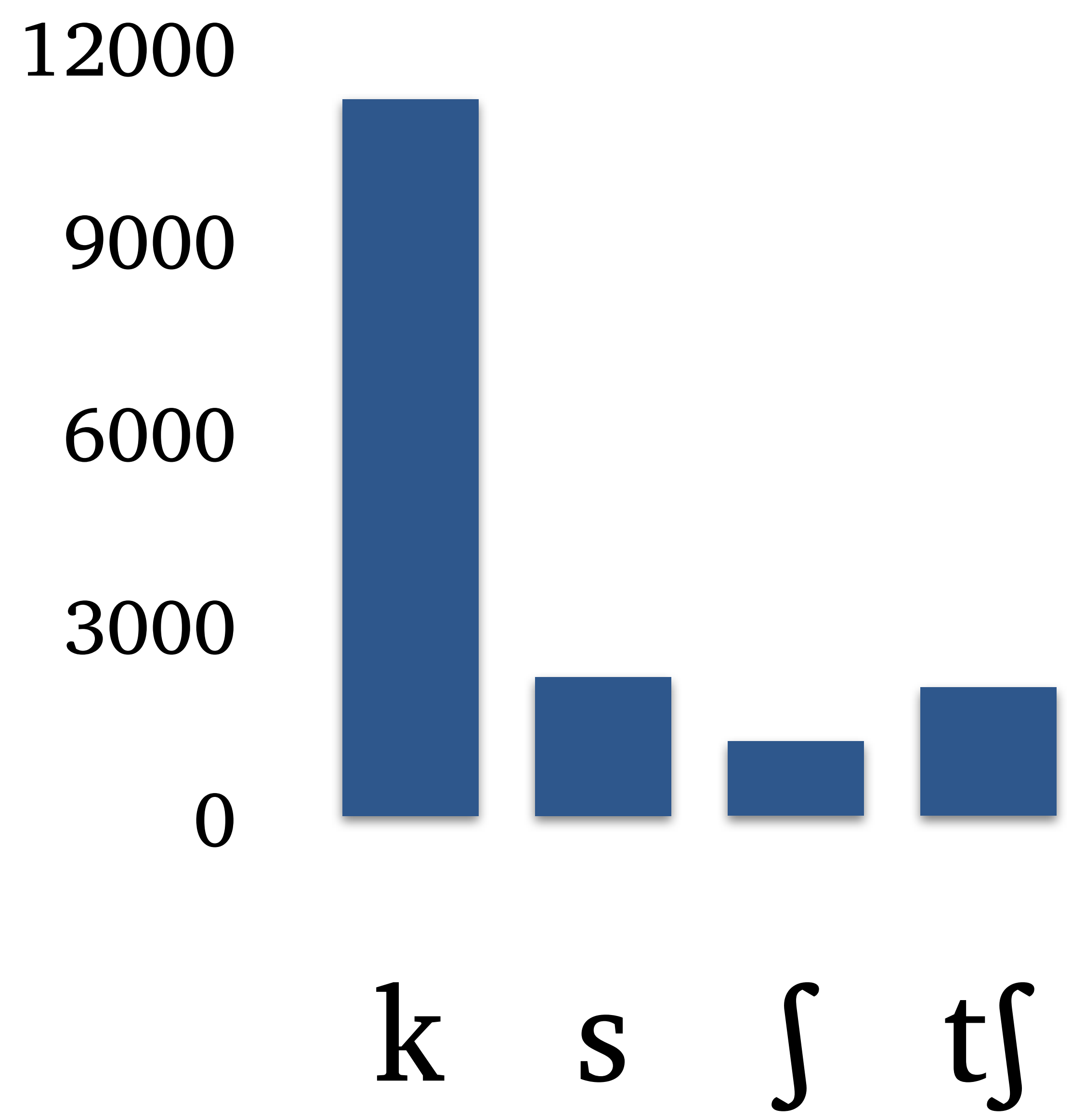

```
ieced s
ercei s
__cen s
recei s
vaced k
licen s
incer s
_scen s
```

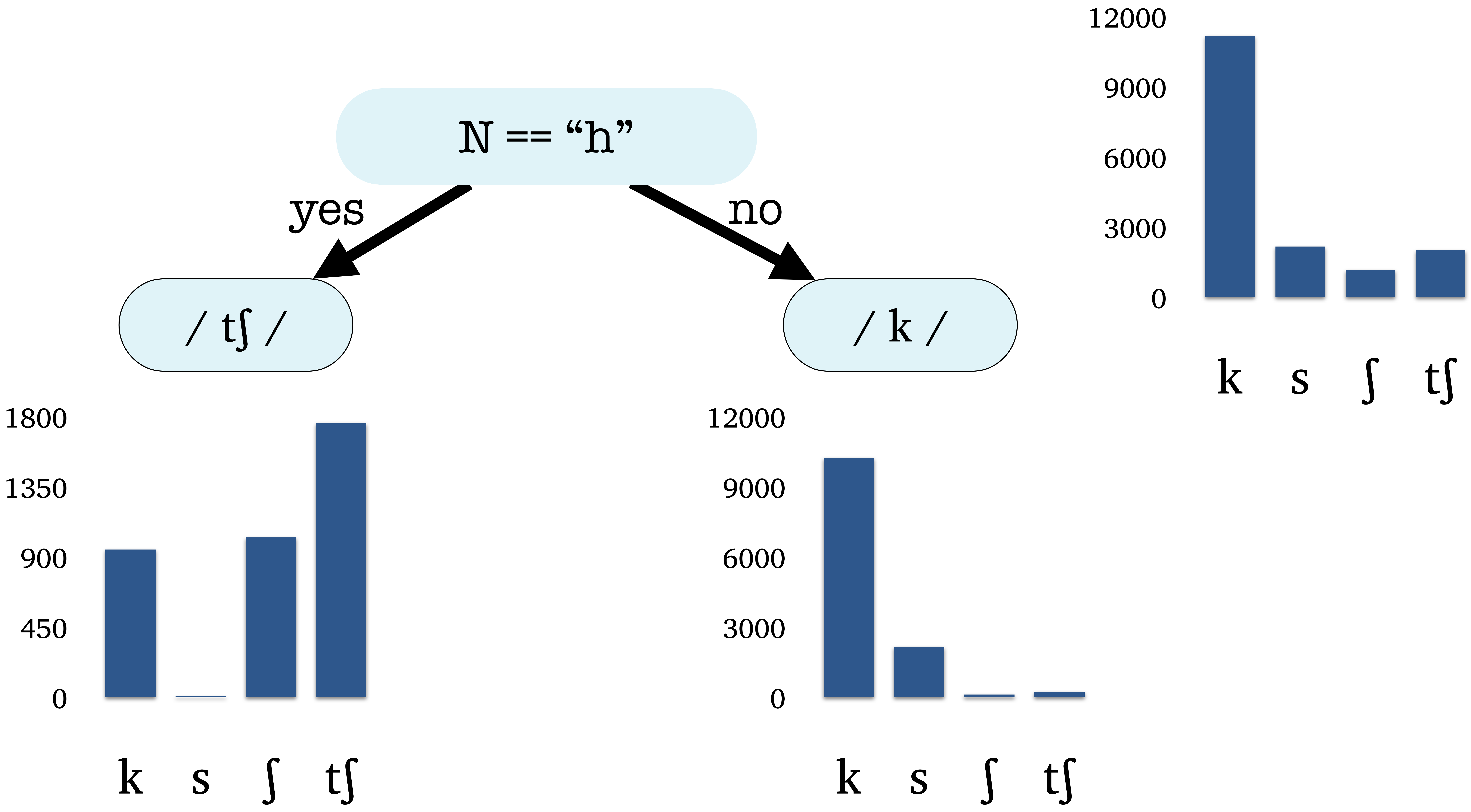

N == "e"
yes
/ s /
1400
1050
700
350
0
k s ʃ tʃ
12000
9000
6000
3000
0
k s ʃ tʃ


N == “e”
yes
no
/ s /
1400
1050
700
350
0
k s ʃ tʃ
12000
9000
6000
3000
0
k s ʃ tʃ
_scho ʃ
__cal k
gic__ k
arcos k
__che tʃ
orca_ k
duca_ k
__cir s

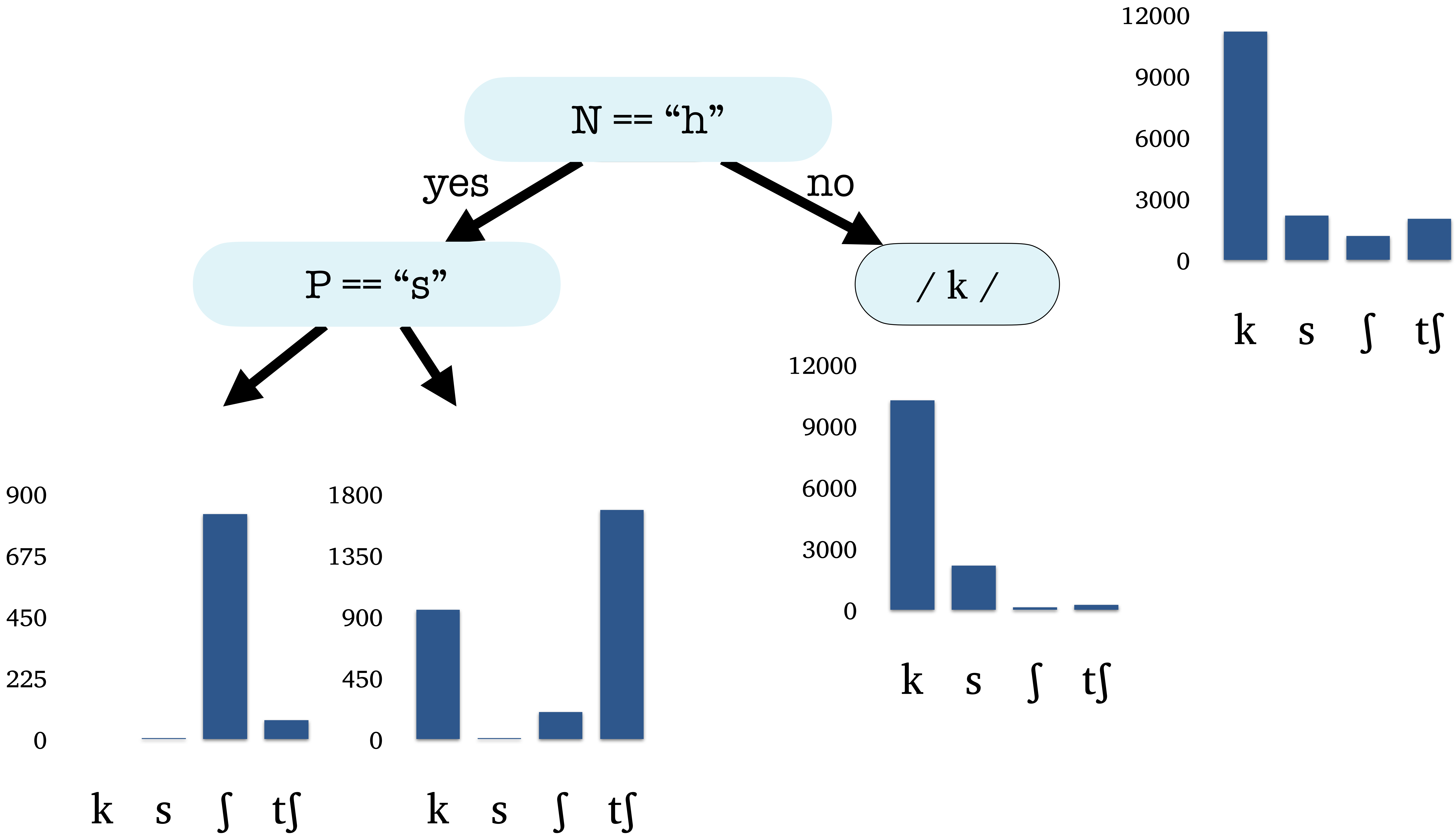

N == "e"
yes
no
/ s /
/ k /
12000
9000
6000
3000
0
k
s
ʃ
tʃ
1400
1050
700
350
0
k
s
ʃ
tʃ
12000
9000
6000
3000
0
k
s
ʃ
tʃ

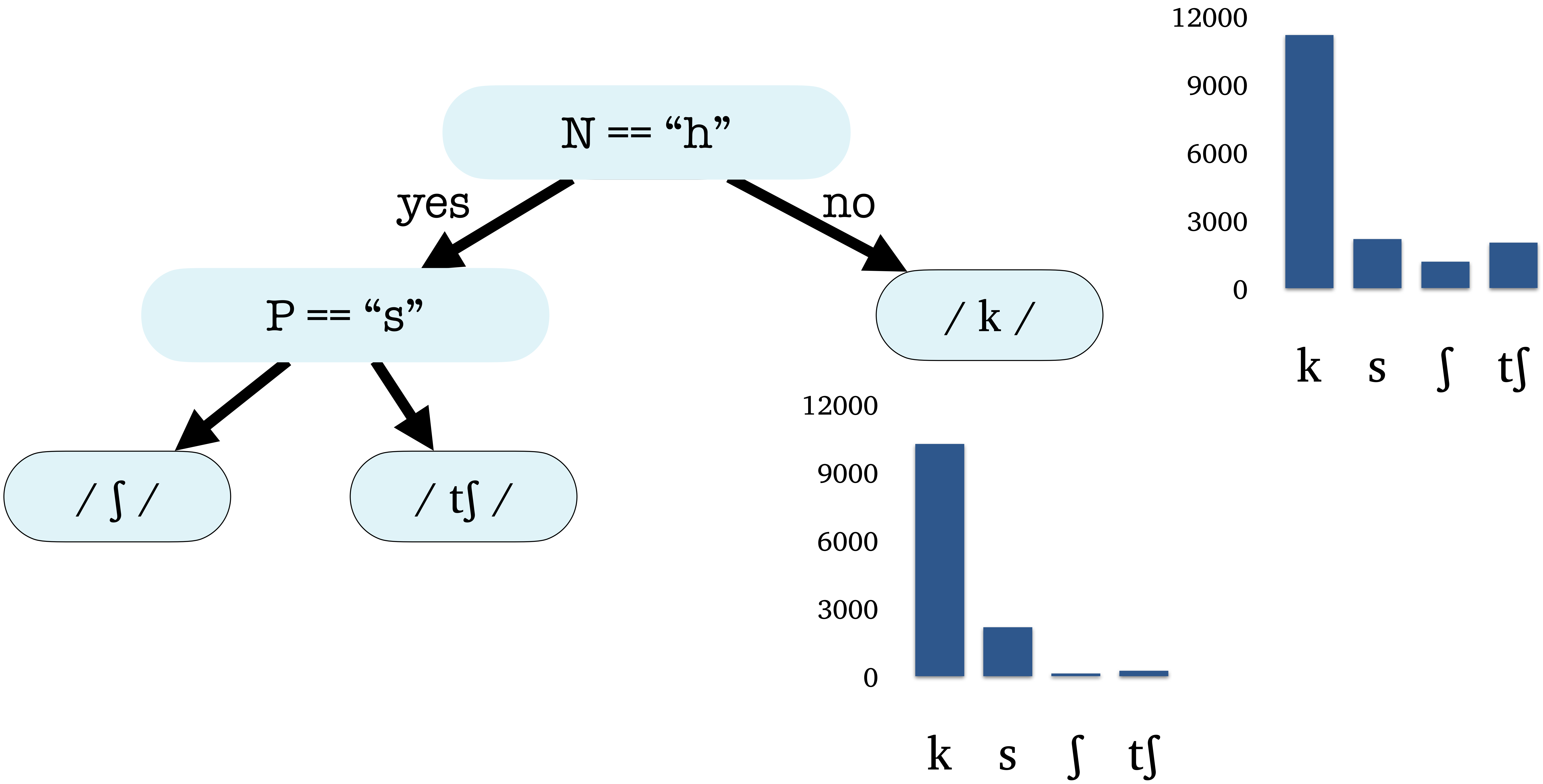
/ k /

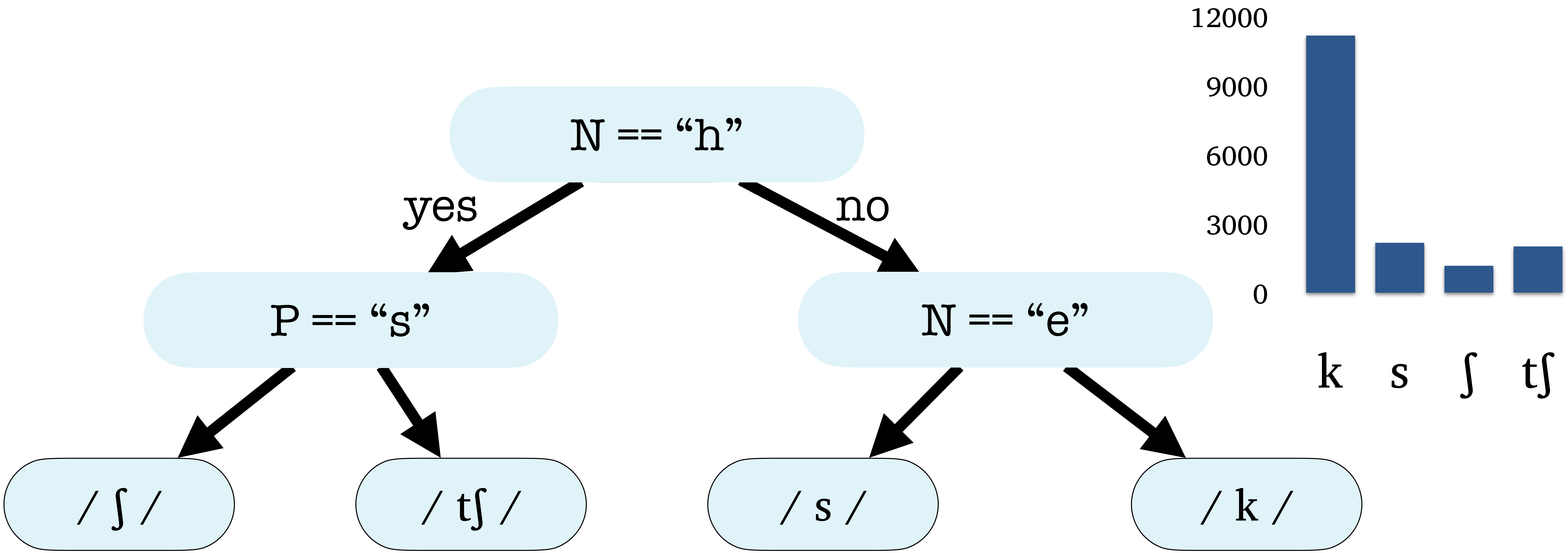


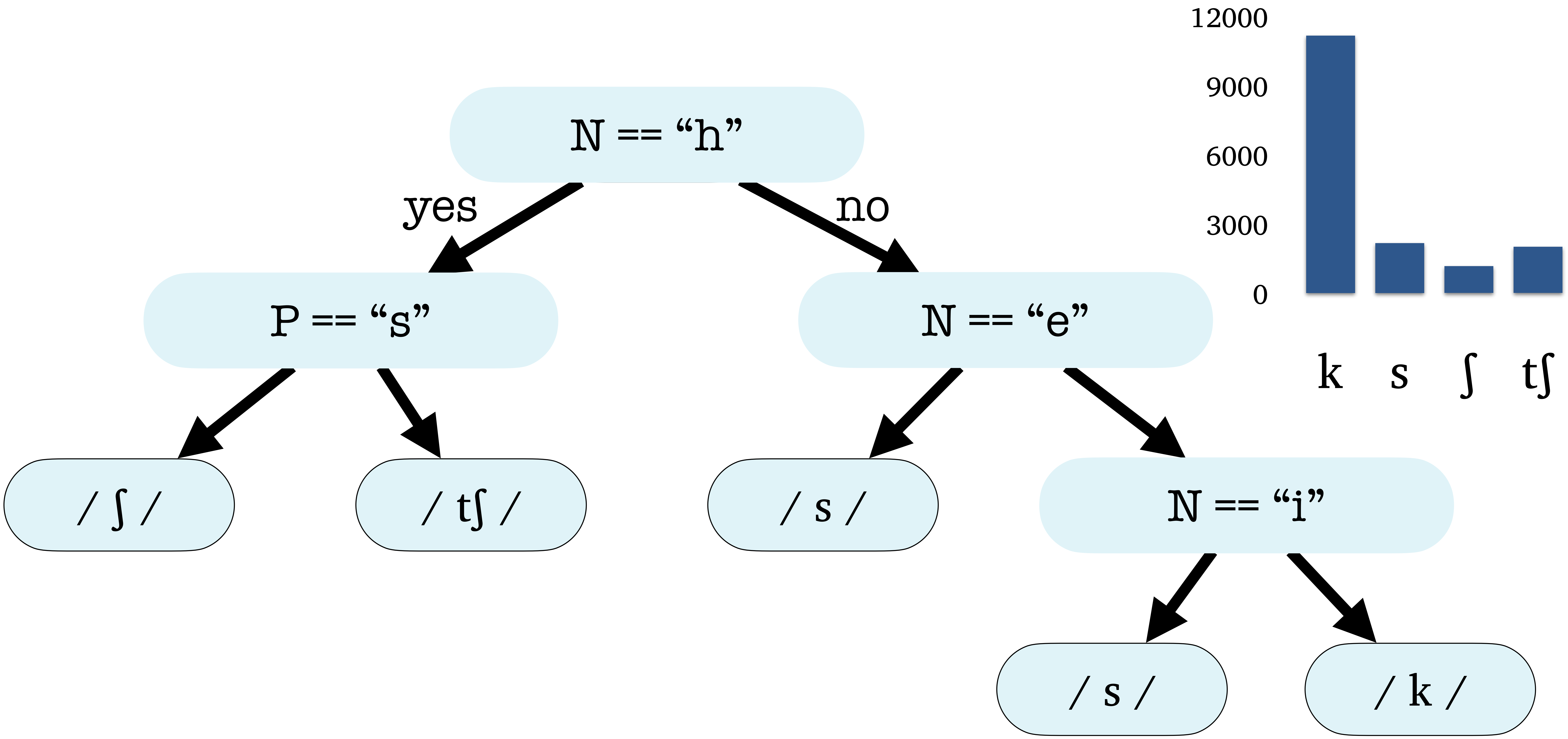


N == “h”
yes
no
/ tʃ /
/ k /
12000
9000
6000
3000
0
k s ʃ tʃ
1800
1350
900
450
0
k s ʃ tʃ
12000
9000
6000
3000
0
k s ʃ tʃ

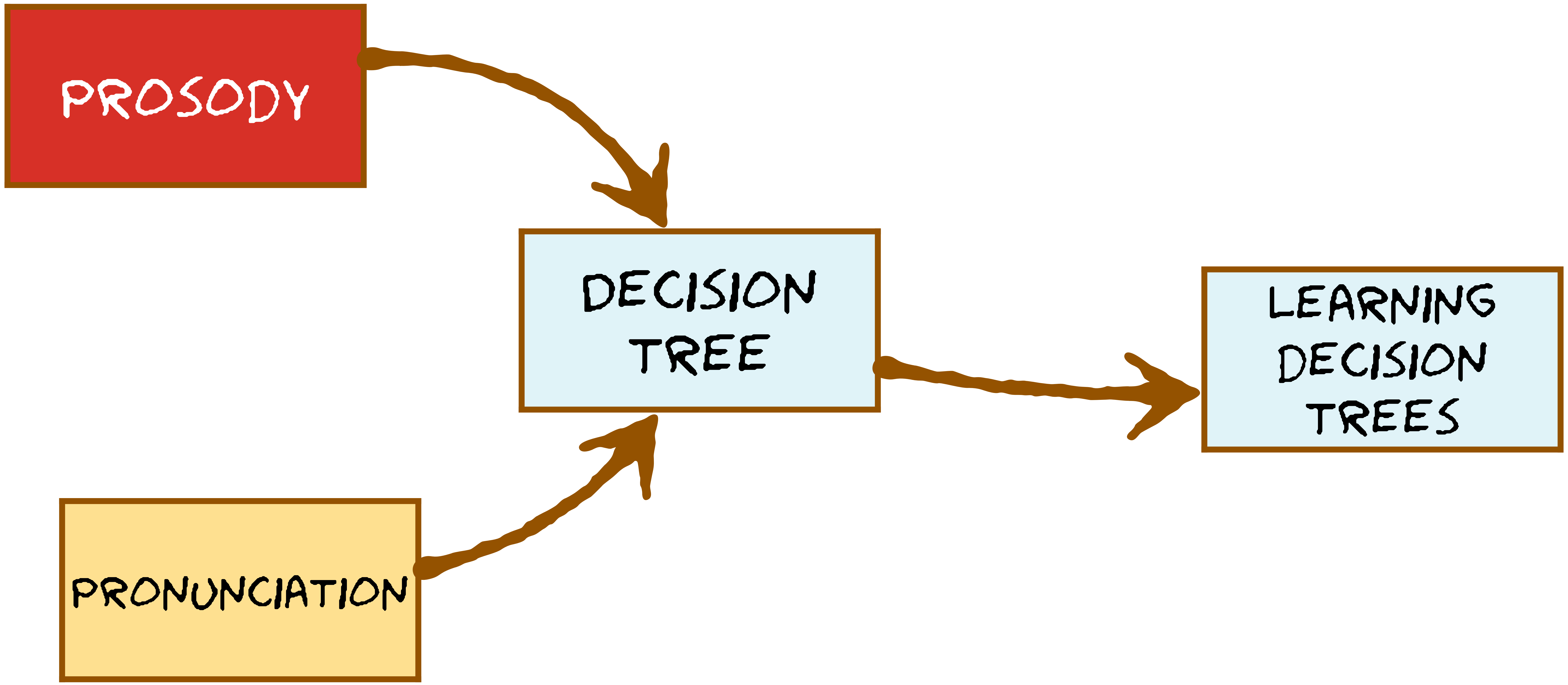

N == "h"
yes
no
P == "s"
/ k /
12000
9000
6000
3000
0
900
675
450
225
1800
1350
k
s
ʃ
tʃ


N == “h”
yes
no
P == “s”
/ k /
/ ʃ /
/ tʃ /
12000
9000
6000
3000
0
k
s
ʃ
tʃ
12000
9000
6000
3000
0
k
s
ʃ
tʃ


N == “h”
yes
no
P == “s”
N == “e”
/ ʃ /
/ tʃ /
/ s /
/ k /
12000
9000
6000
3000
0
k
s
ʃ
tʃ

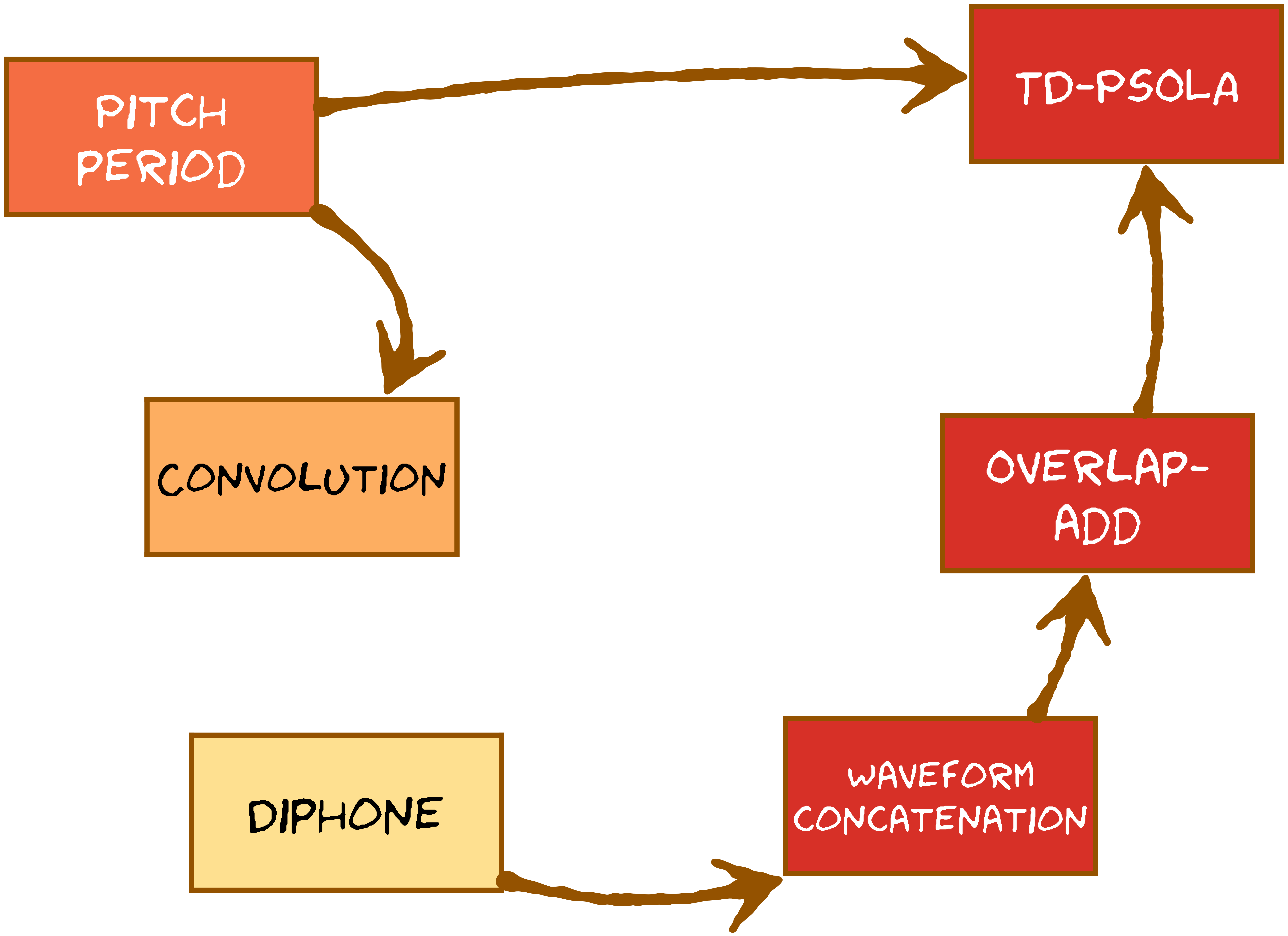

N == “h”
yes
no
P == “s”
N == “e”
/ ʃ /
/ tʃ /
/ s /
N == “i”
/ s /
/ k /
12000
9000
6000
3000
0
k
s
ʃ
tʃ

What you can learn next



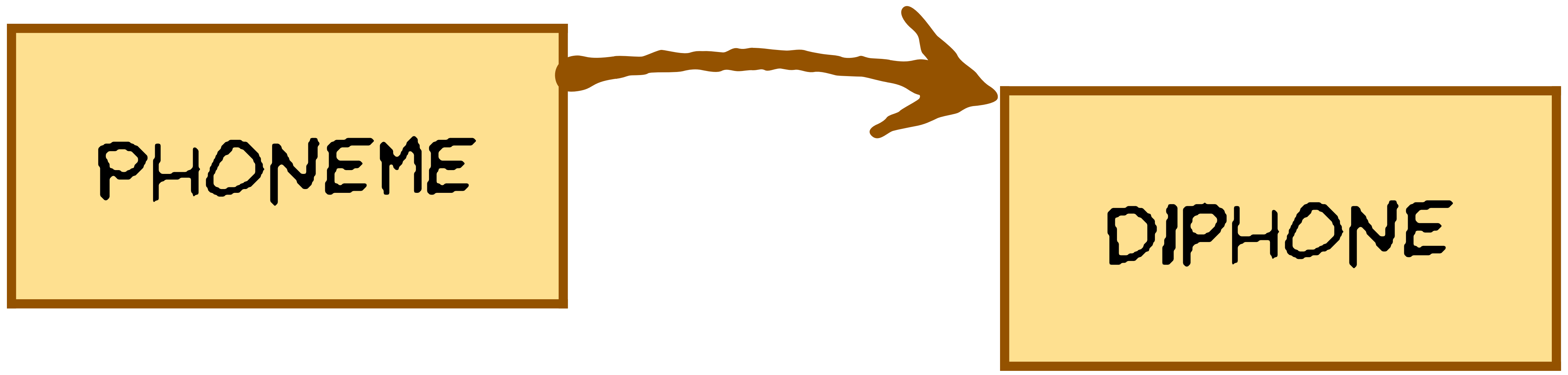
Module 5

Waveform generation

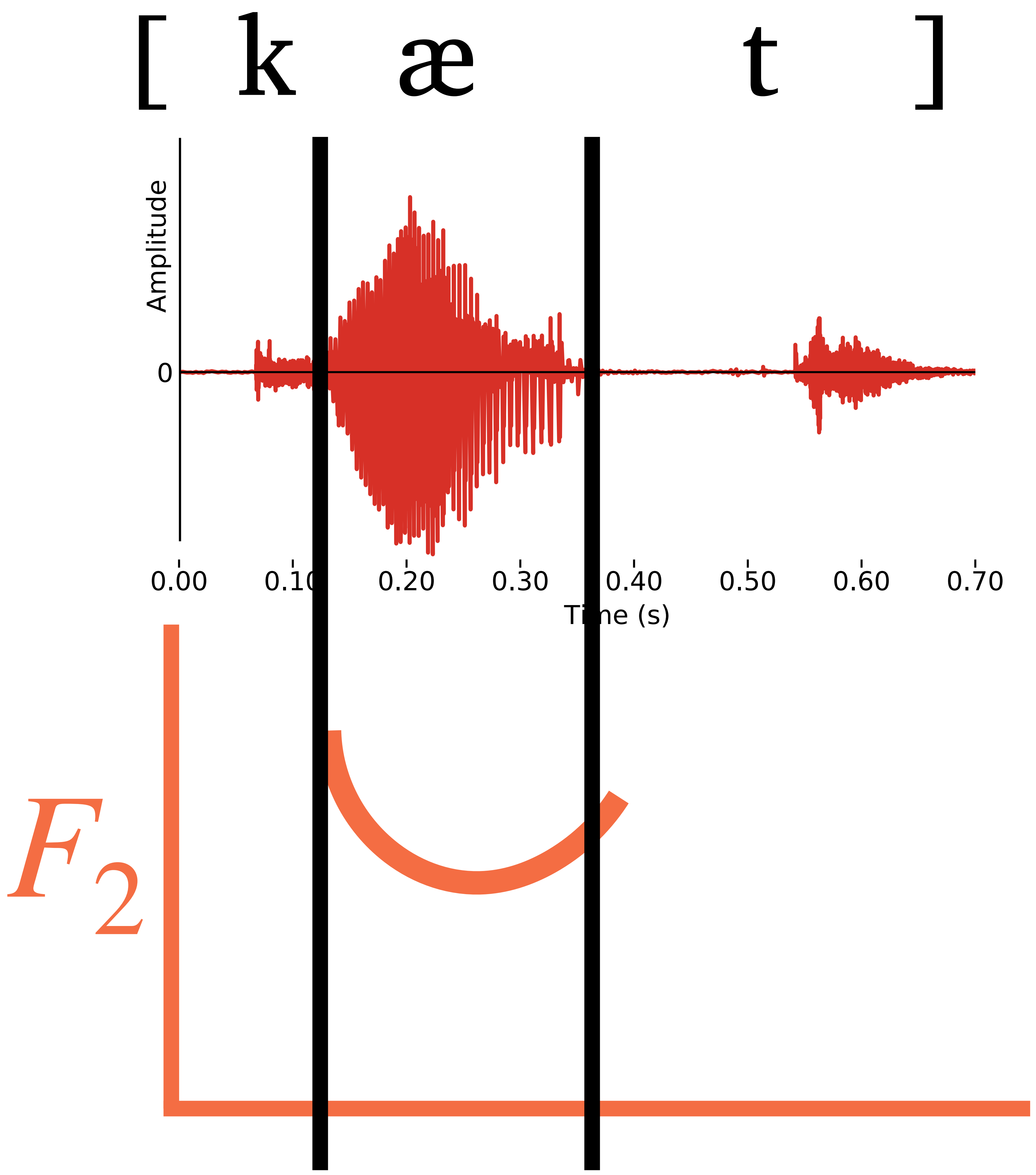

PITCH PERIOD
TD-PSOLA
CONVOLUTION
OVERLAP-ADD
DIPHONE
WAVEFORM CONCATENATION

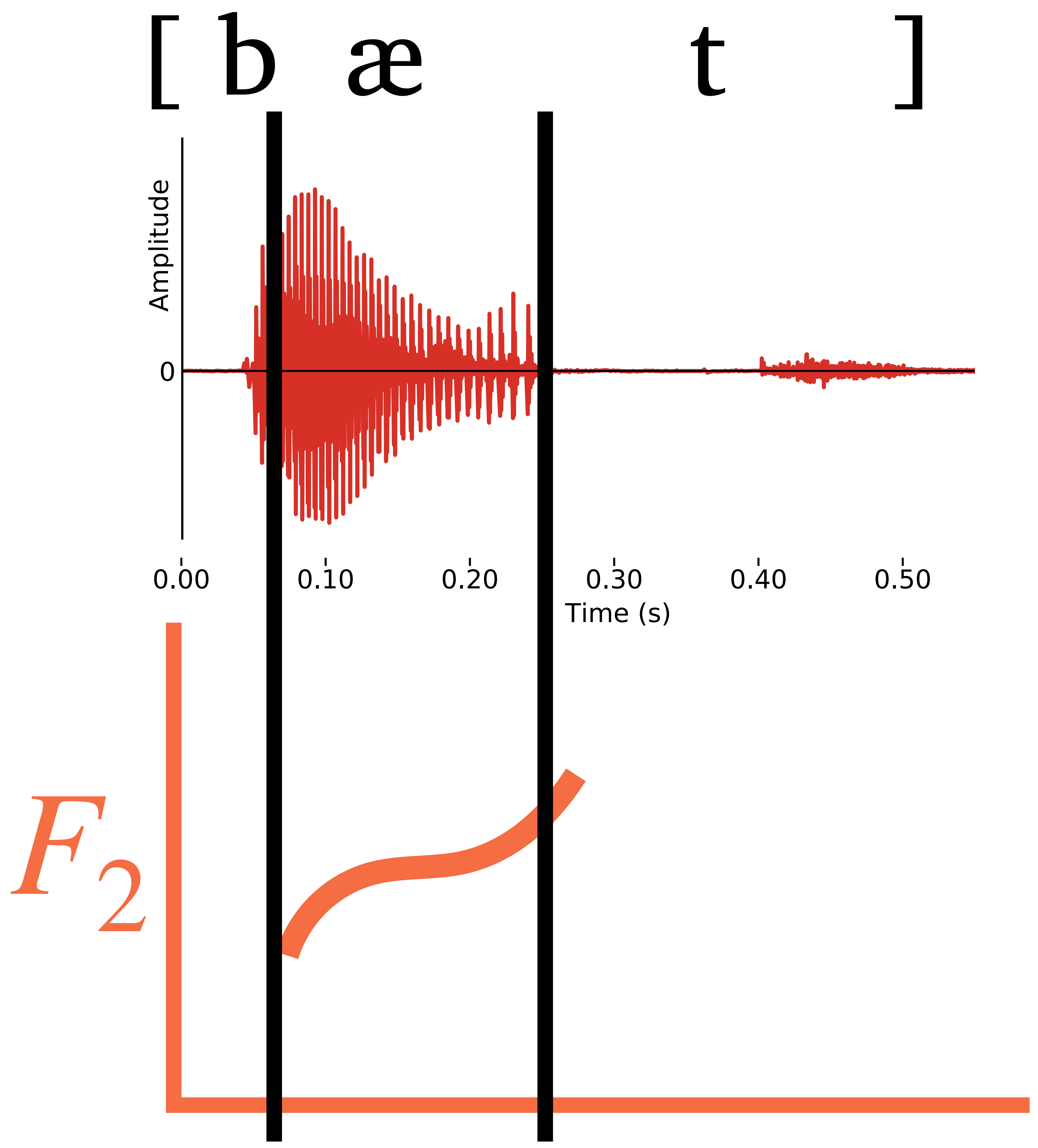
DIPHONE

What you need to know already

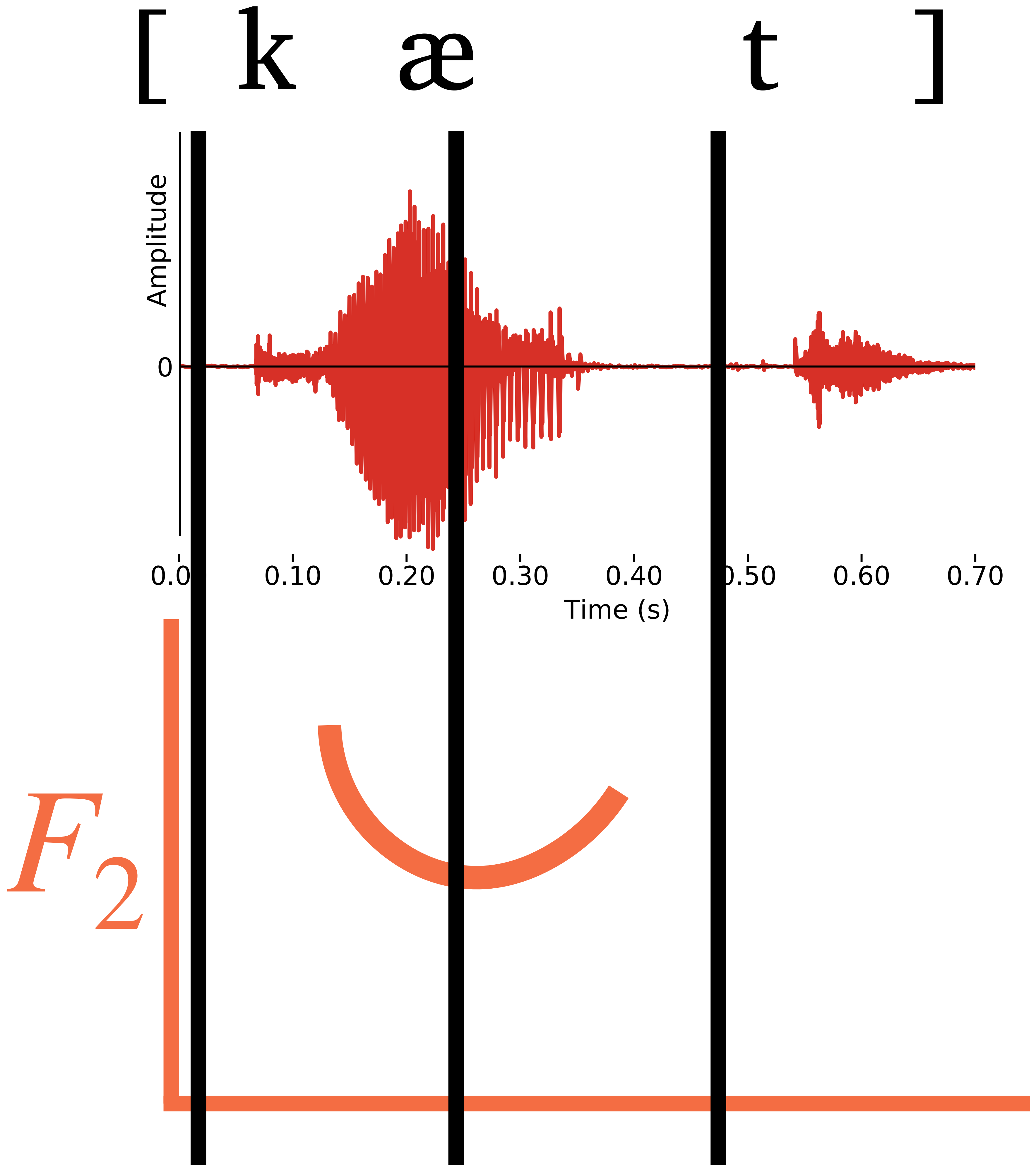


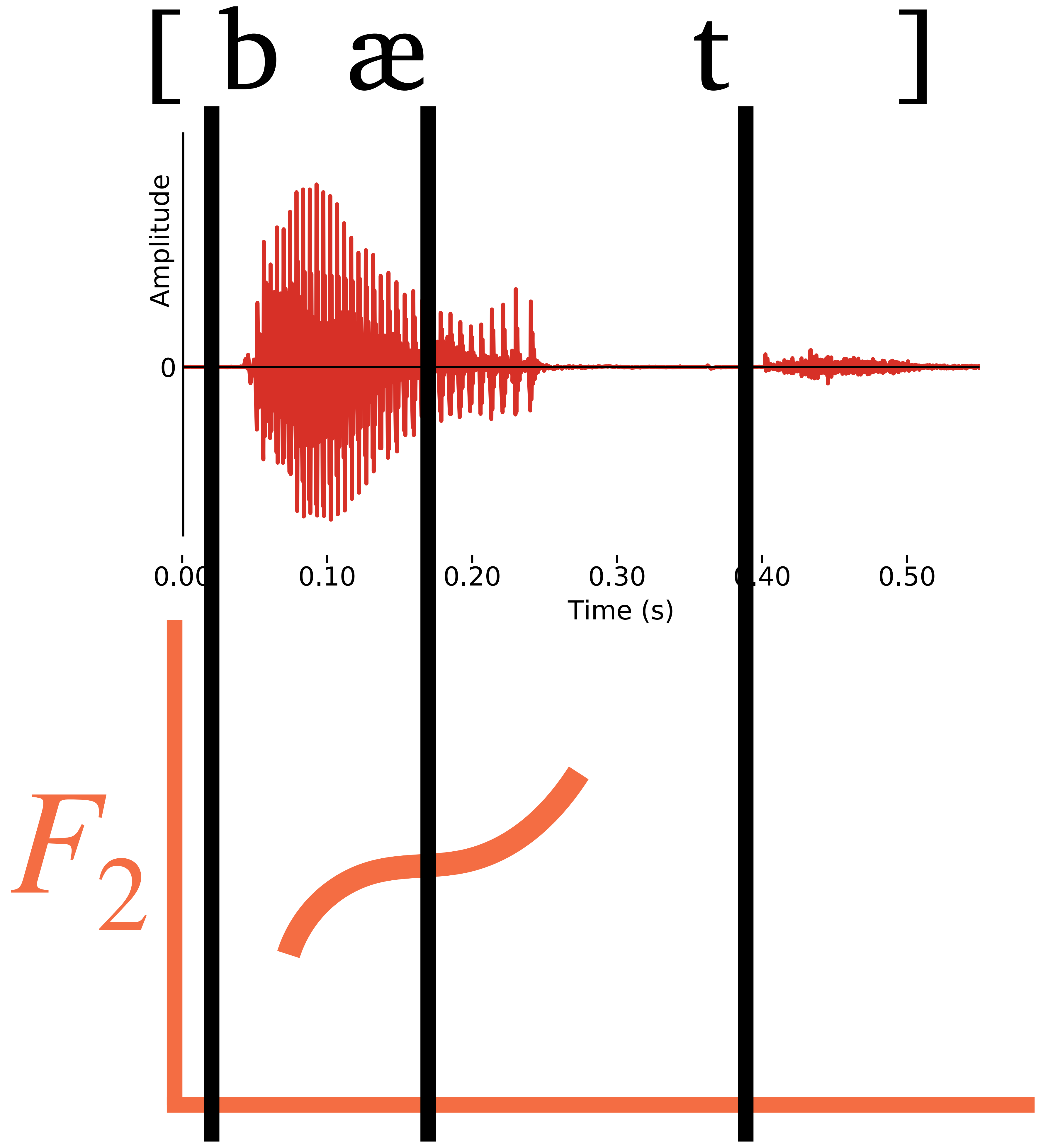
Co-articulation





Co-articulation





Diphones

Phones	[	k	æ	t	]
Diphones	sil - k	k - æ	æ - t	t - sil	

Phones	[	b	æ	t	]
Diphones	sil - b	b - æ	æ - t	t - sil	

Synthesising new utterances from a database of diphone units

sil-f f-æ æ-t t-k k-æ æ-t t-sil

sil-æ æ-n n-ɪ ɪ-m m-ə ə-l l-sil

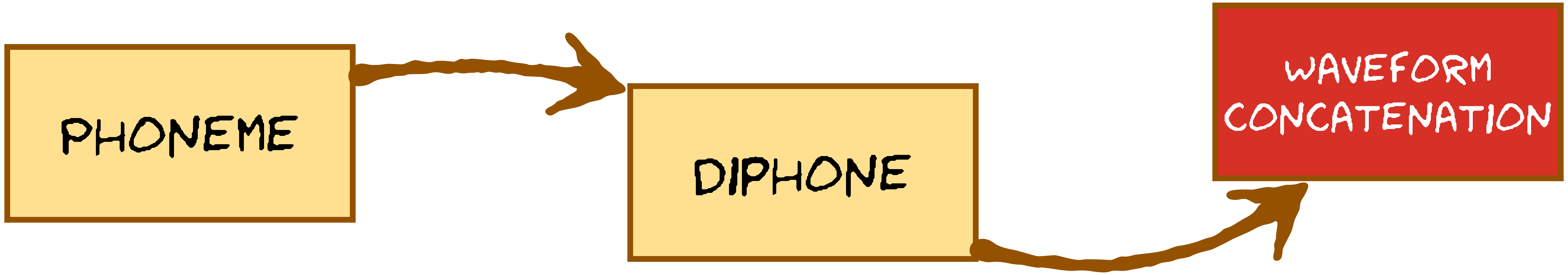
sil-h h-ɒ ɒ-t t-p p-æ æ-n n-sil

sil-f f-æ æ-n n-sil

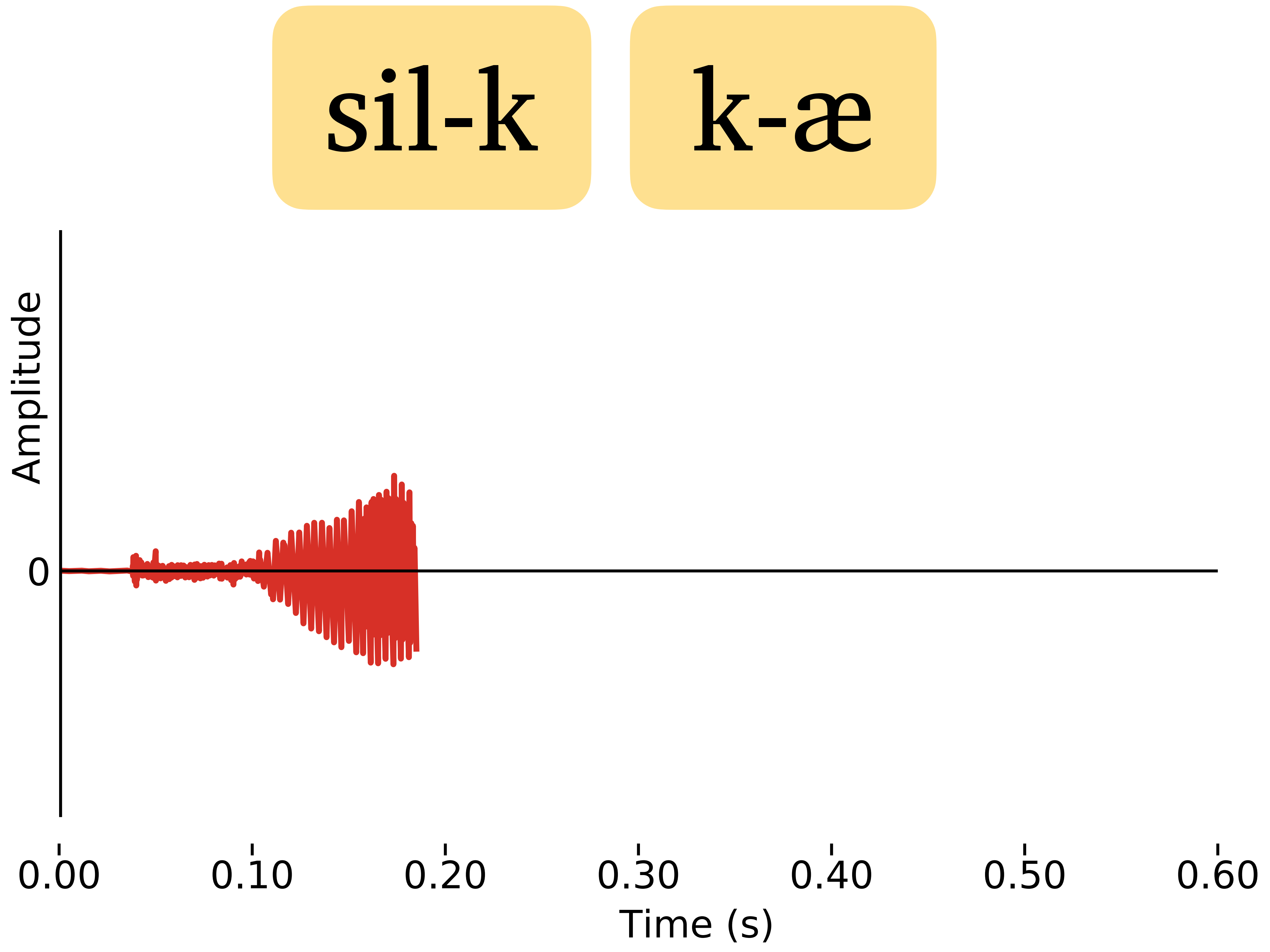
What you can learn next

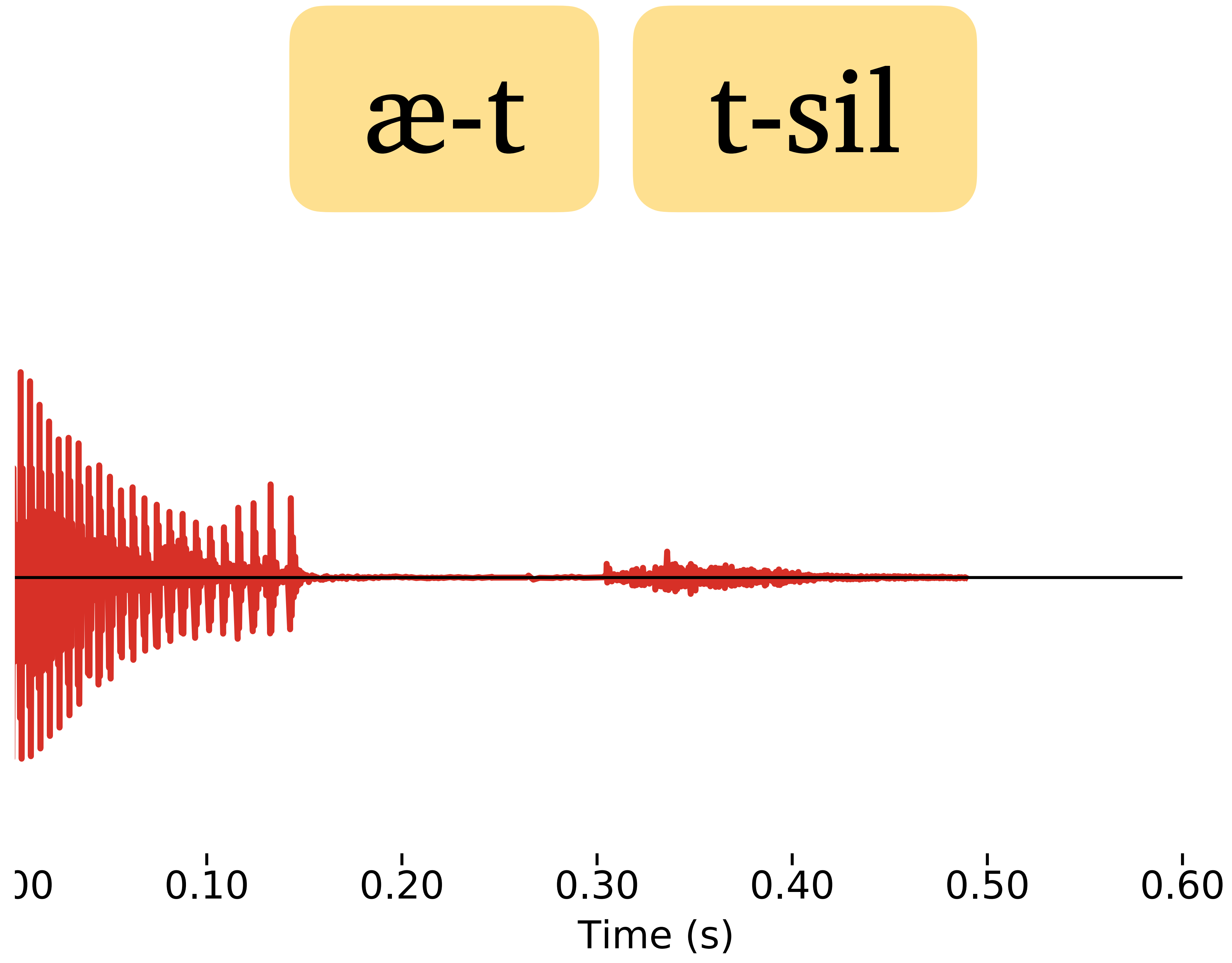
WAVEFORM CONCATENATION

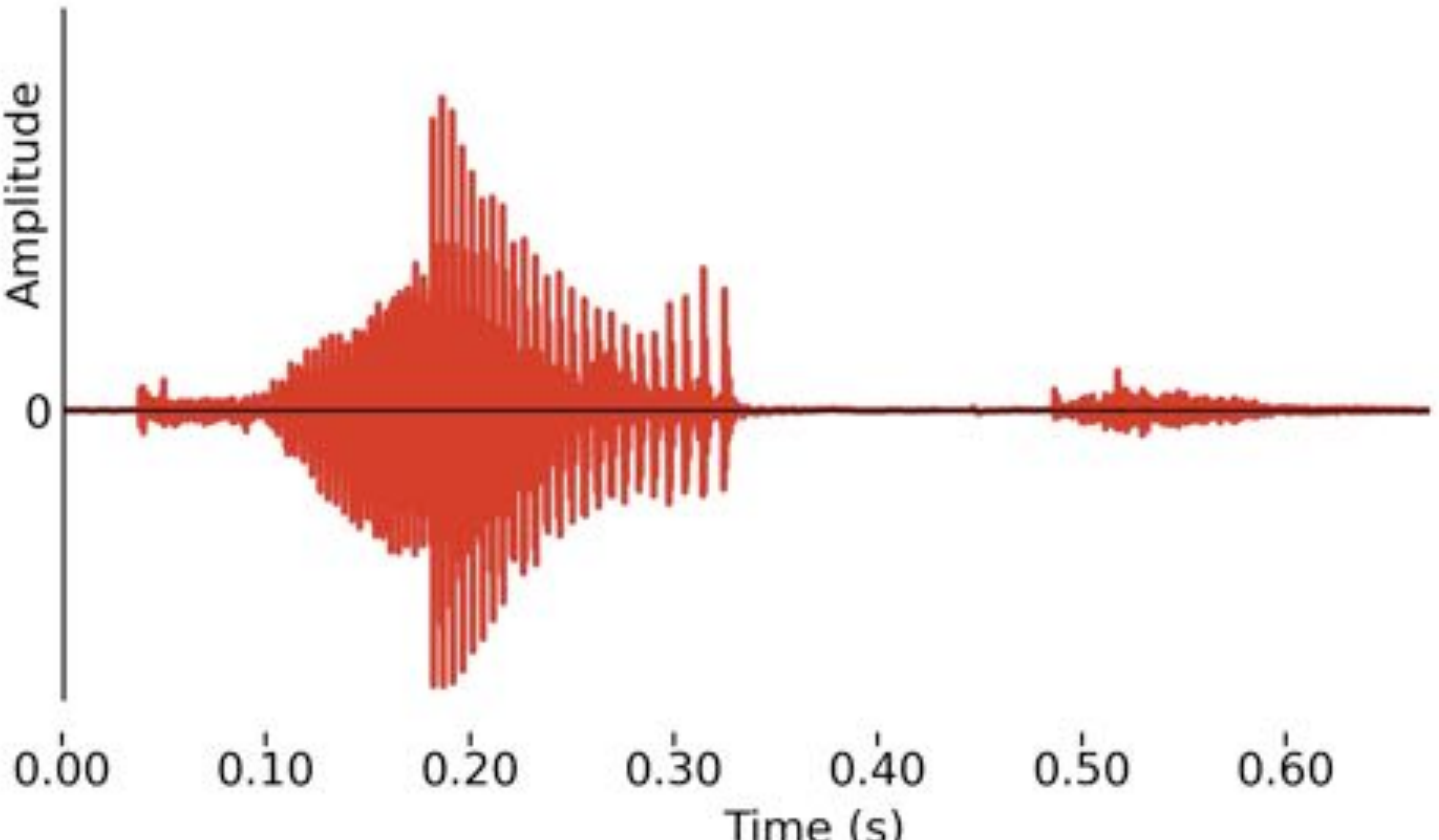
What you need to know already

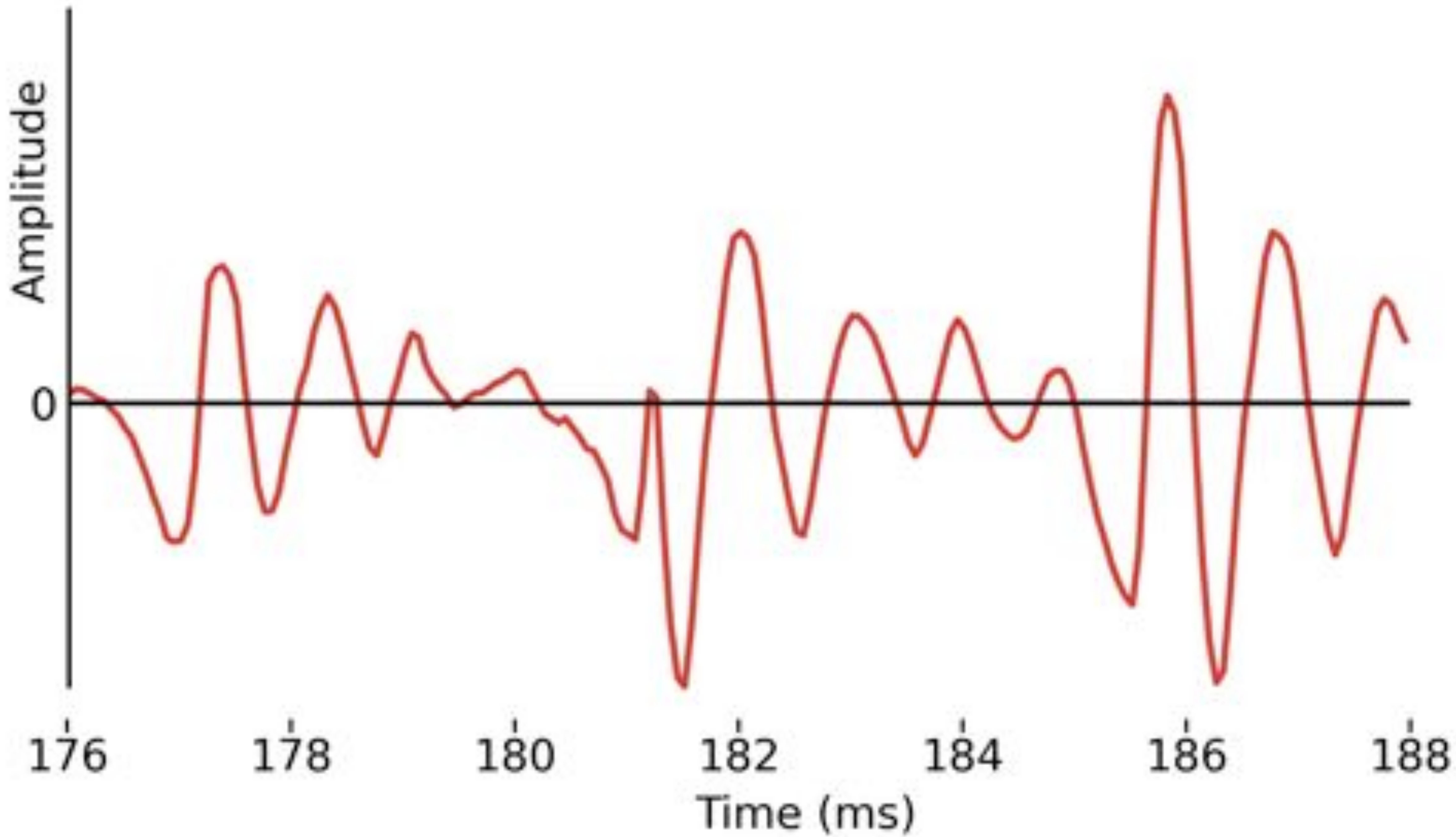


Naive concatenation



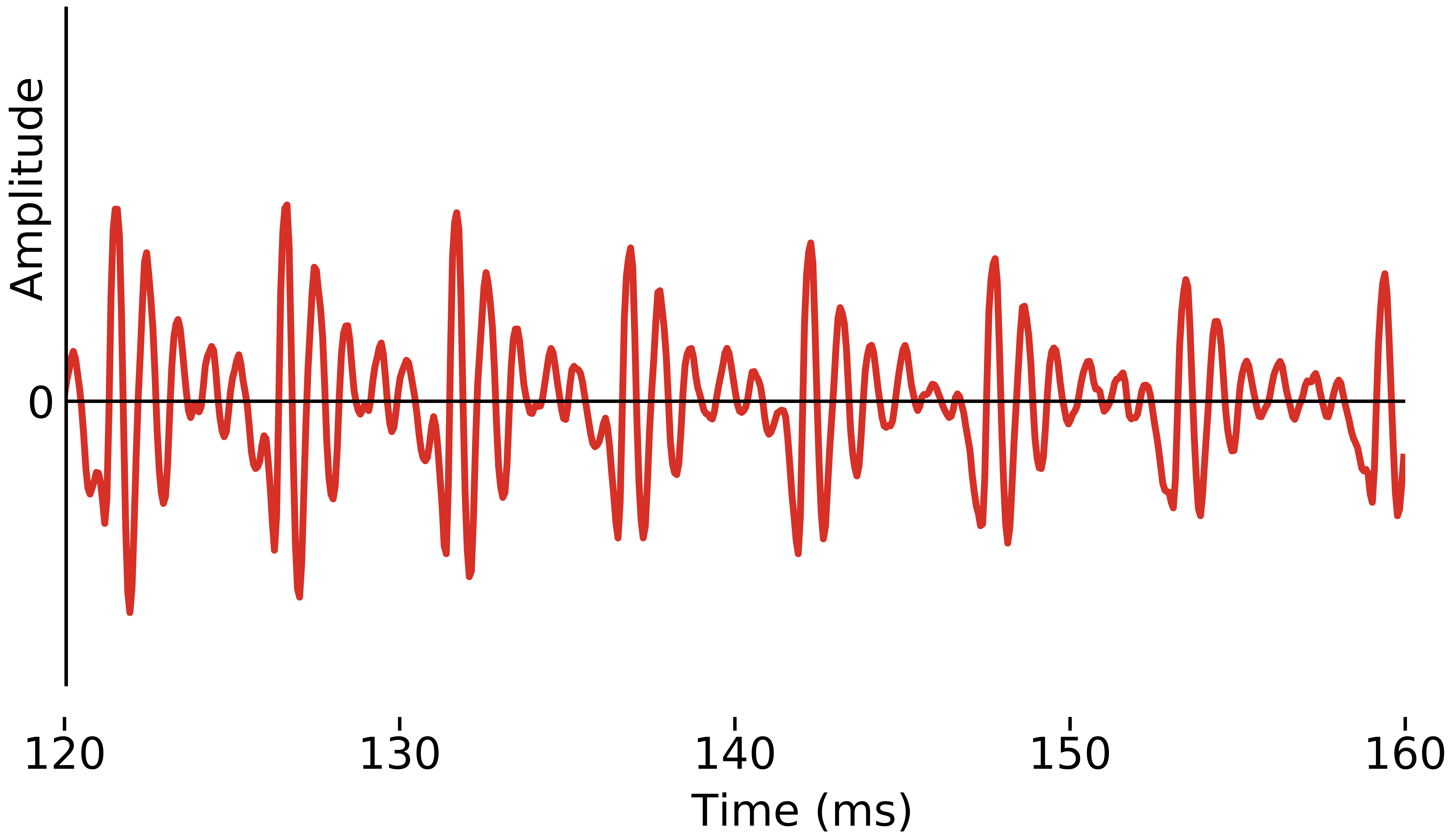



Amplitude
0
0.00
0.10
0.20
0.30
0.40
0.50
0.60
Time (s)



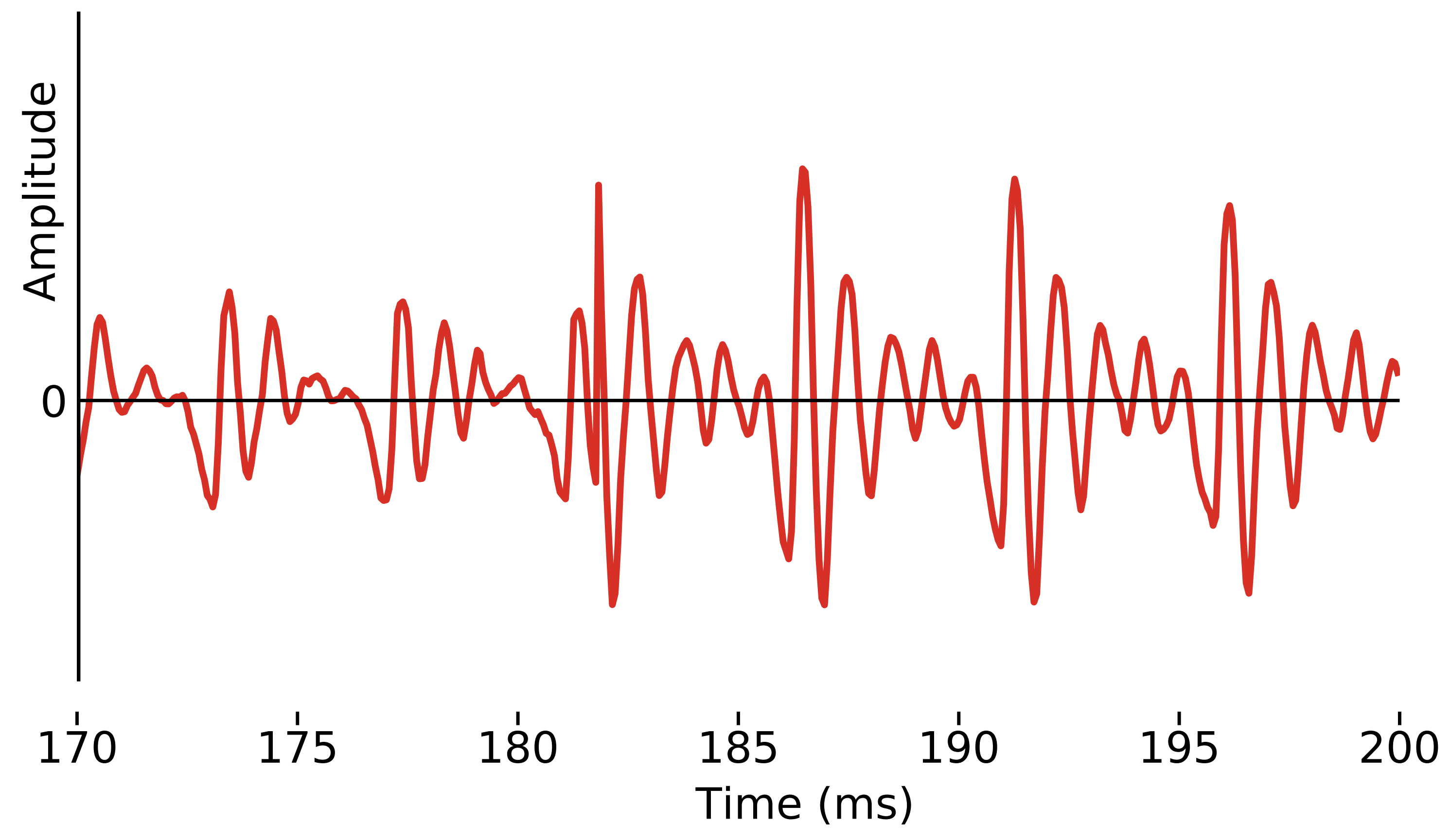
Amplitude
0
176
178
180
182
184
186
188
Time (ms)

Epochs (pitch marks)

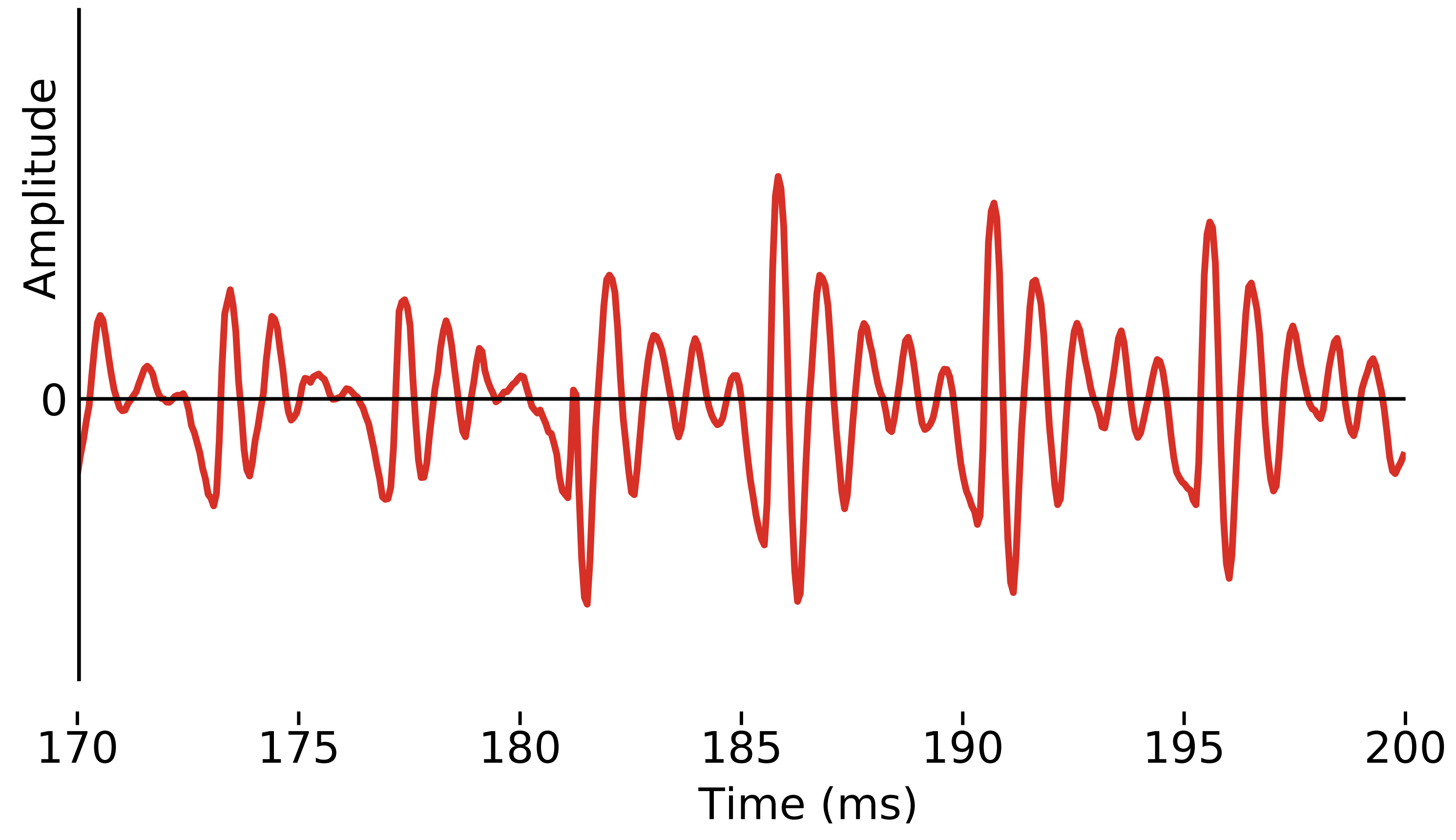


Concatenating waveforms (three ways)

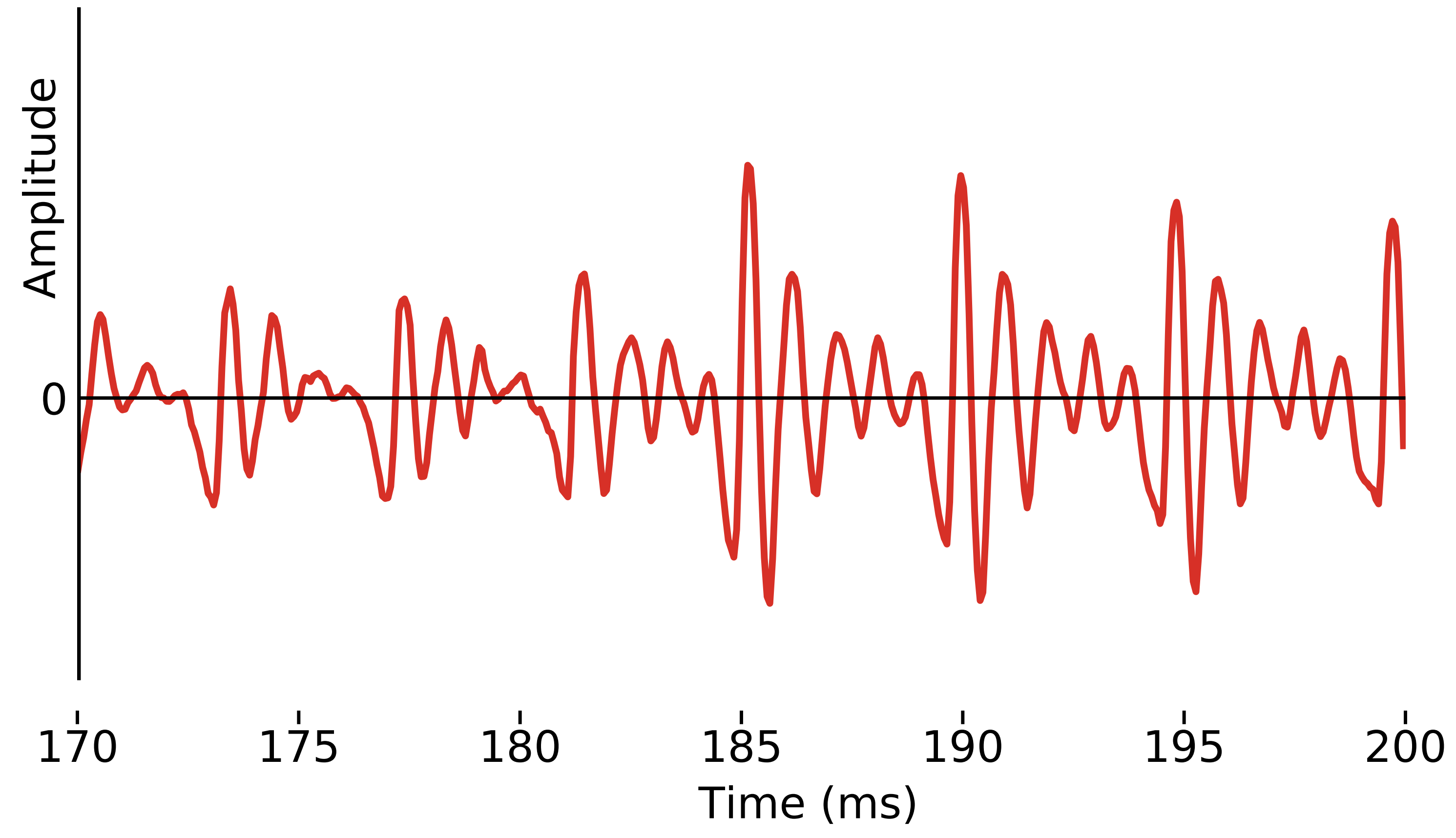
Naive



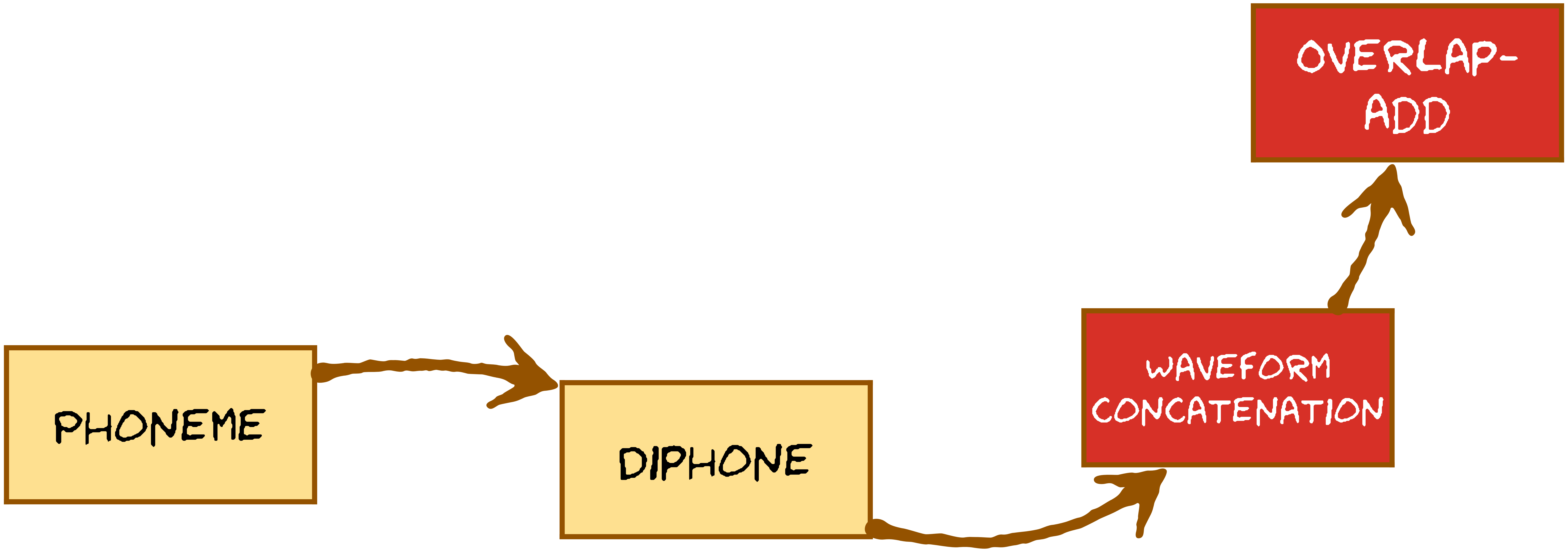
Zero-crossing



Pitch-synchronous

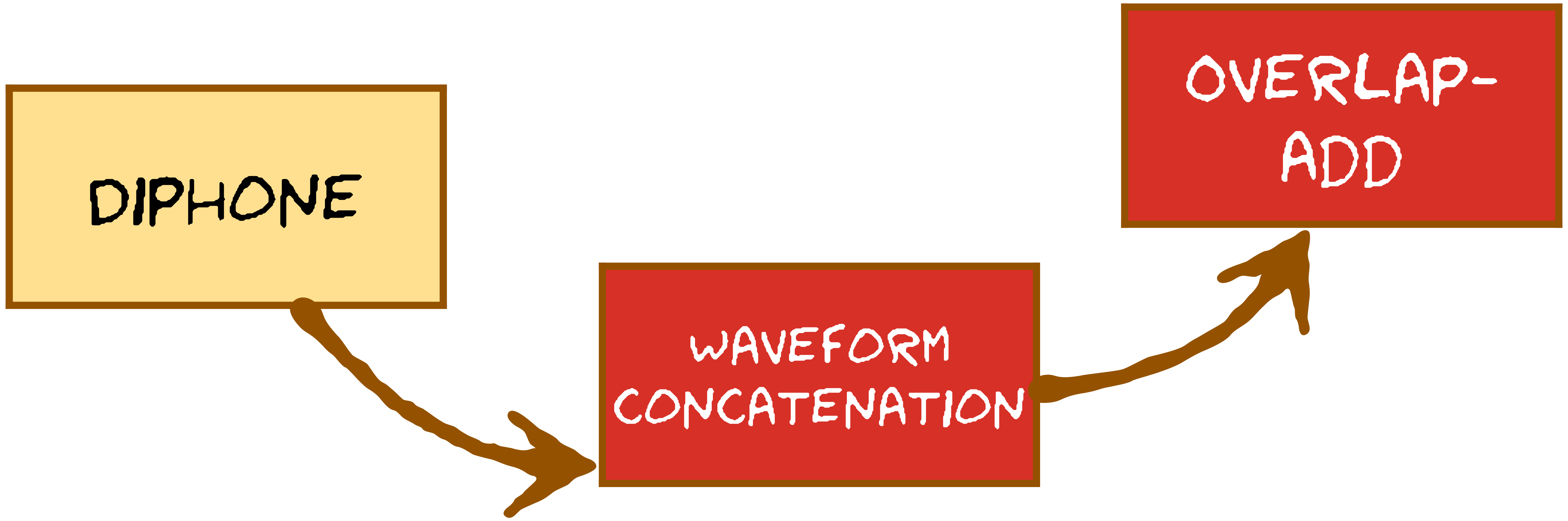


What you can learn next

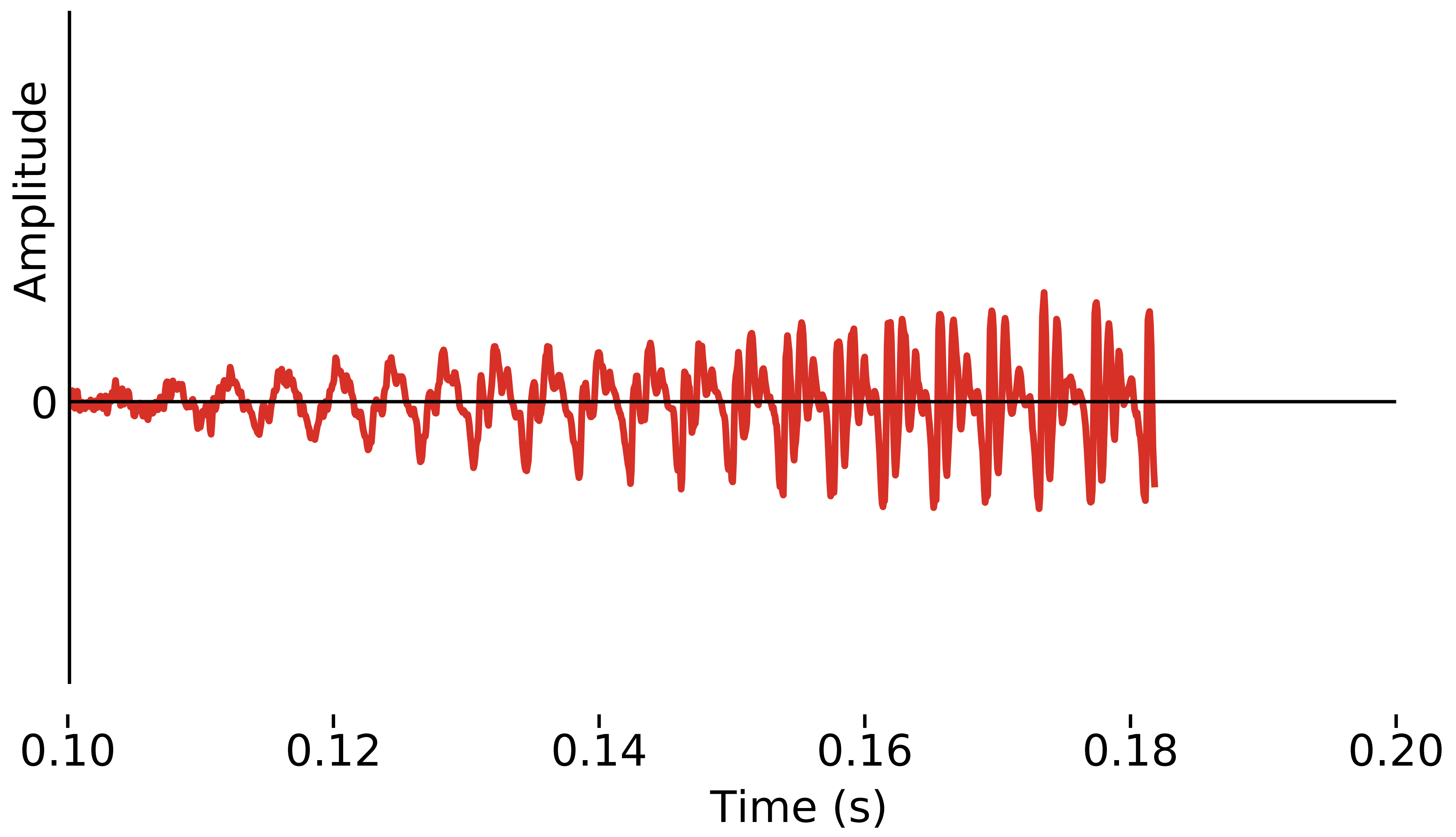


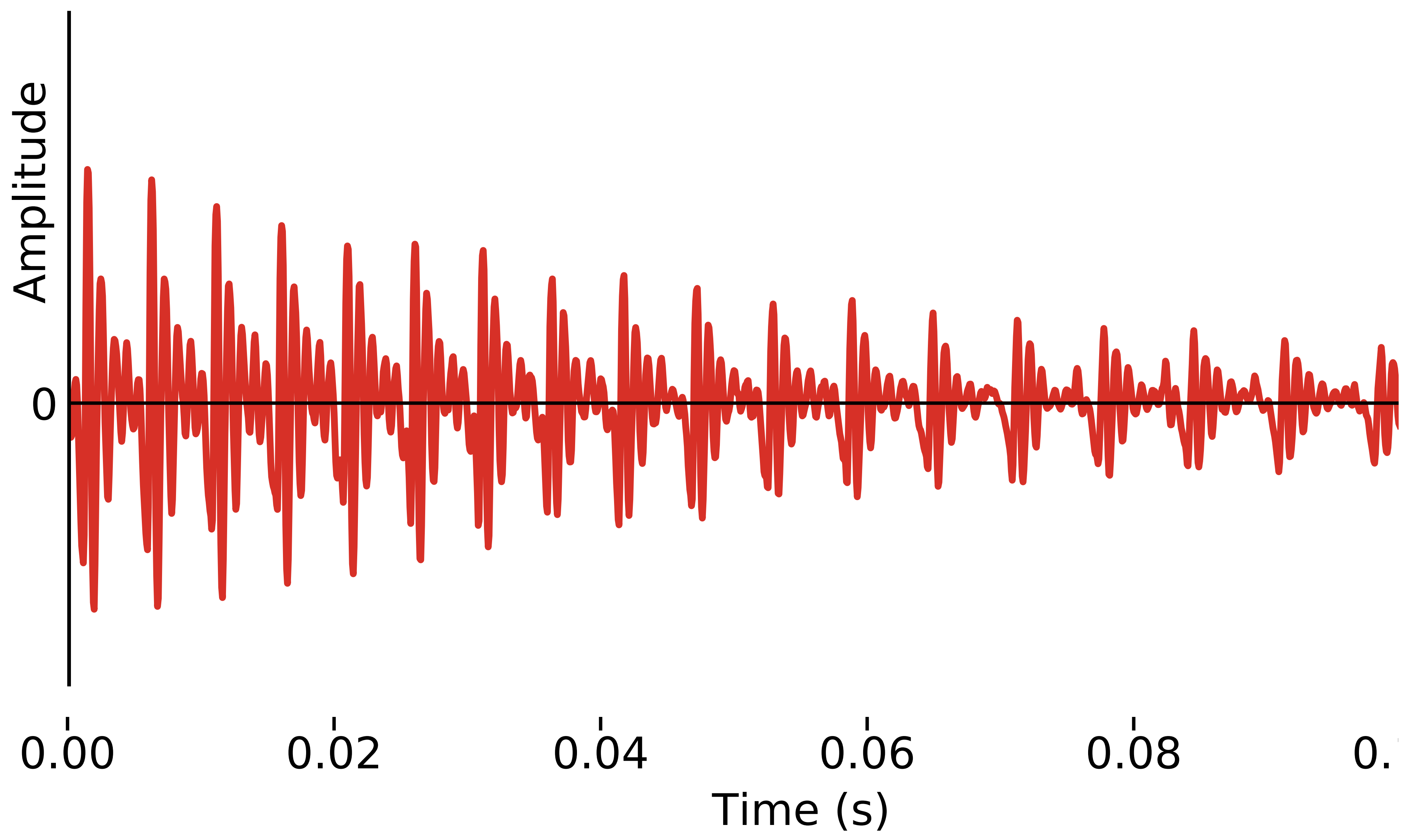
OVERLAP-ADD

What you need to know already

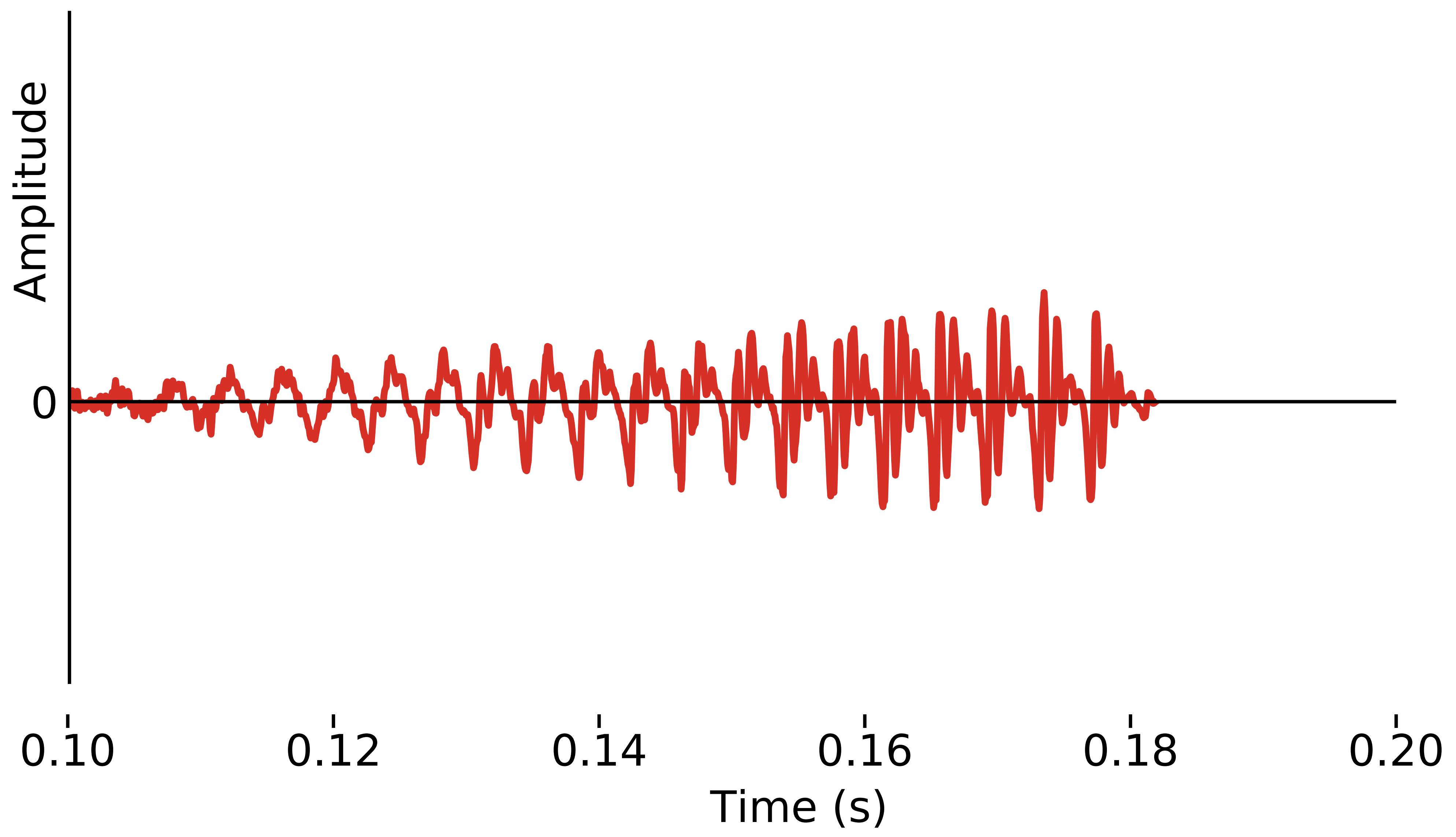


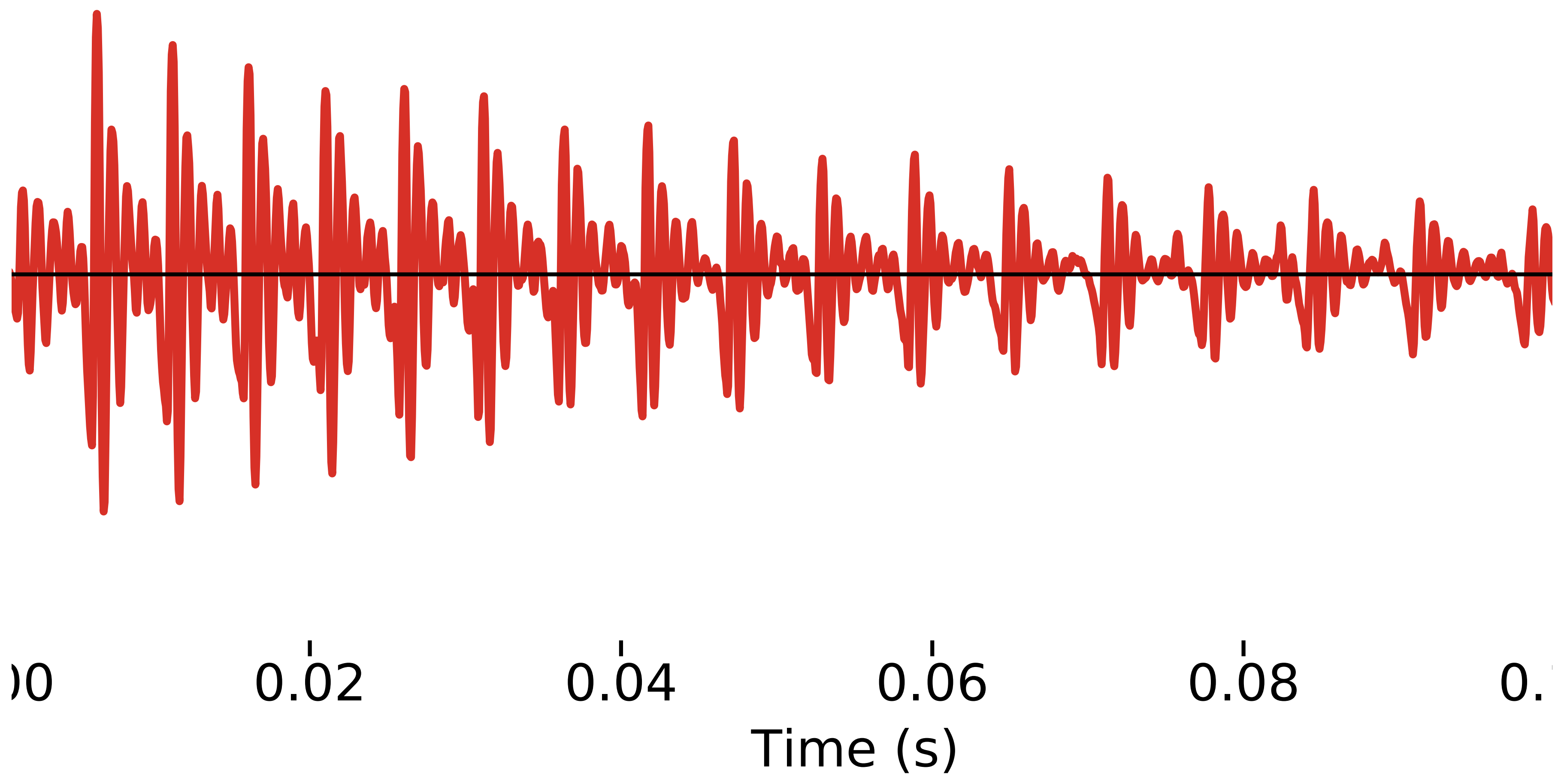
Overlap-add (or, cross-fade)



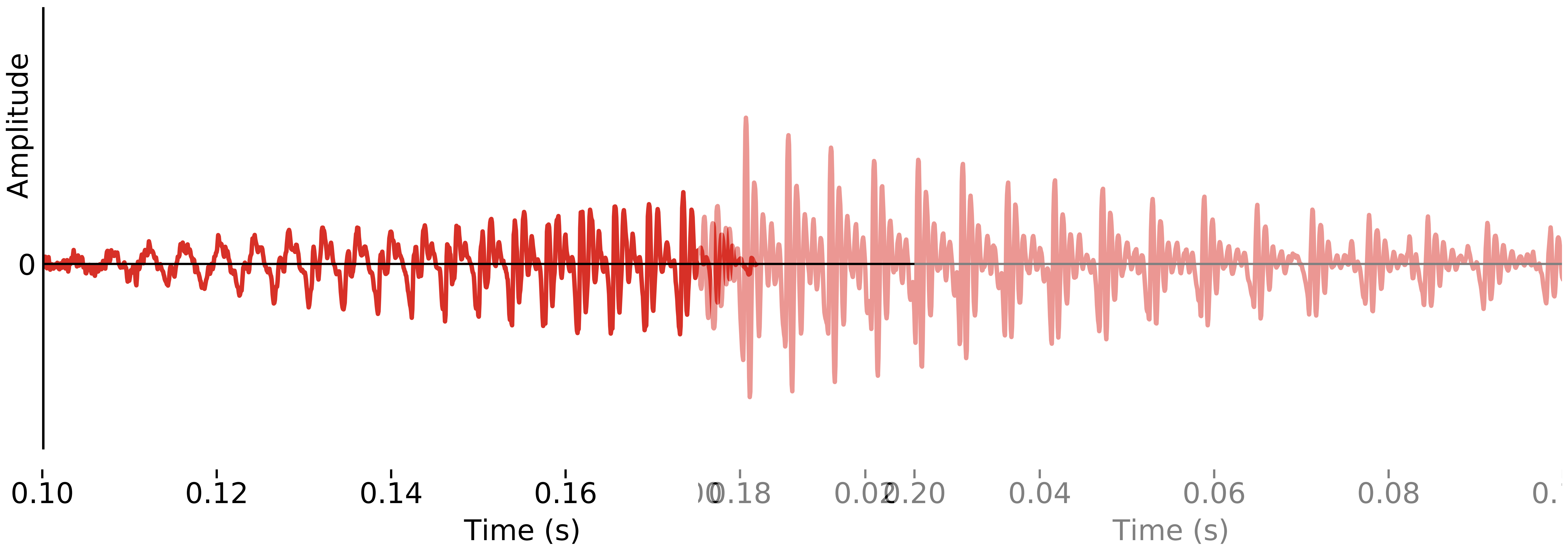


Overlap-add (or, cross-fade)

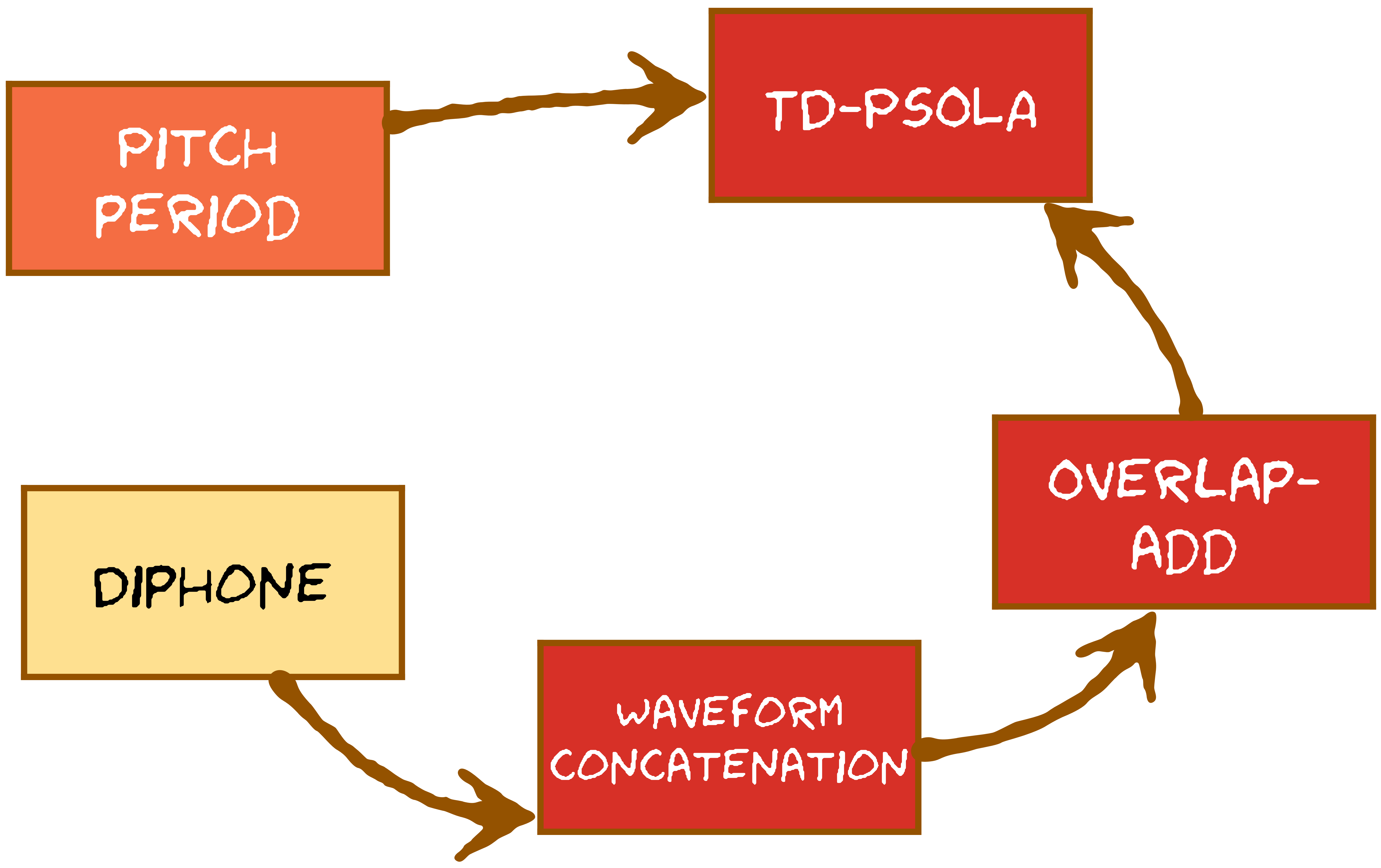




Overlap-add (or, cross-fade)

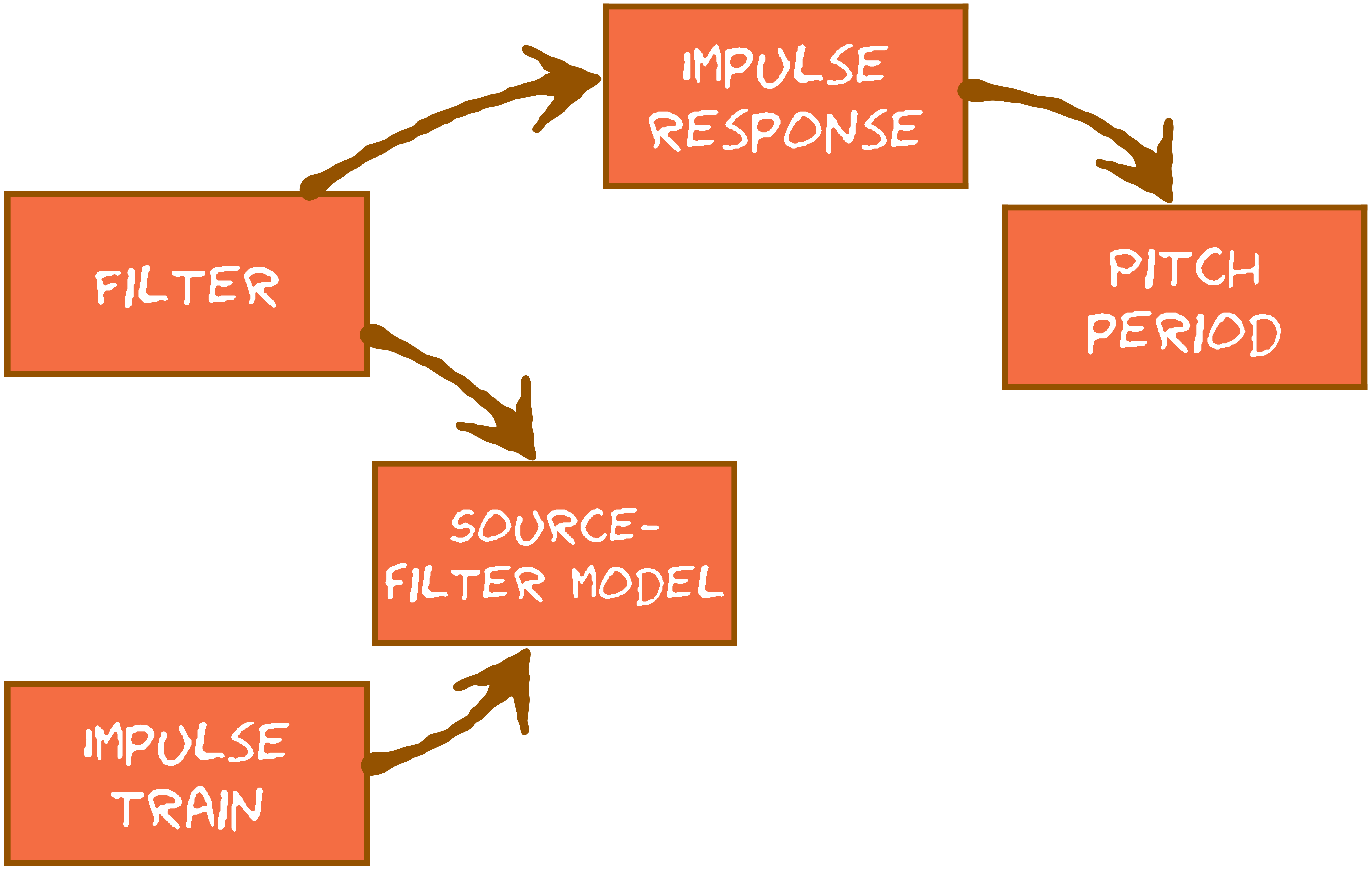


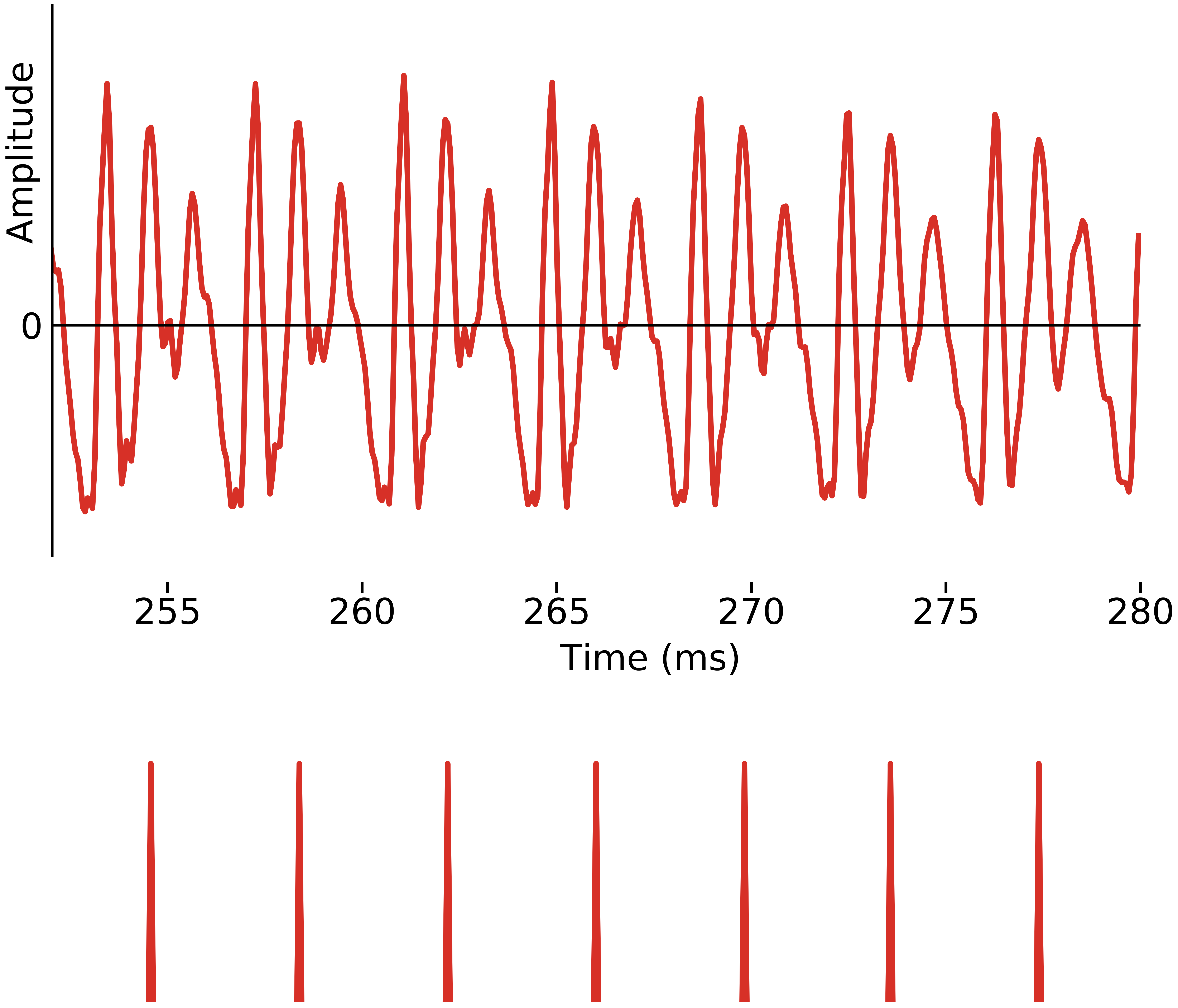
What you can learn next



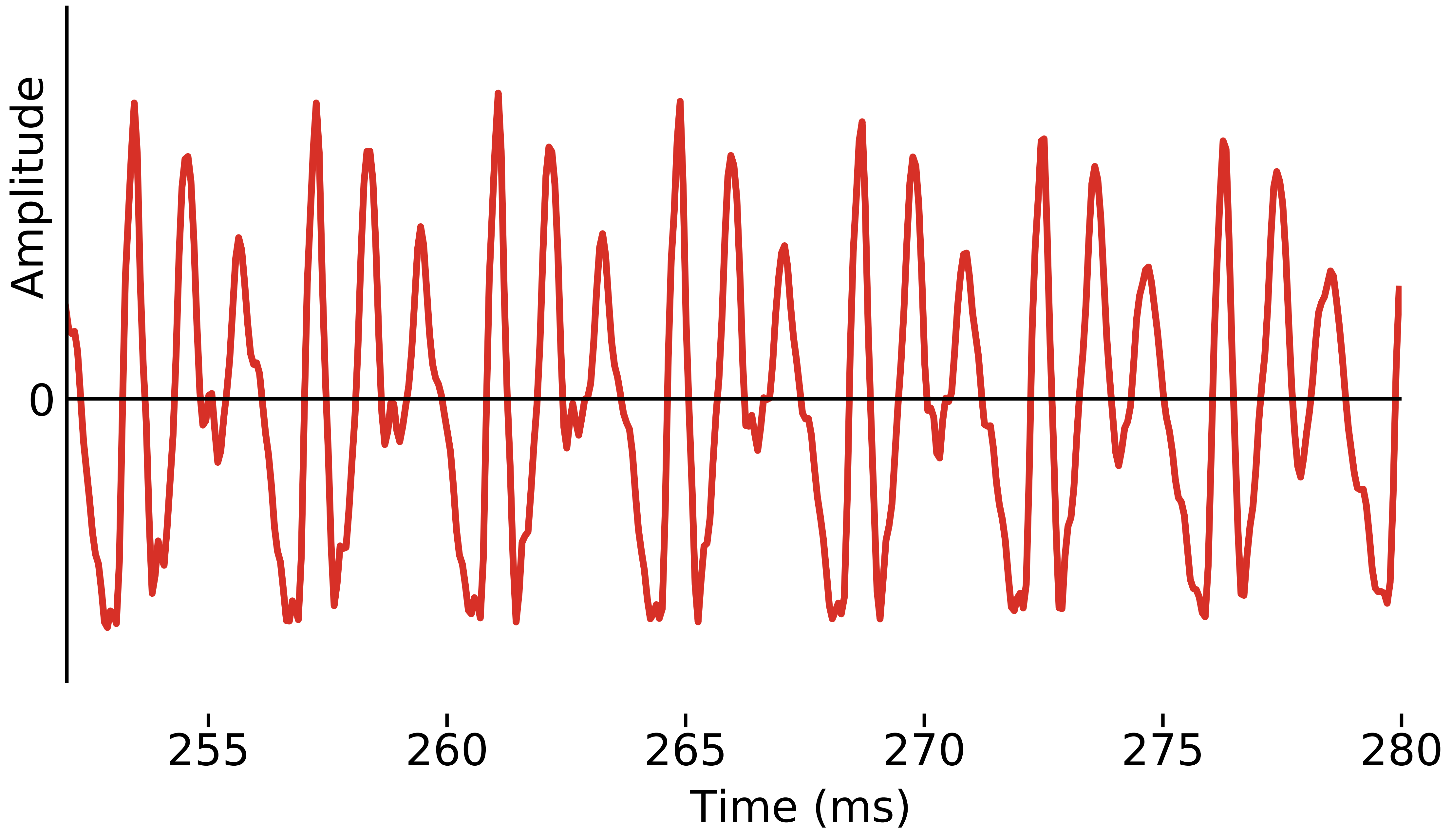
PITCH PERIOD

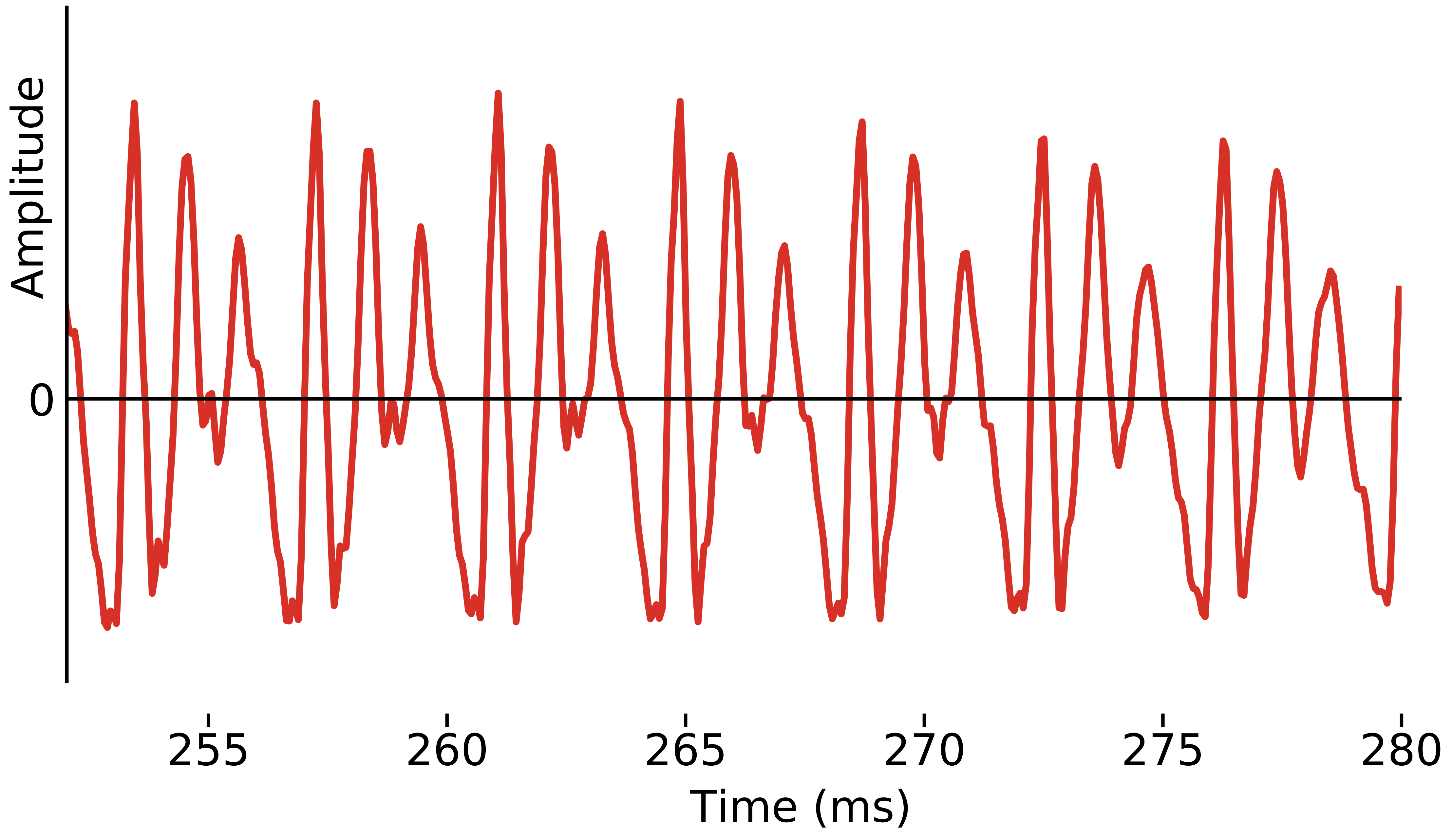
What you need to know already

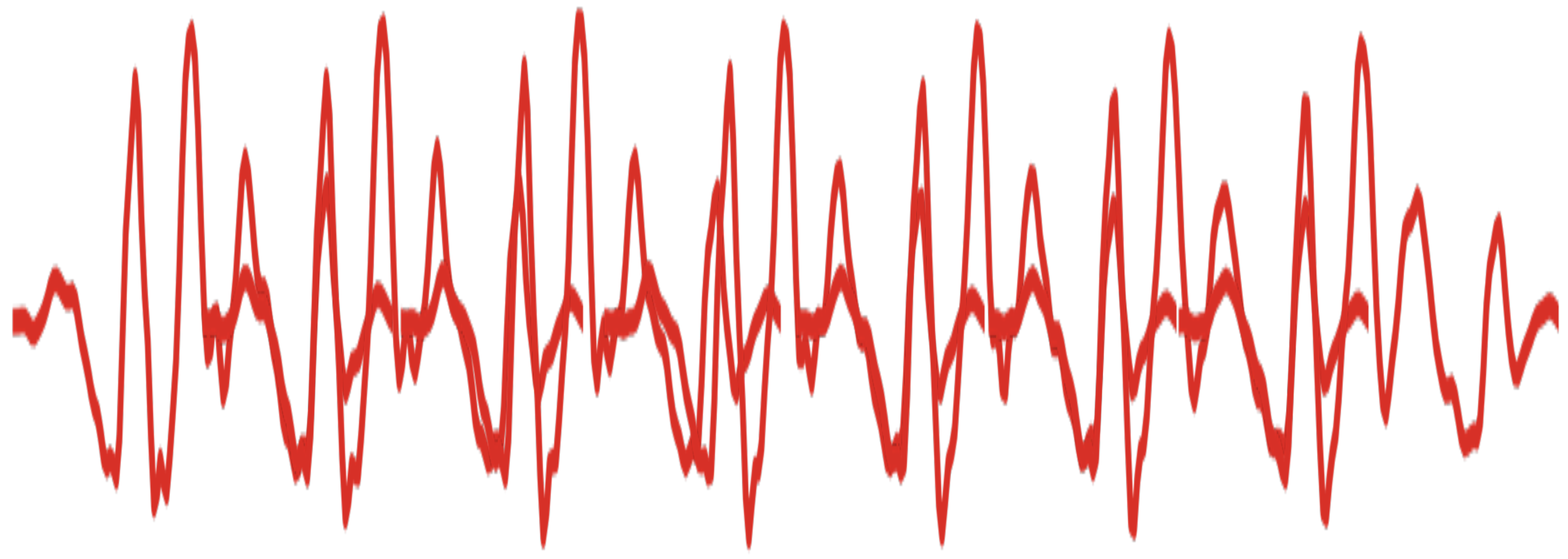




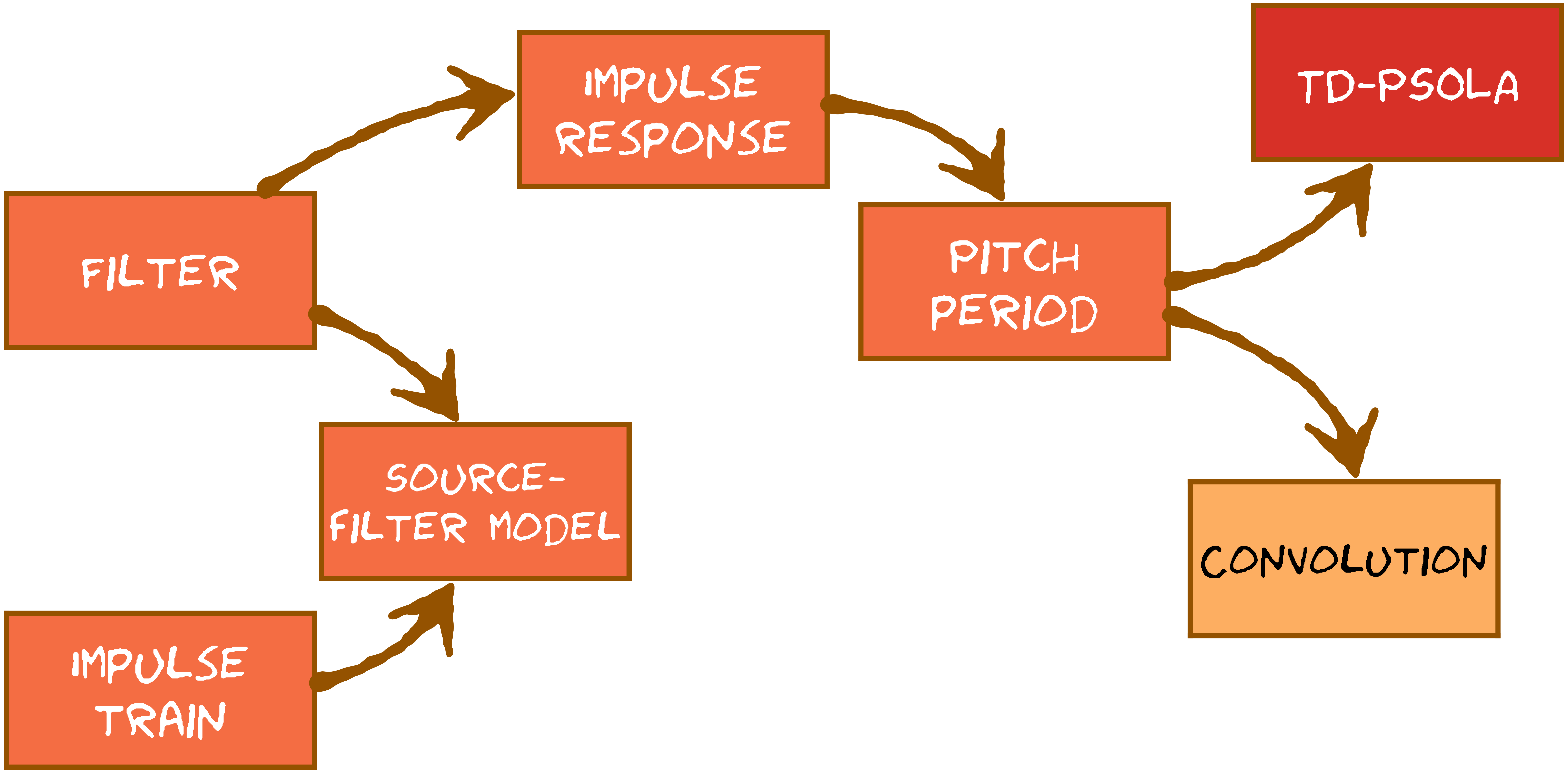
Amplitude
0
255
260
265
270
275
280
Time (ms)


Amplitude
0
255
260
265
270
275
280
Time (ms)


Amplitude
0
255
260
265
270
275
280
Time (ms)



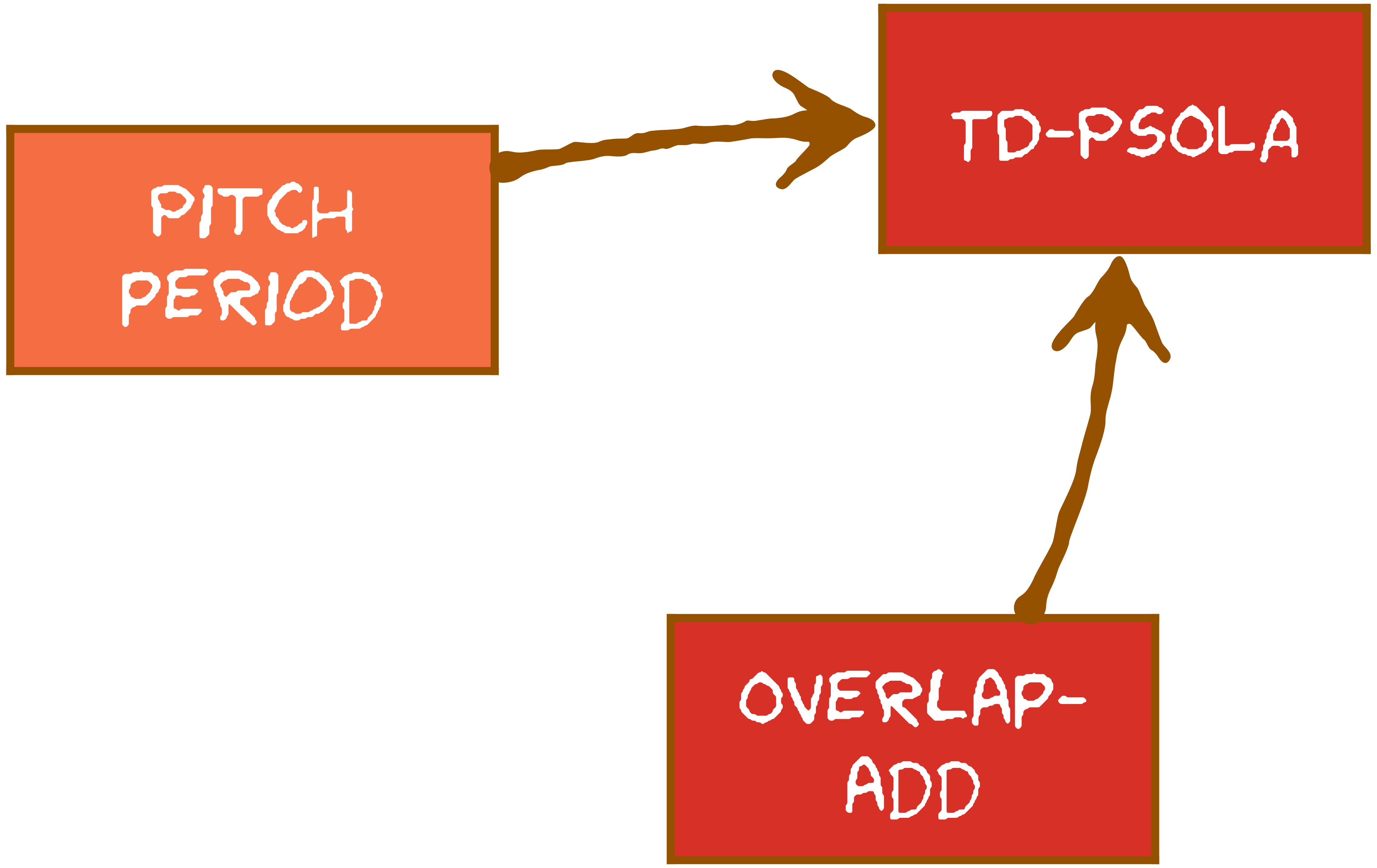
What you can learn next



TD-PSOLA

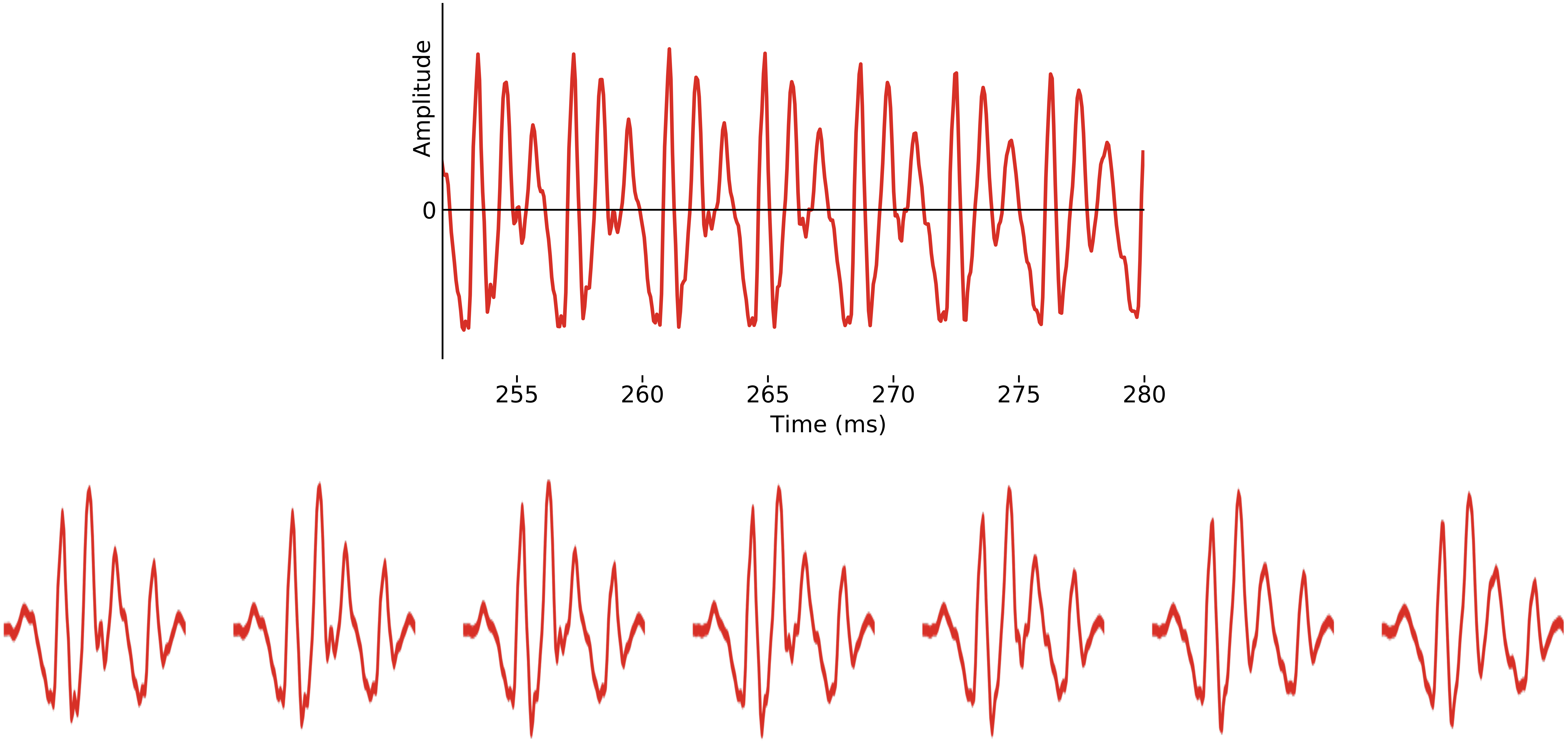
PERIODIC SIGNALS IN THE TIME DOMAIN

What you need to know already

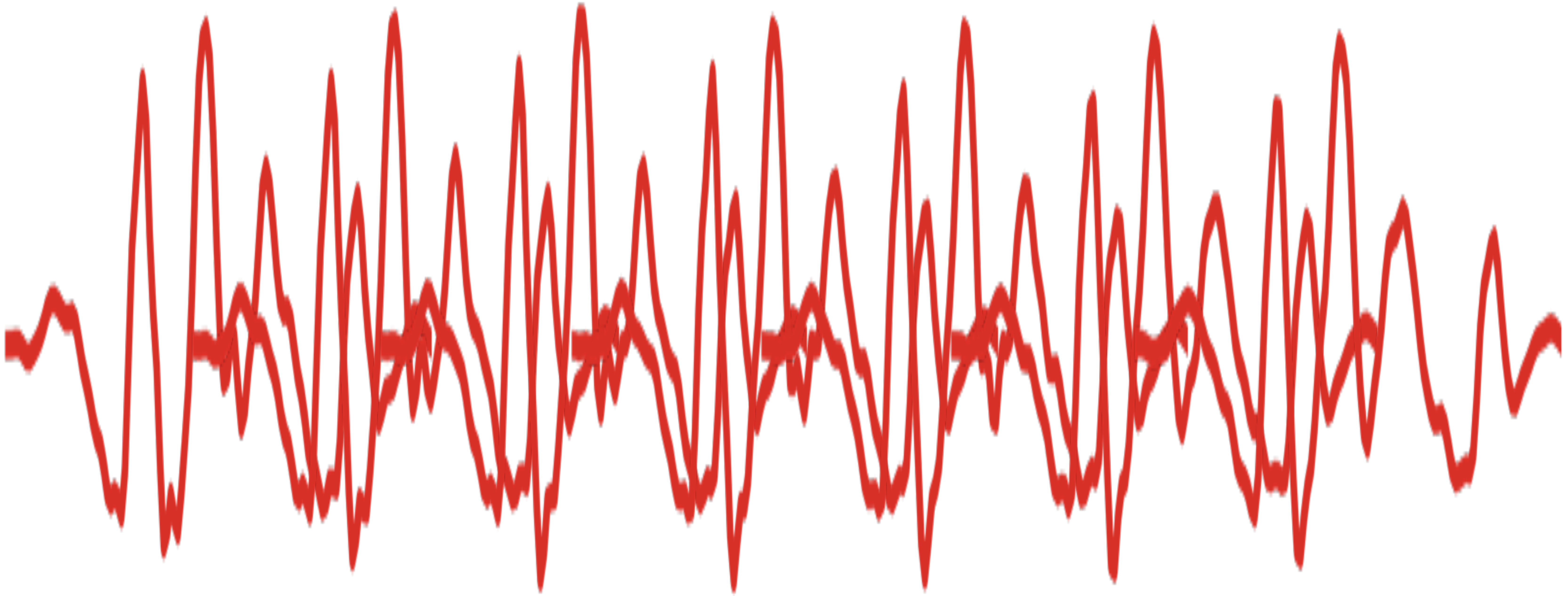


TD-PSOLA

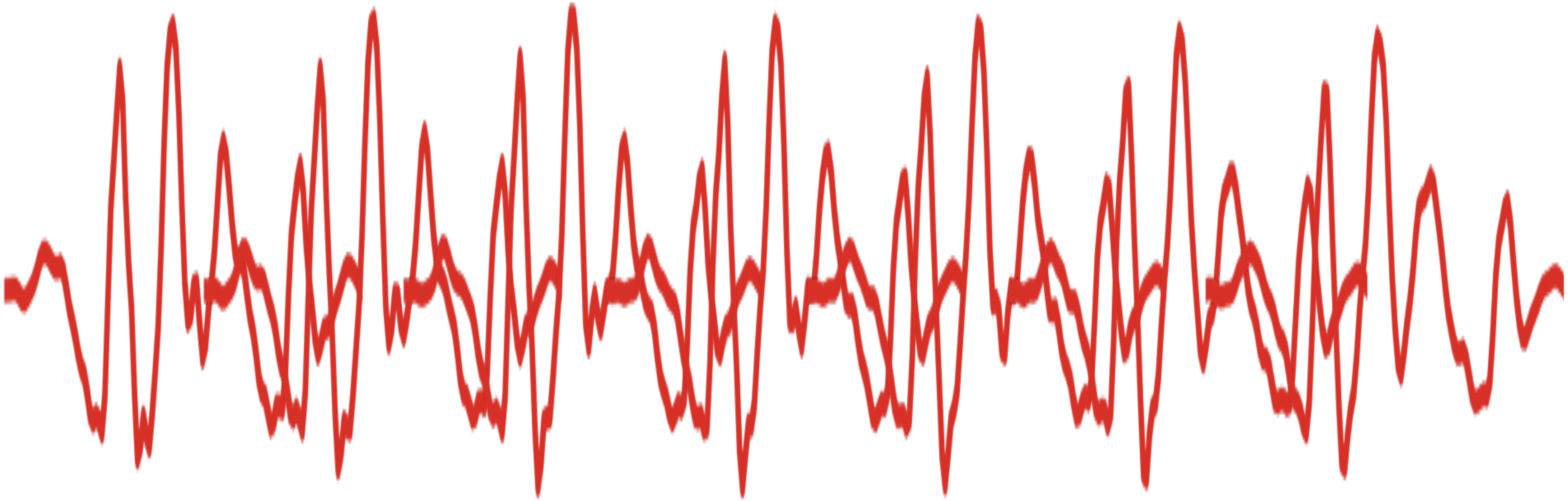
Time-domain pitch-synchronous overlap-and-add


Amplitude
0
255
260
265
270
275
280
Time (ms)

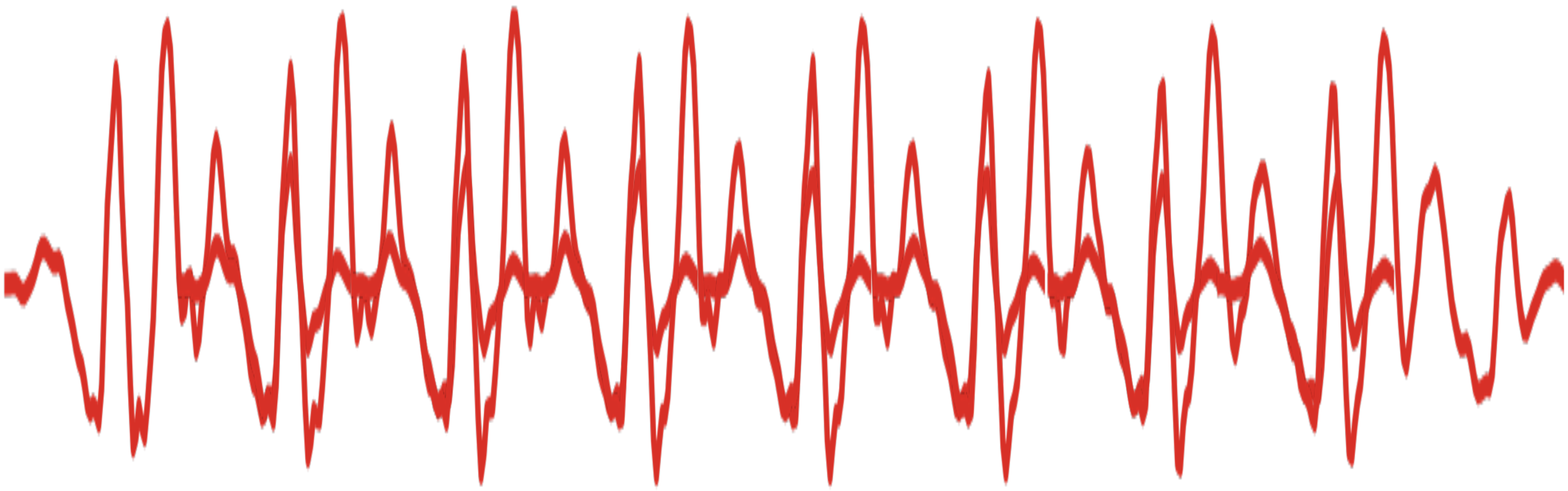
Increase F0



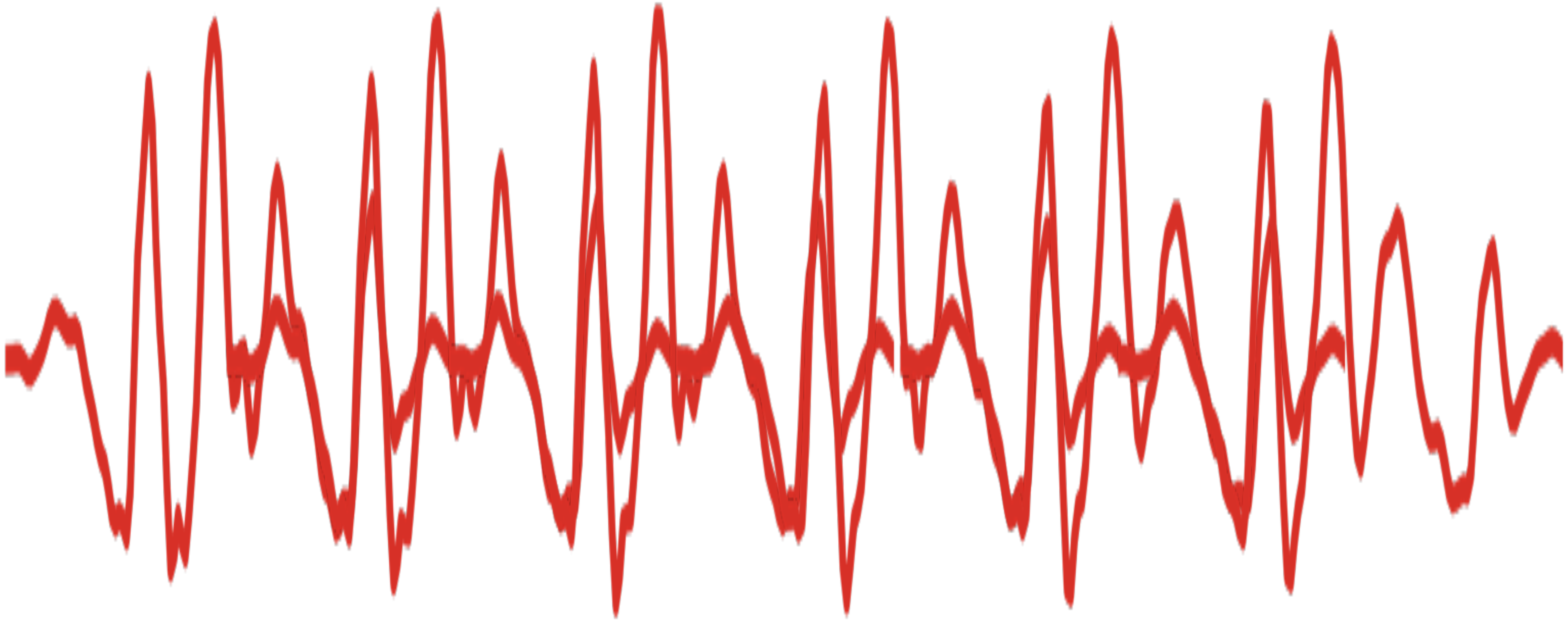
Decrease F0



Increase duration



Decrease duration



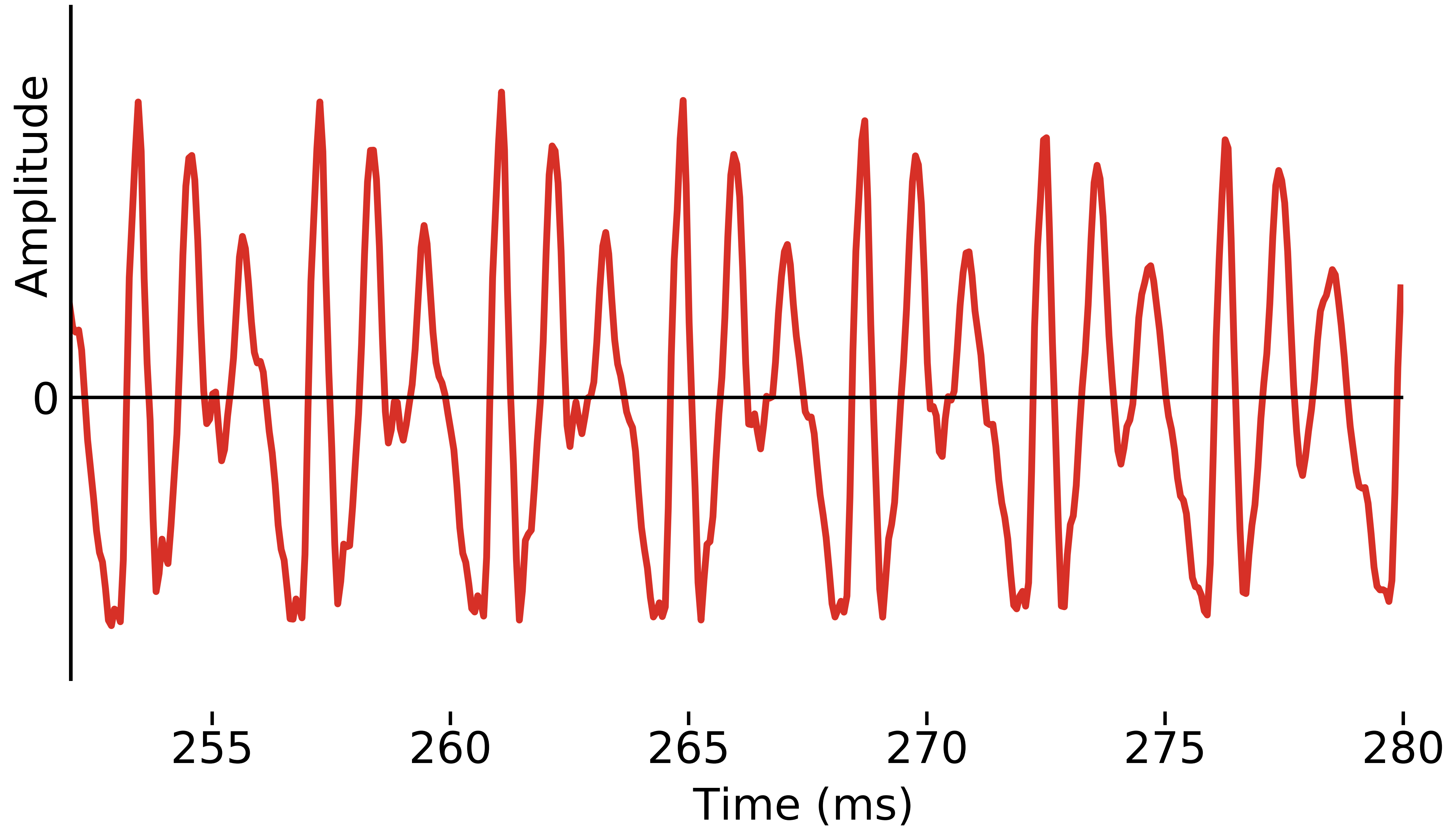
Putting it all together

sil-f	f-æ	æ-t	t-k	k-æ	æ-t	t-sil
sil-æ	æ-n	n-ɪ	ɪ-m	m-ə	ə-l	l-sil
sil-h	h-ɒ	ɒ-t	t-p	p-æ	æ-n	n-sil

sil-f	f-æ	æ-n	n-sil

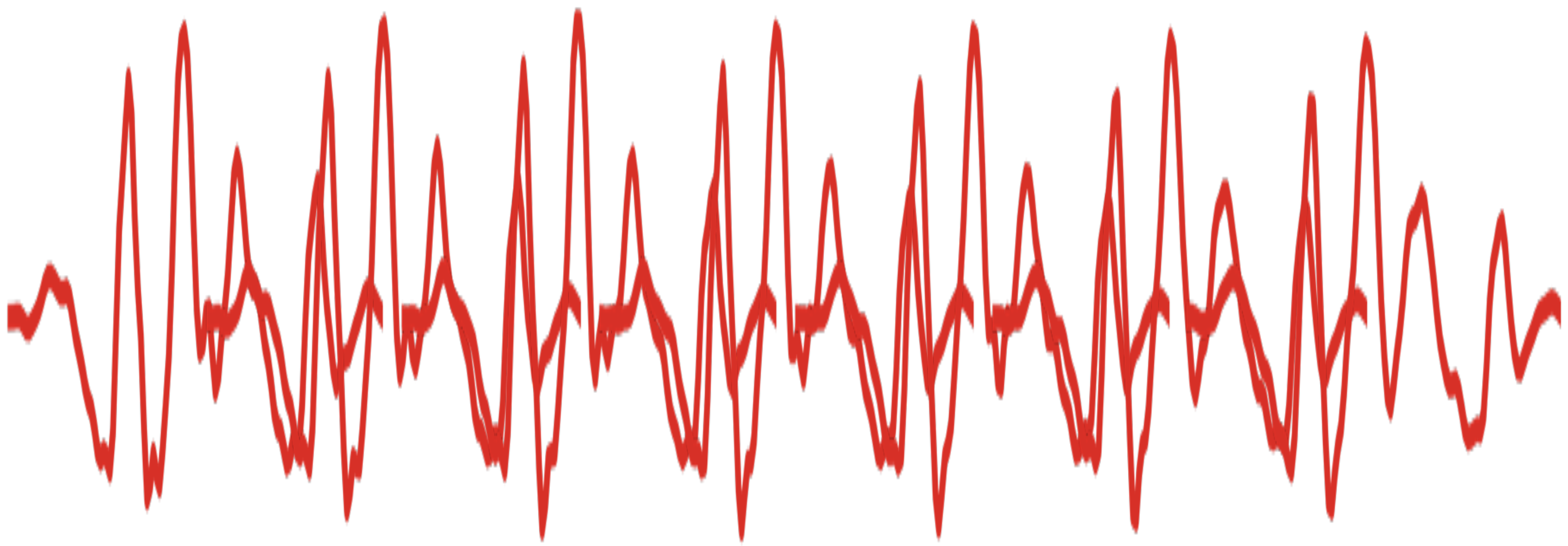
Putting it all together

sil-f | f-æ | æ-n | n-sil



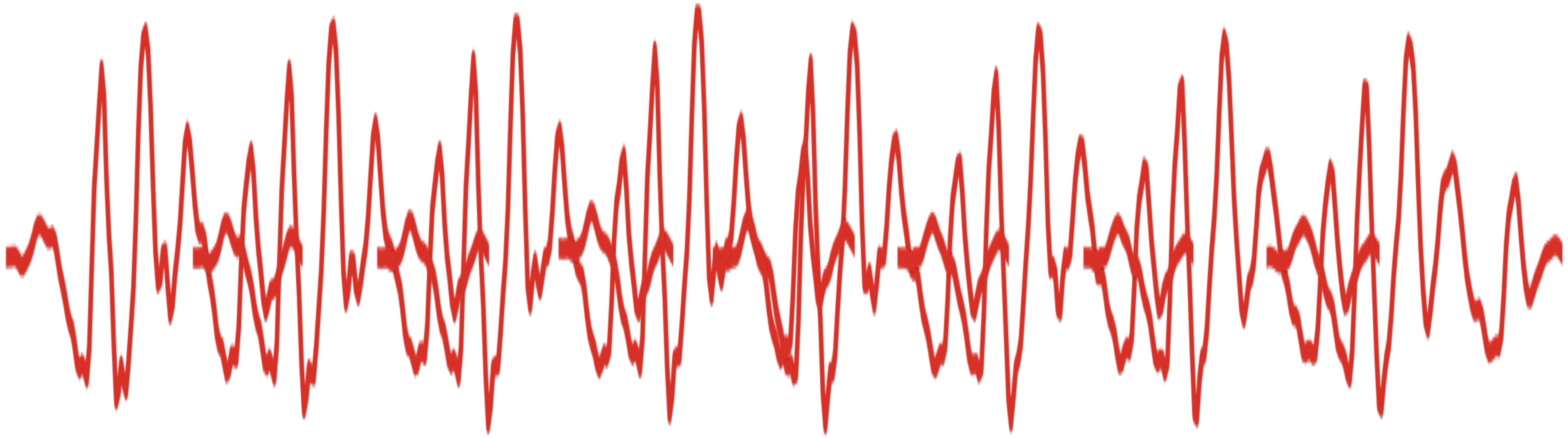
Putting it all together

sil-f | f-æ | æ-n | n-sil

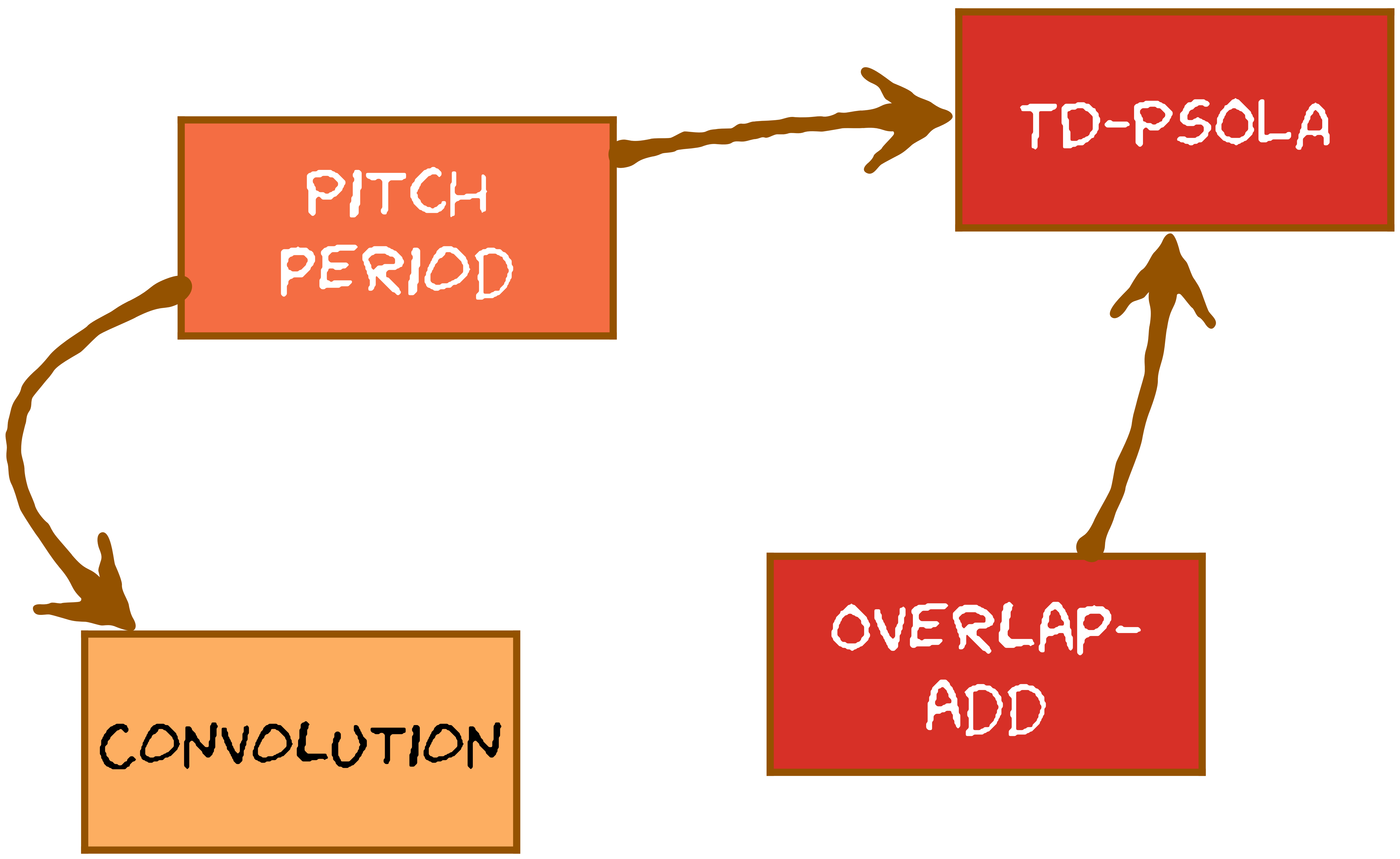


Putting it all together

sil-f f-æ æ-n n-sil



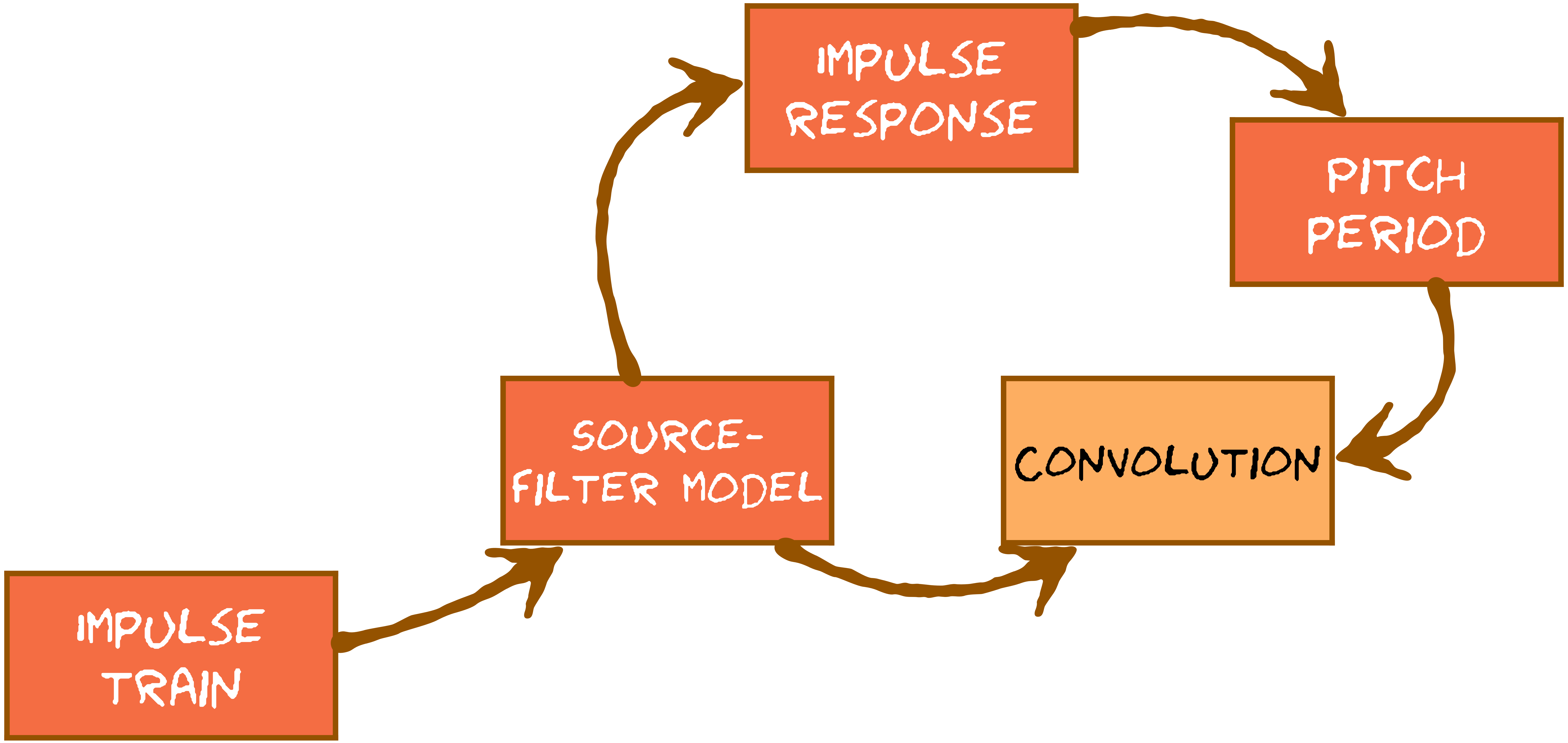
What you can learn next



CONVOLUTION

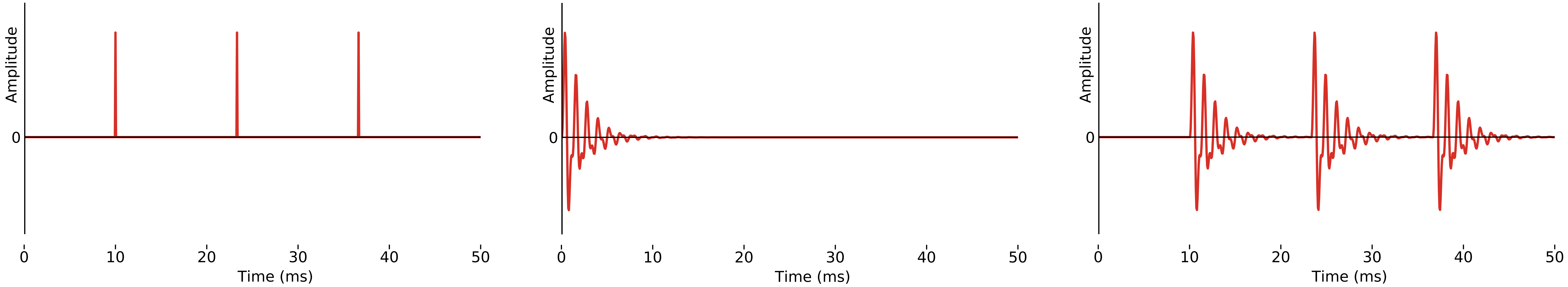
FREQUENCY DOMAIN AND BEYOND

What you need to know already



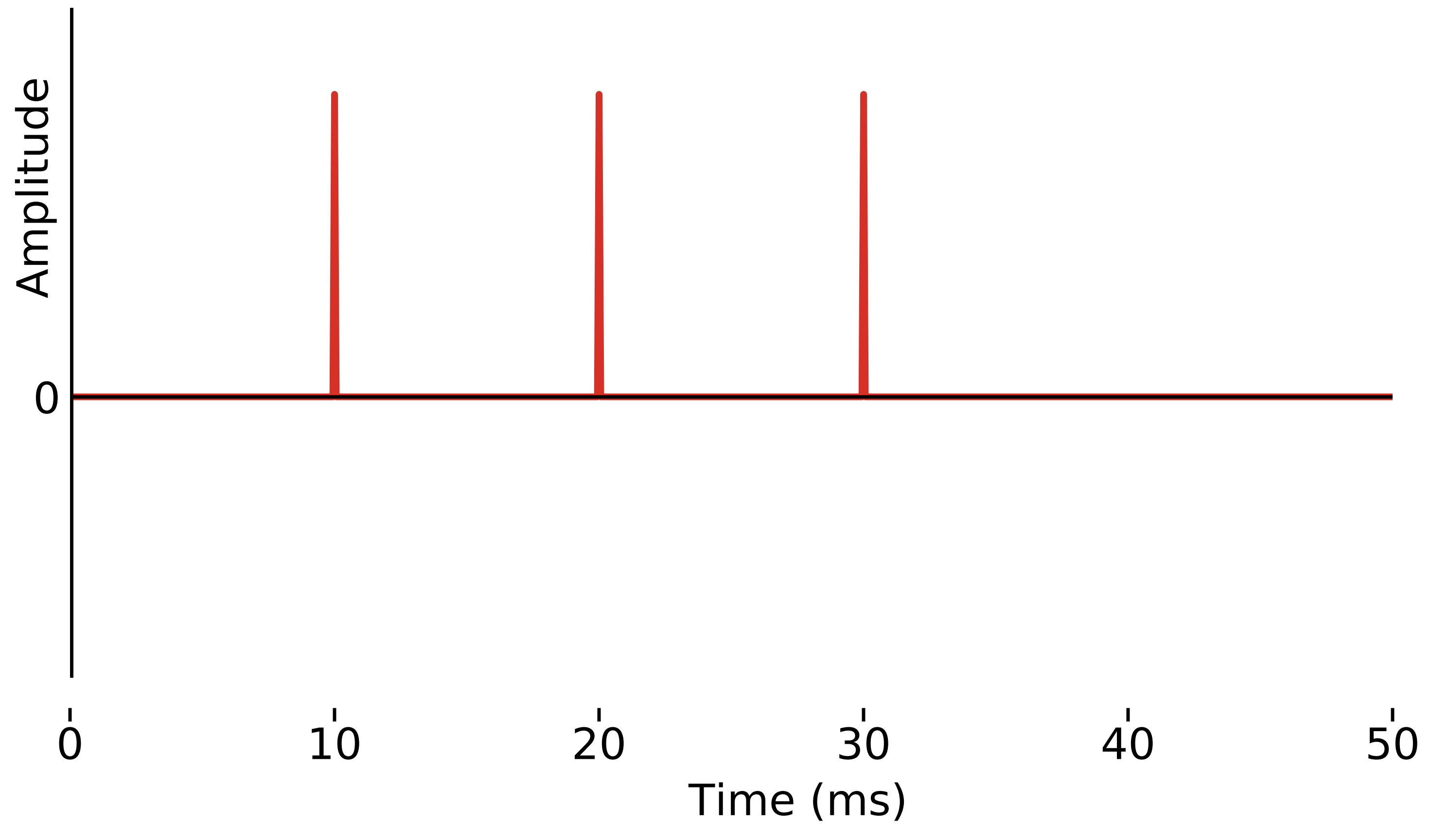
Source-filter model

excitation filter speech

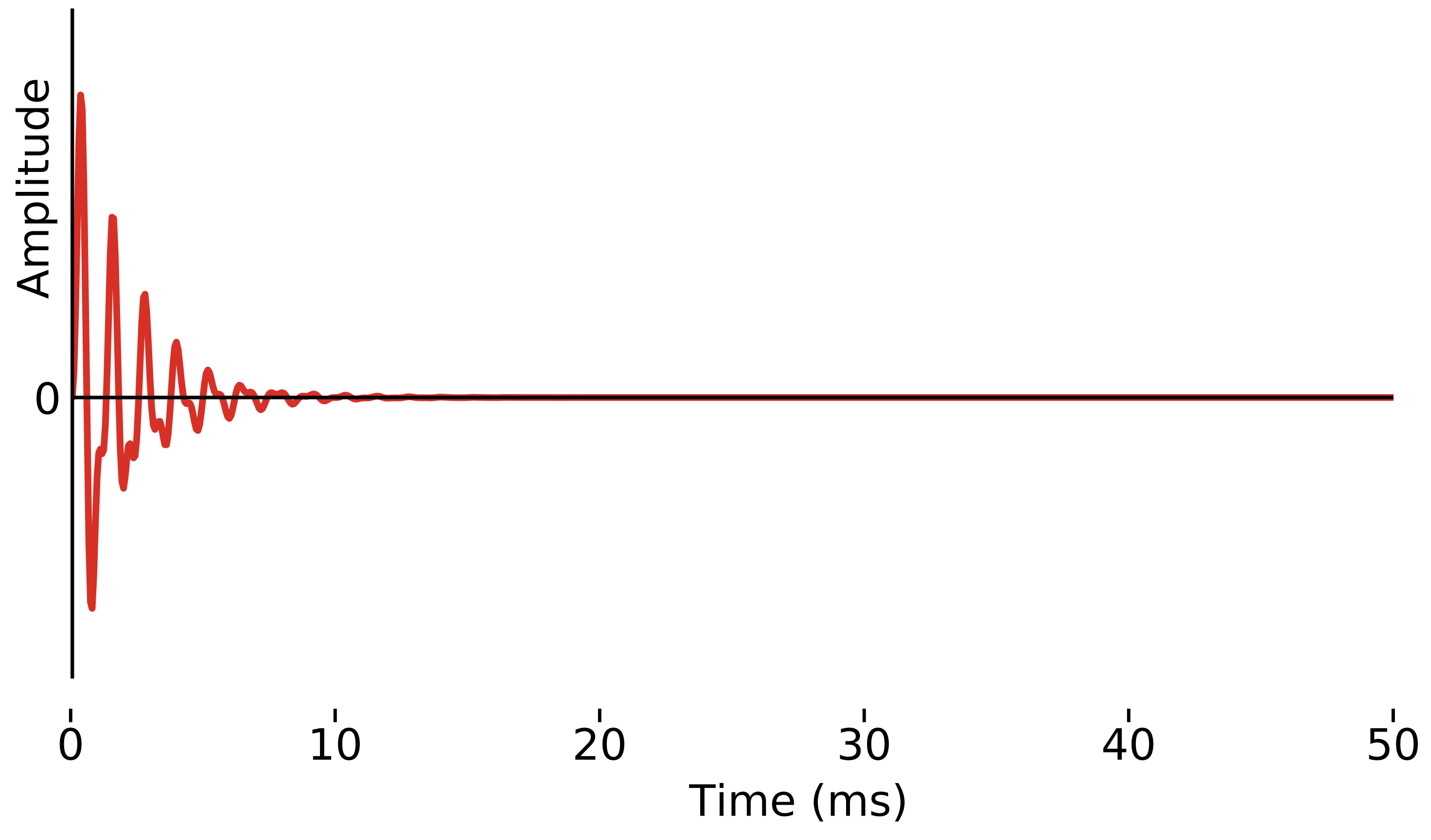


Source-filter model

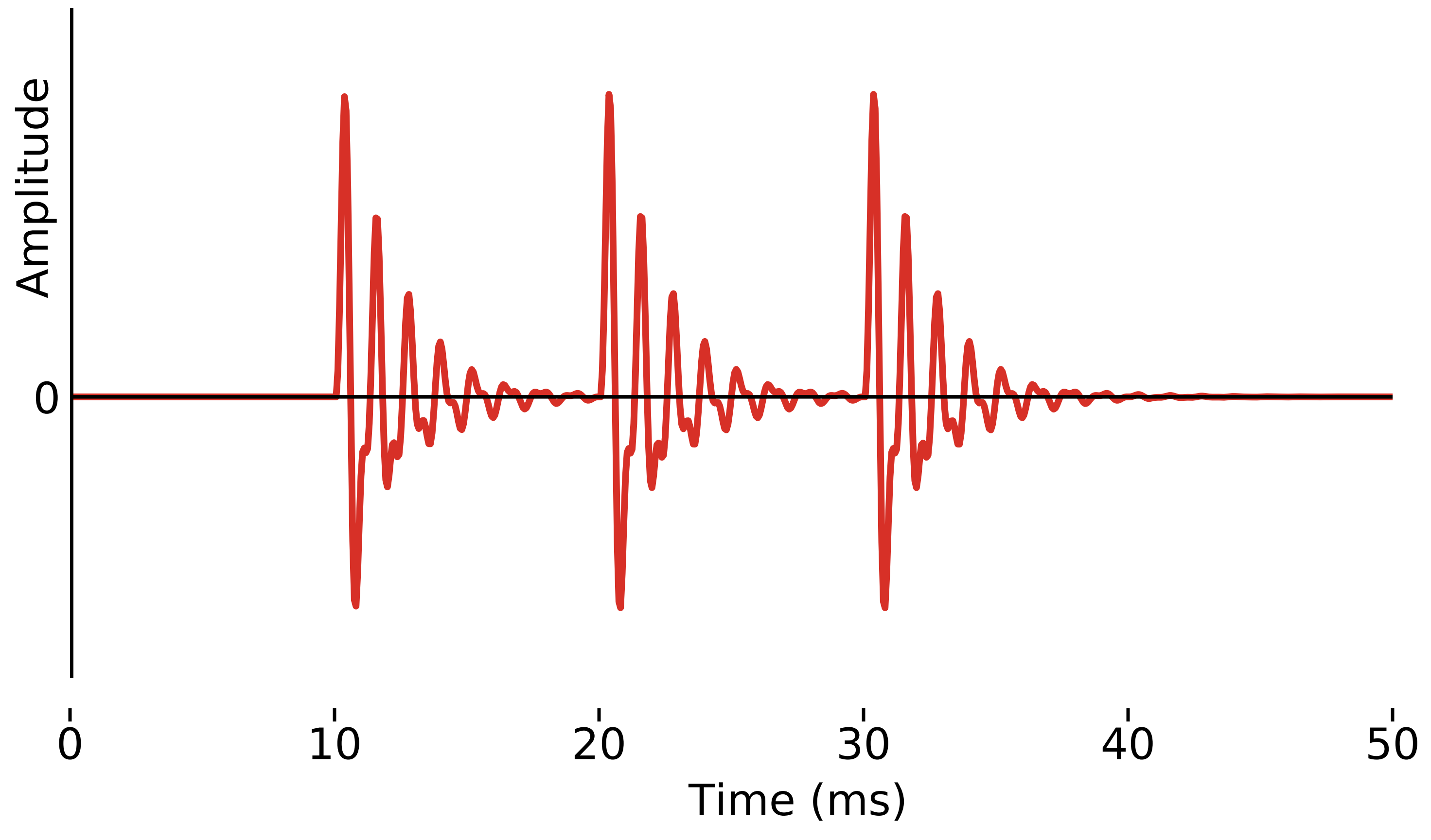
excitation



filter

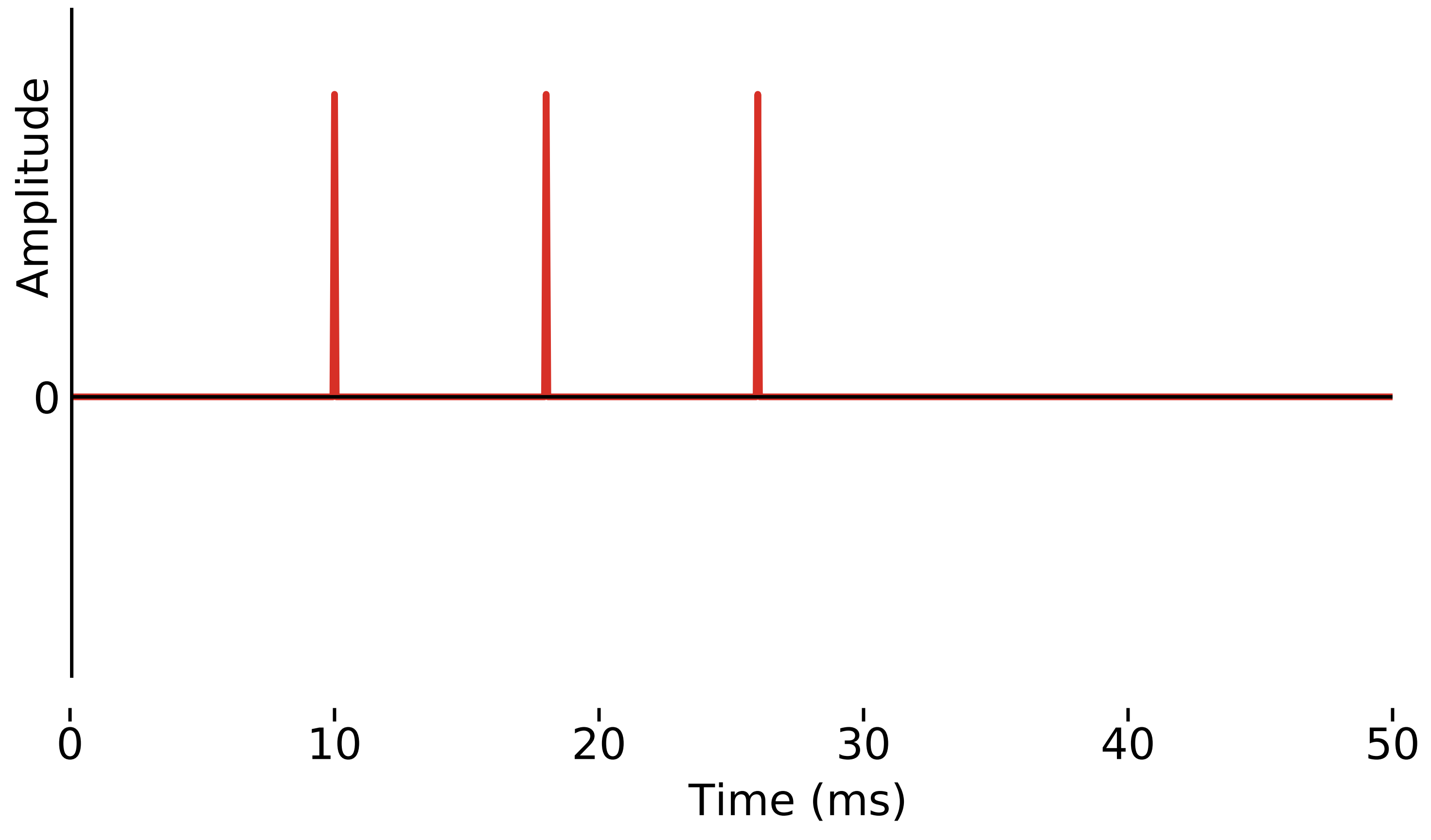


speech

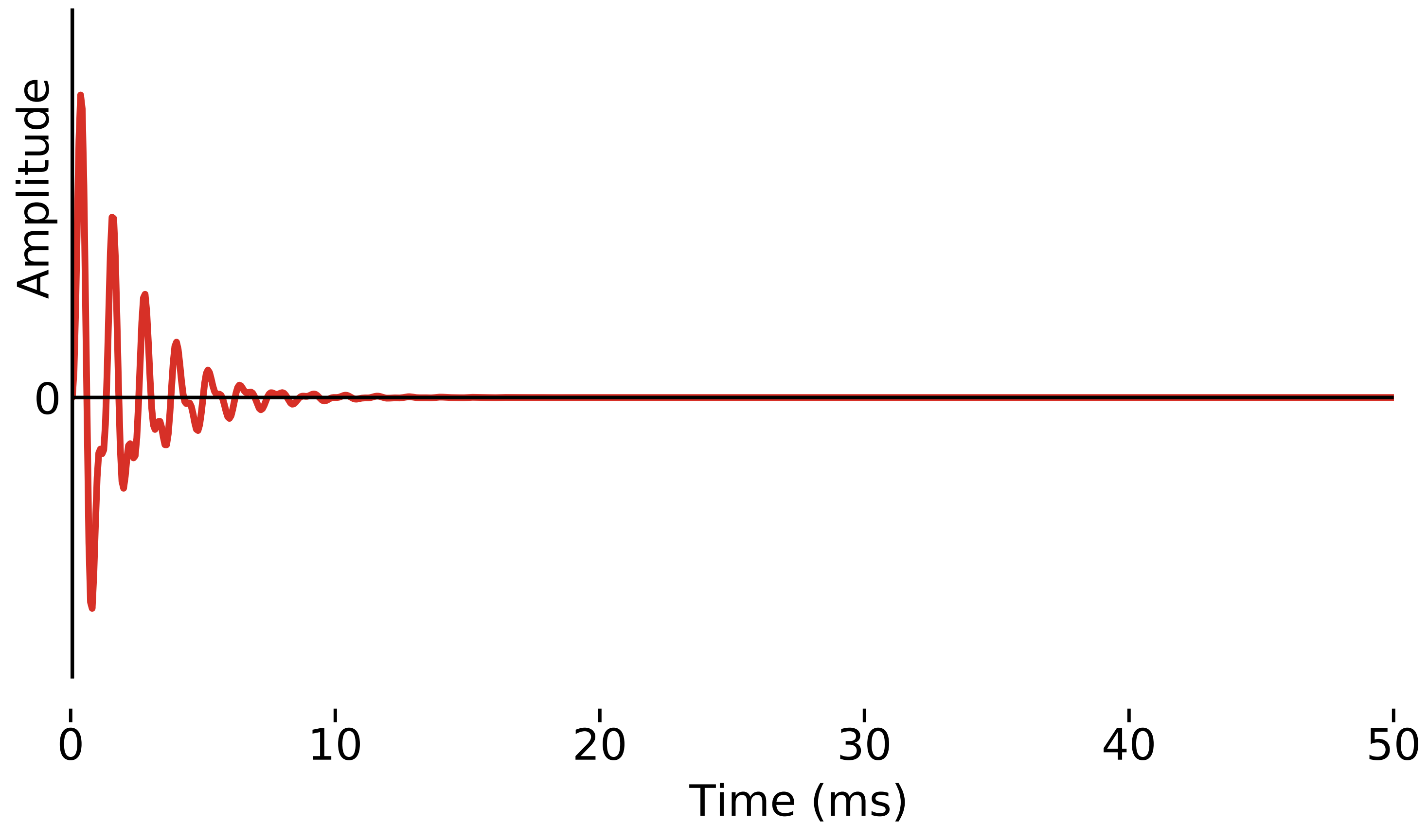


Source-filter model

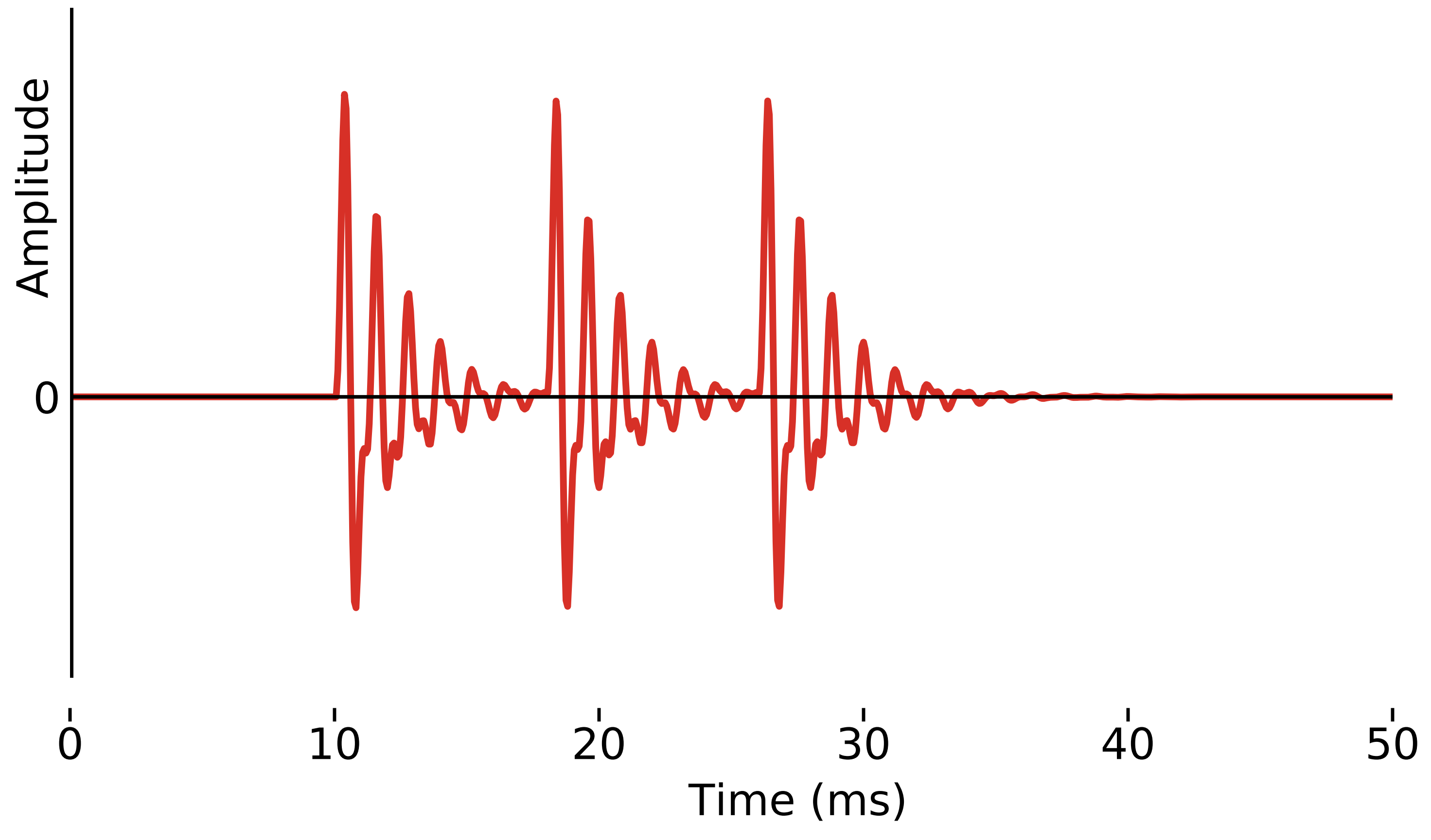
excitation



filter

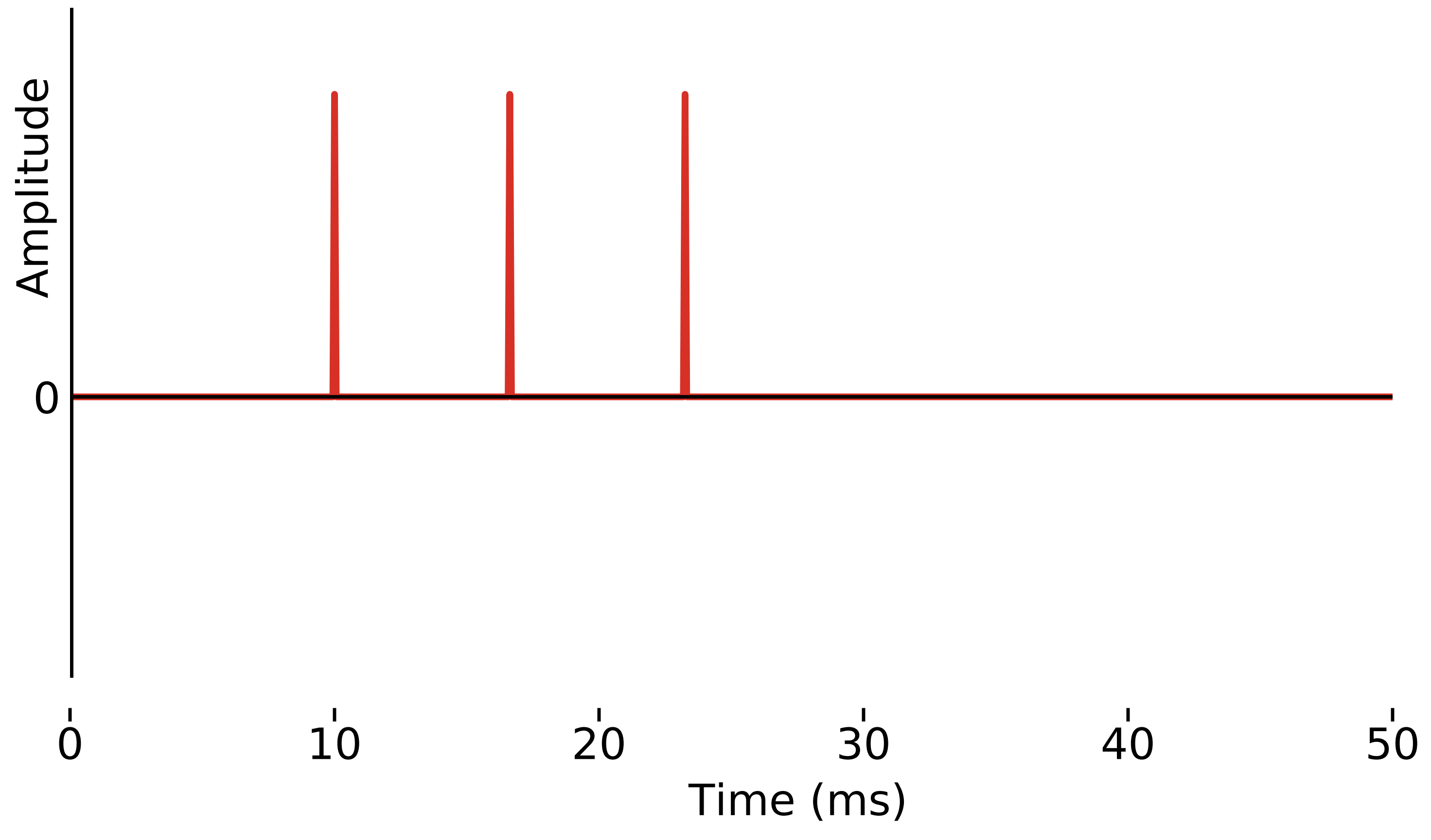


speech

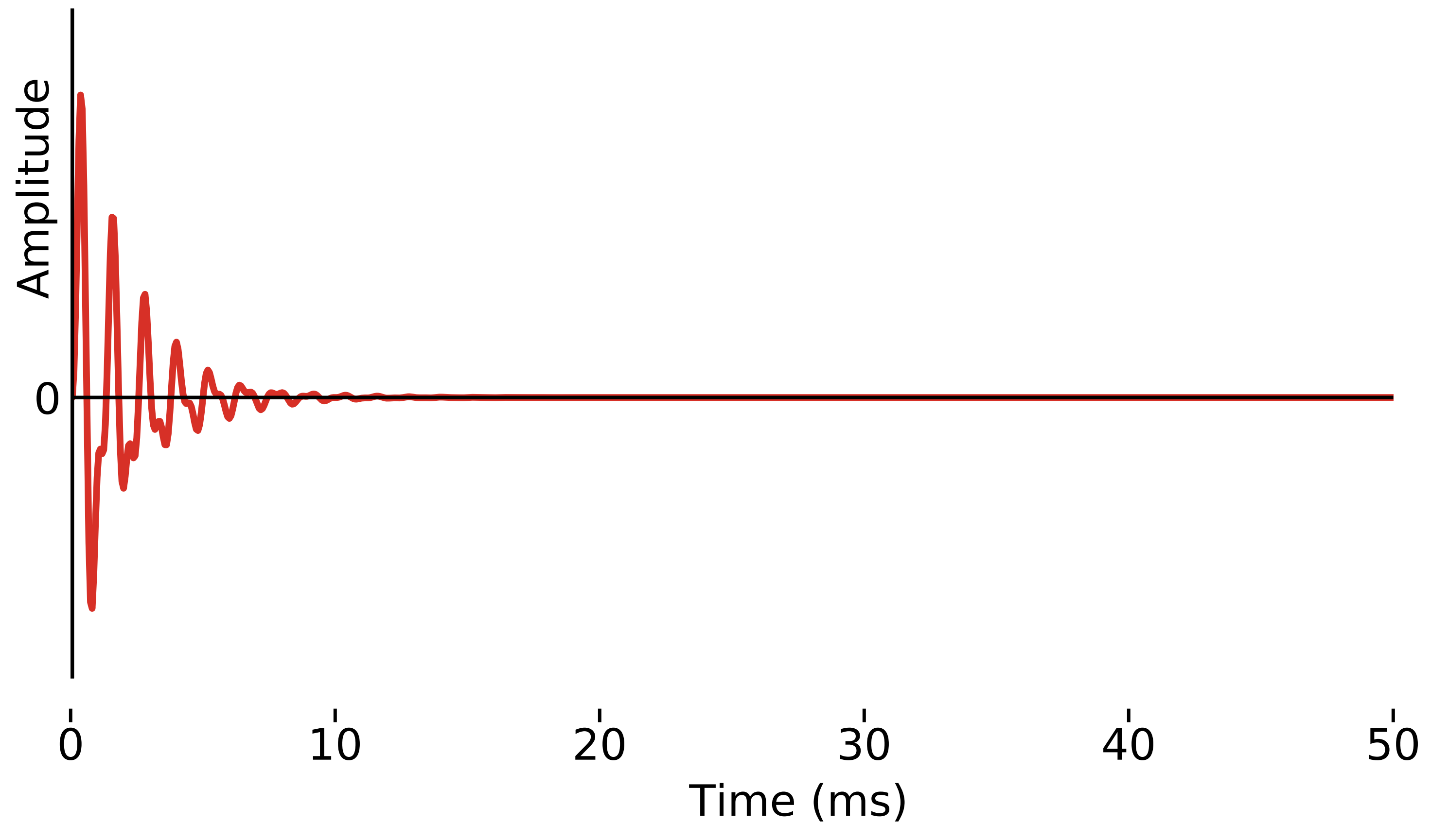


Source-filter model

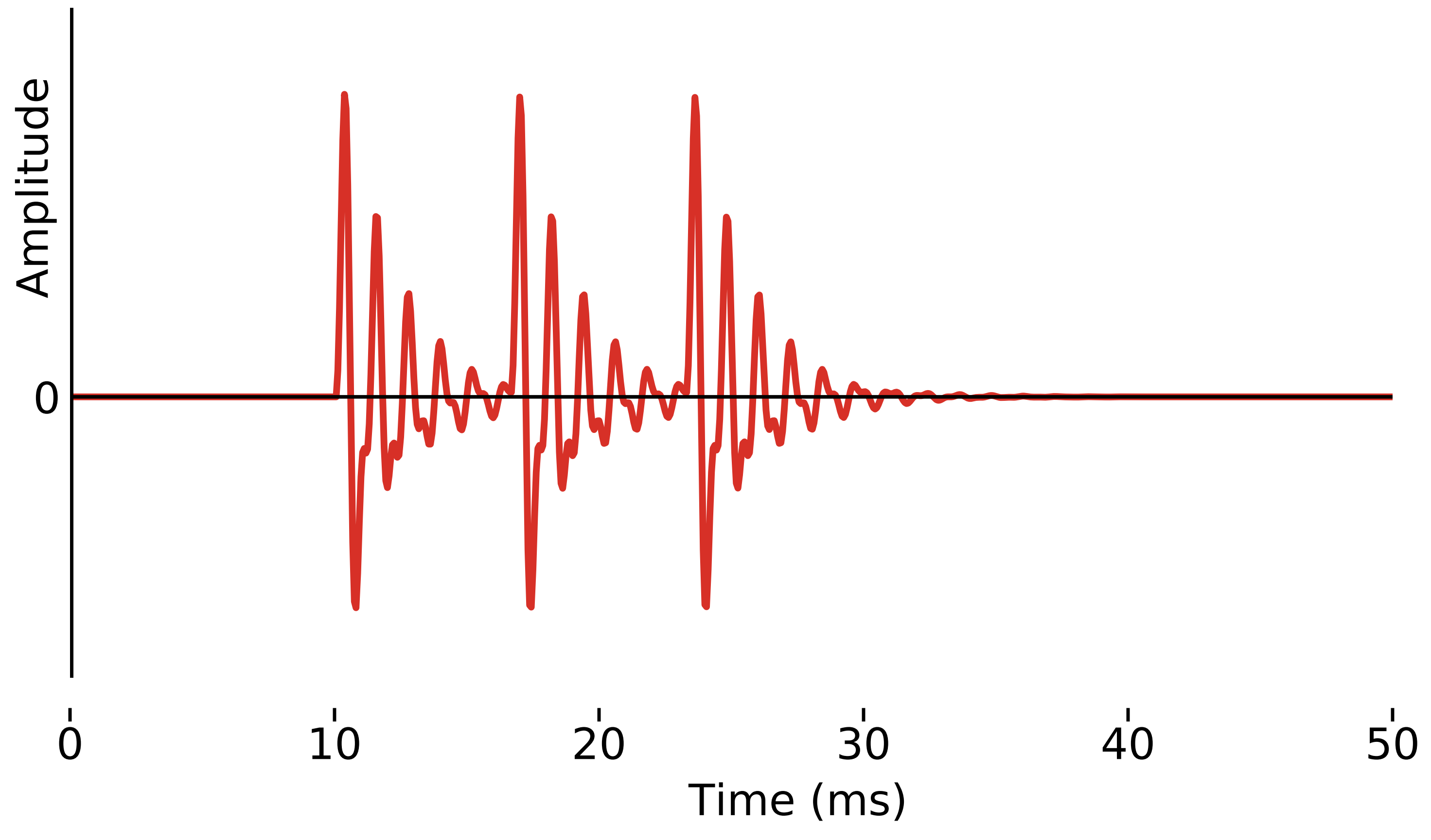
excitation

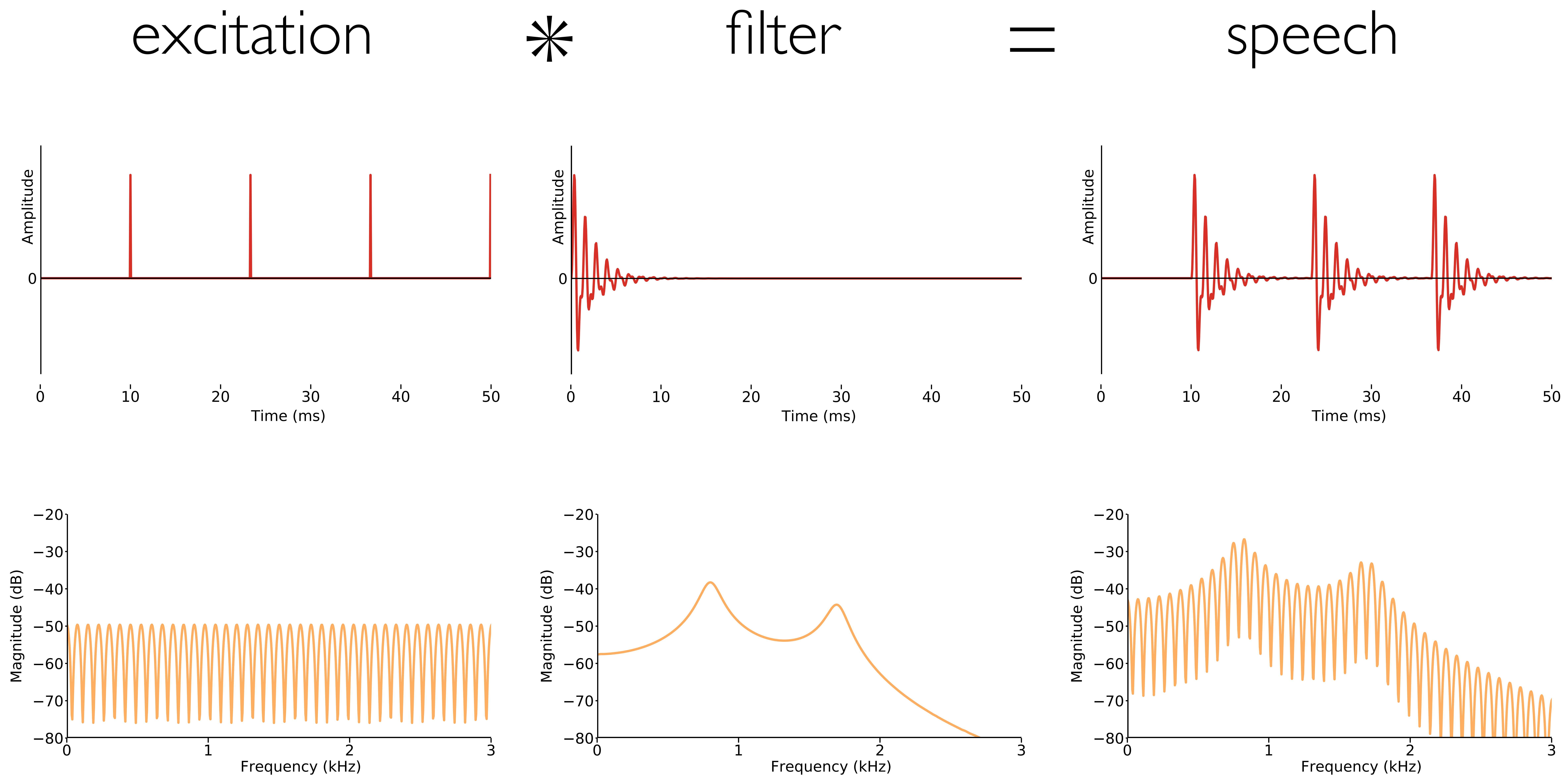


filter

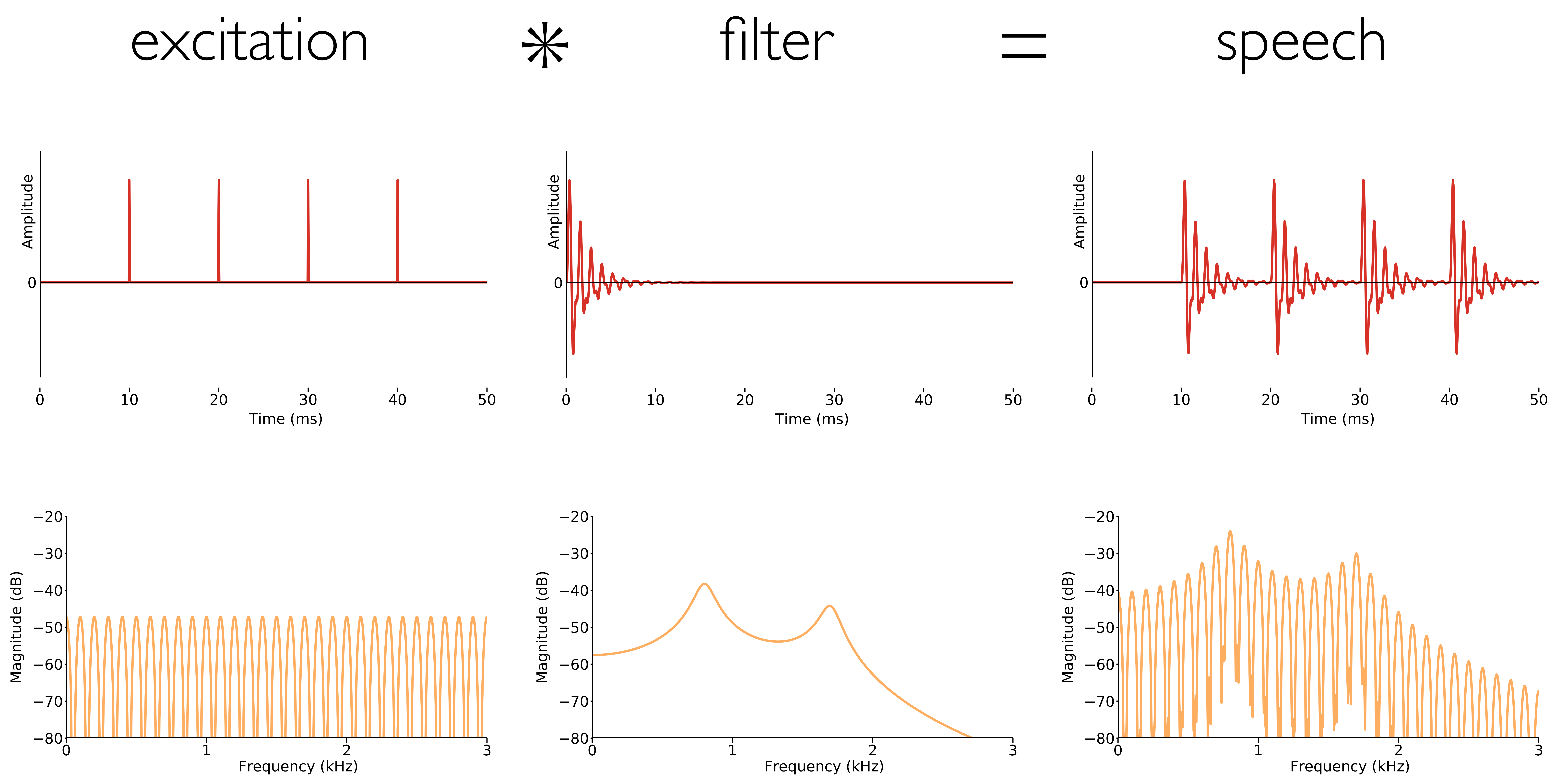


speech



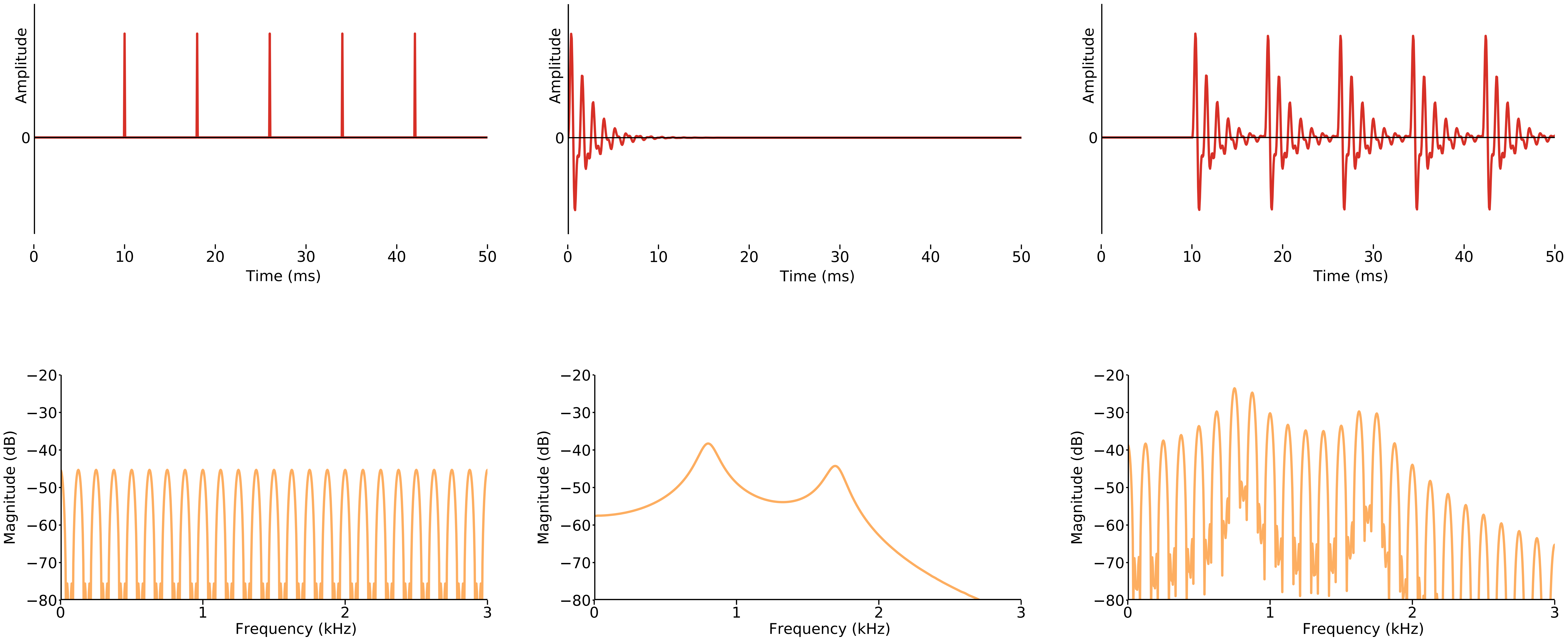


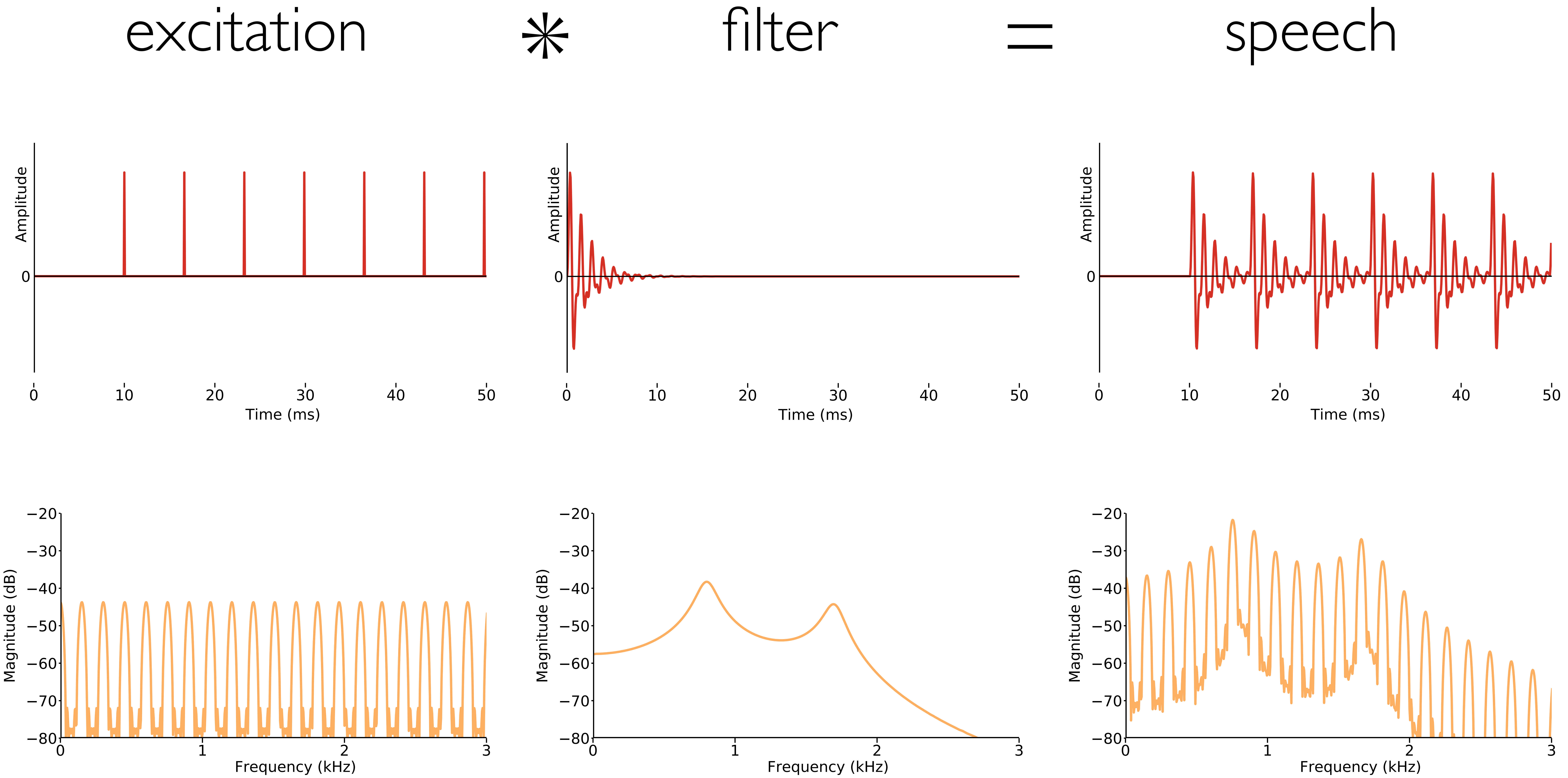
excitation
∗
filter
=
speech
Amplitude
0
0
10
20
30
40
50
Time (ms)
Magnitude (dB)
−20
−30
−40
−50
−60
−70
−80
0
1
2
3
Frequency (kHz)



excitation
*
filter
=
speech
Amplitude
0
0
10
20
30
40
50
Time (ms)
Amplitude
0
0
10
20
30
40
50
Time (ms)
Amplitude
0
0
10
20
30
40
50
Time (ms)
Magnitude (dB)
−20
−30
−40
−50
−60
−70
−80
0
1
2
3
Frequency (kHz)
Magnitude (dB)
−20
−30
−40
−50
−60
−70
−80
0
1
2
3
Frequency (kHz)
Magnitude (dB)
−20
−30
−40
−50
−60
−70
−80
0
1
2
3
Frequency (kHz)

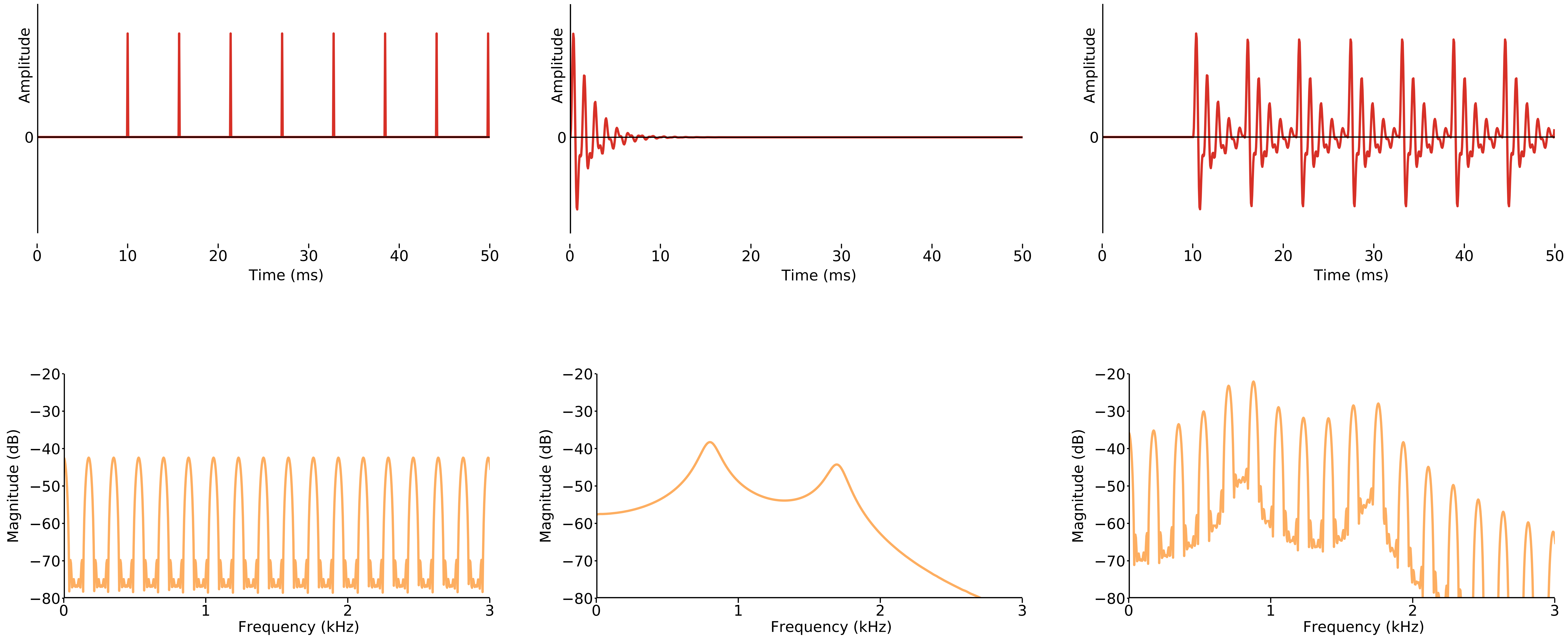
excitation $*$ filter $=$ speech



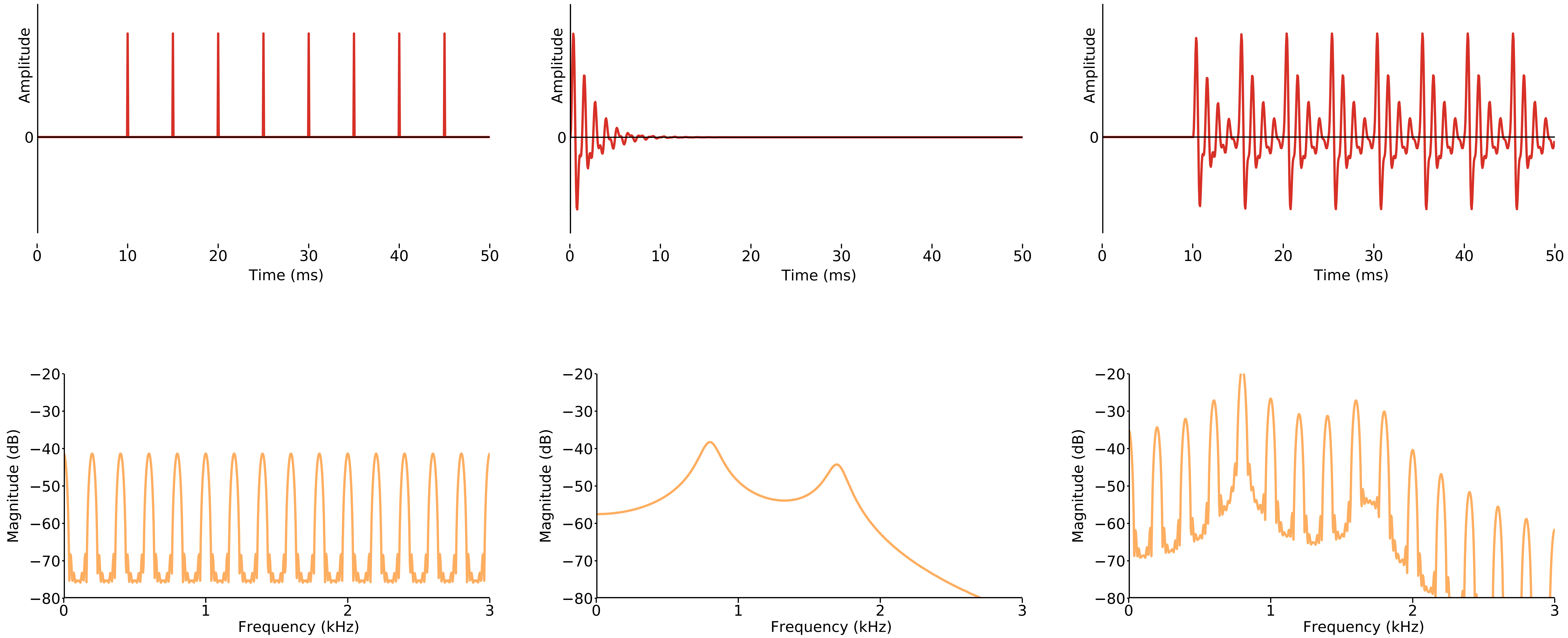


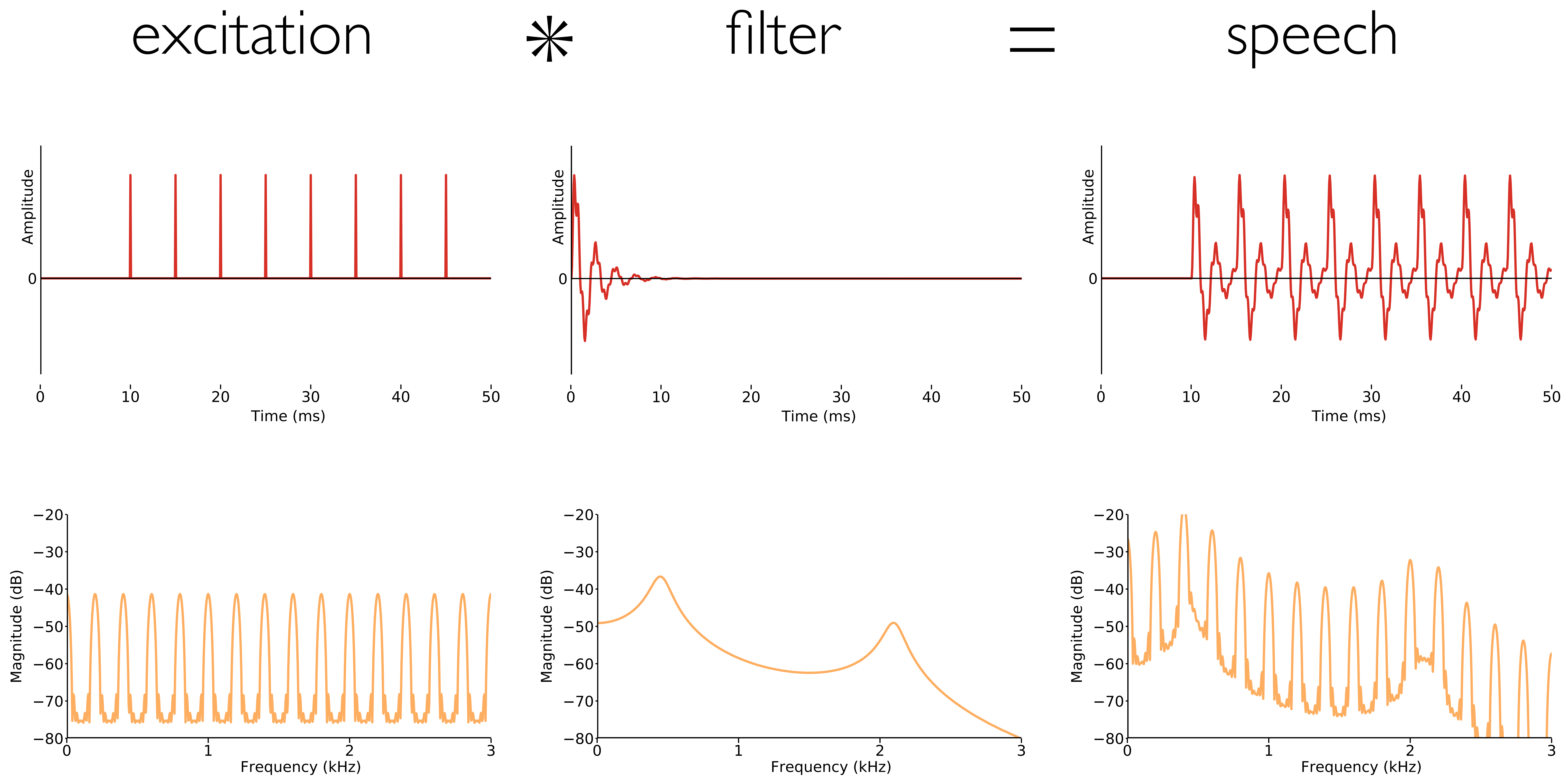
excitation * filter = speech
Amplitude
0
0 10 20 30 40 50
Time (ms)
Magnitude (dB)
−20 −30 −40 −50 −60 −70 −80
0 1 2 3
Frequency (kHz)

excitation ∗ filter = speech

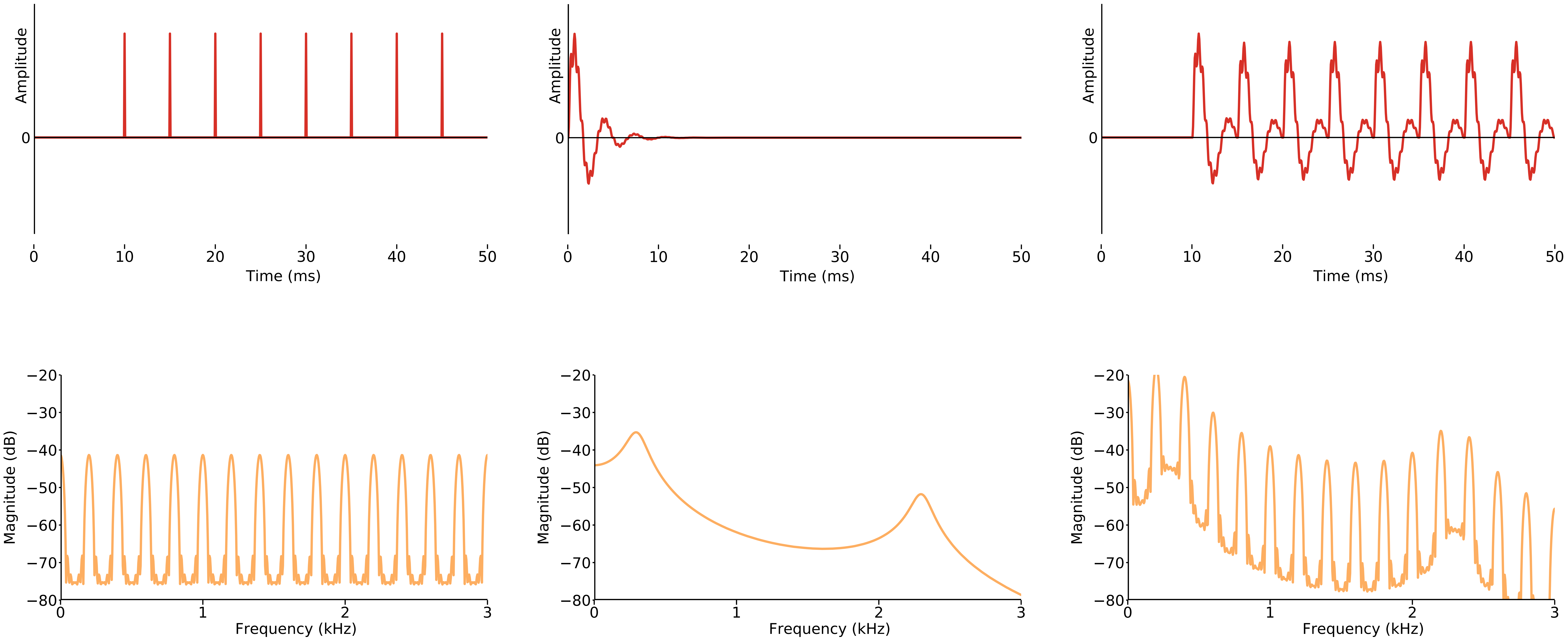


excitation $*$ filter $=$ speech

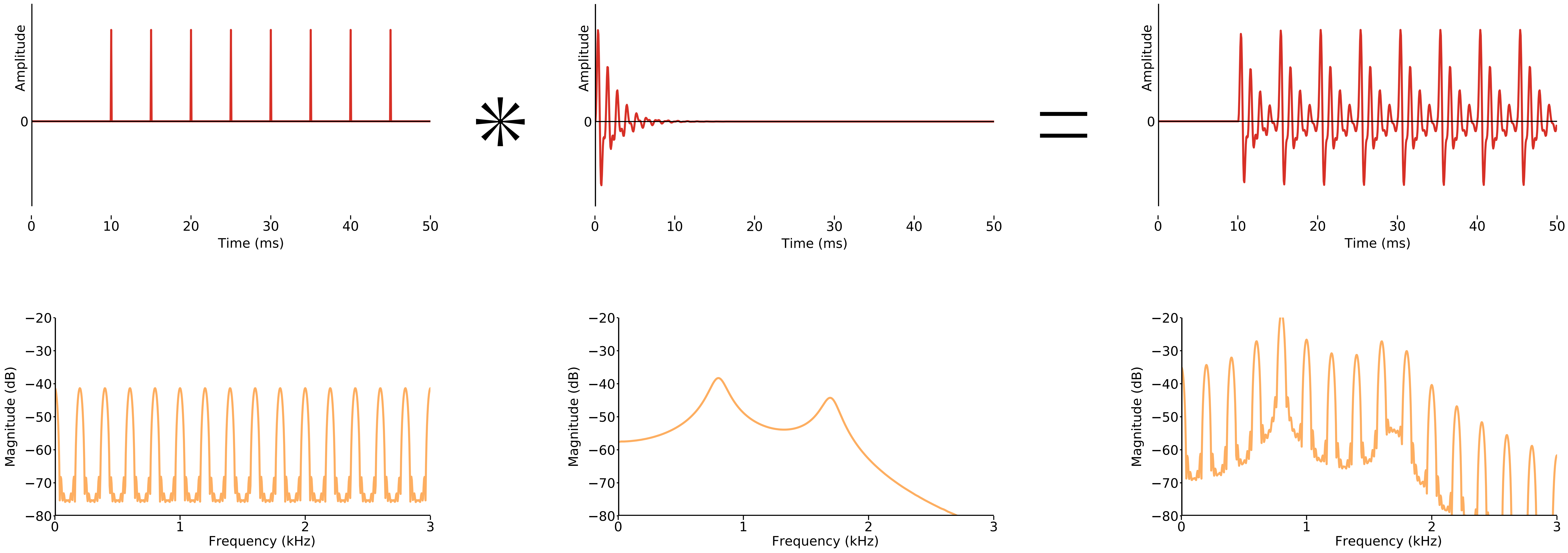



excitation * filter = speech
Amplitude
0
Time (ms)
0 10 20 30 40 50
Magnitude (dB)
−20 −30 −40 −50 −60 −70 −80
Frequency (kHz)
0 1 2 3

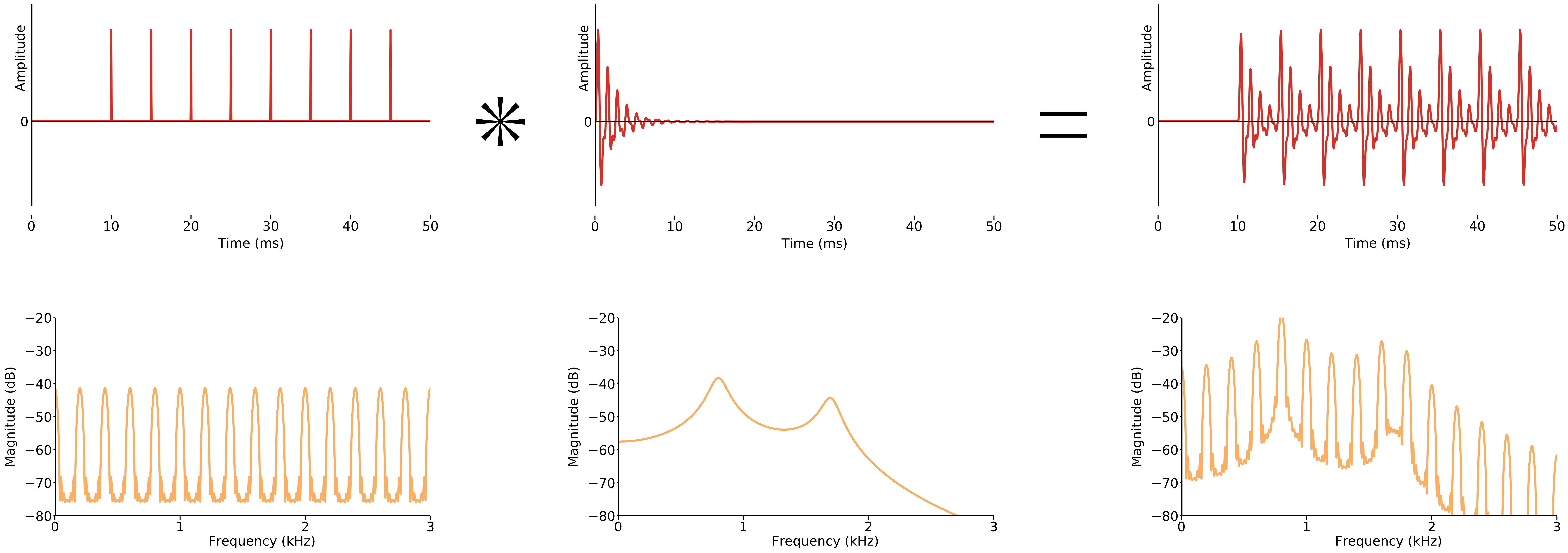
excitation ∗ filter = speech



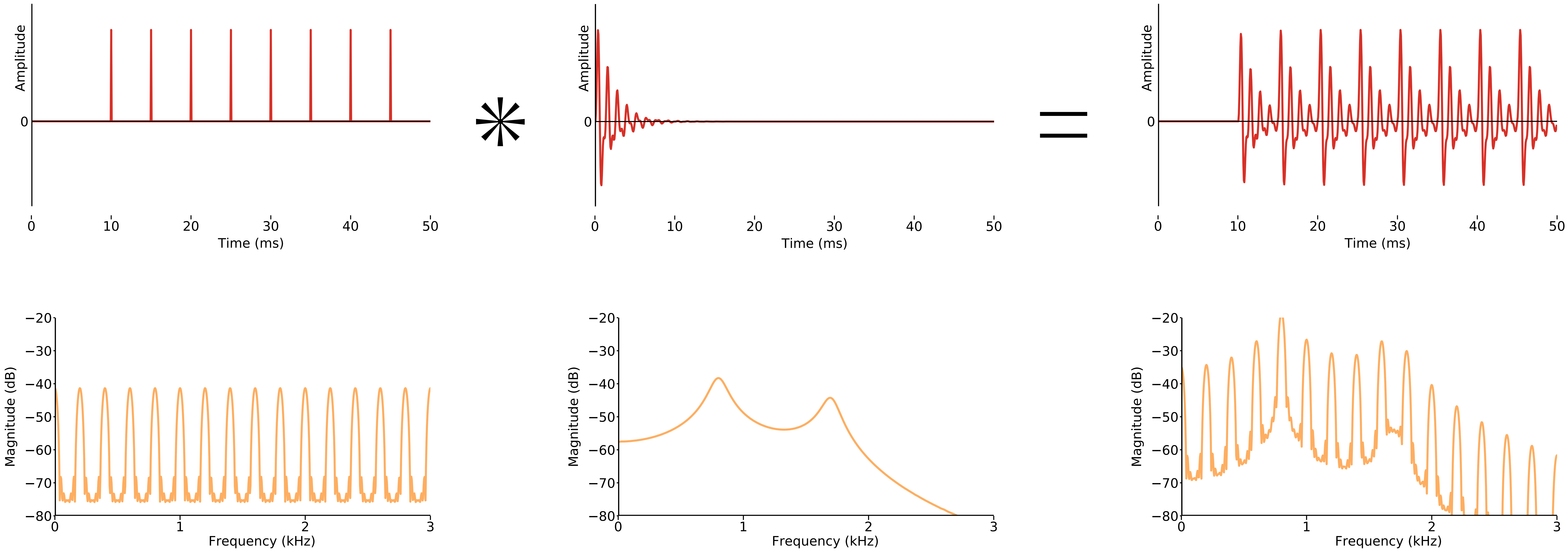
convolution in the time domain = multiplication in the frequency domain



convolution of waveforms = multiplication of magnitude spectra



convolution of waveforms = addition of log magnitude spectra



What you can learn next