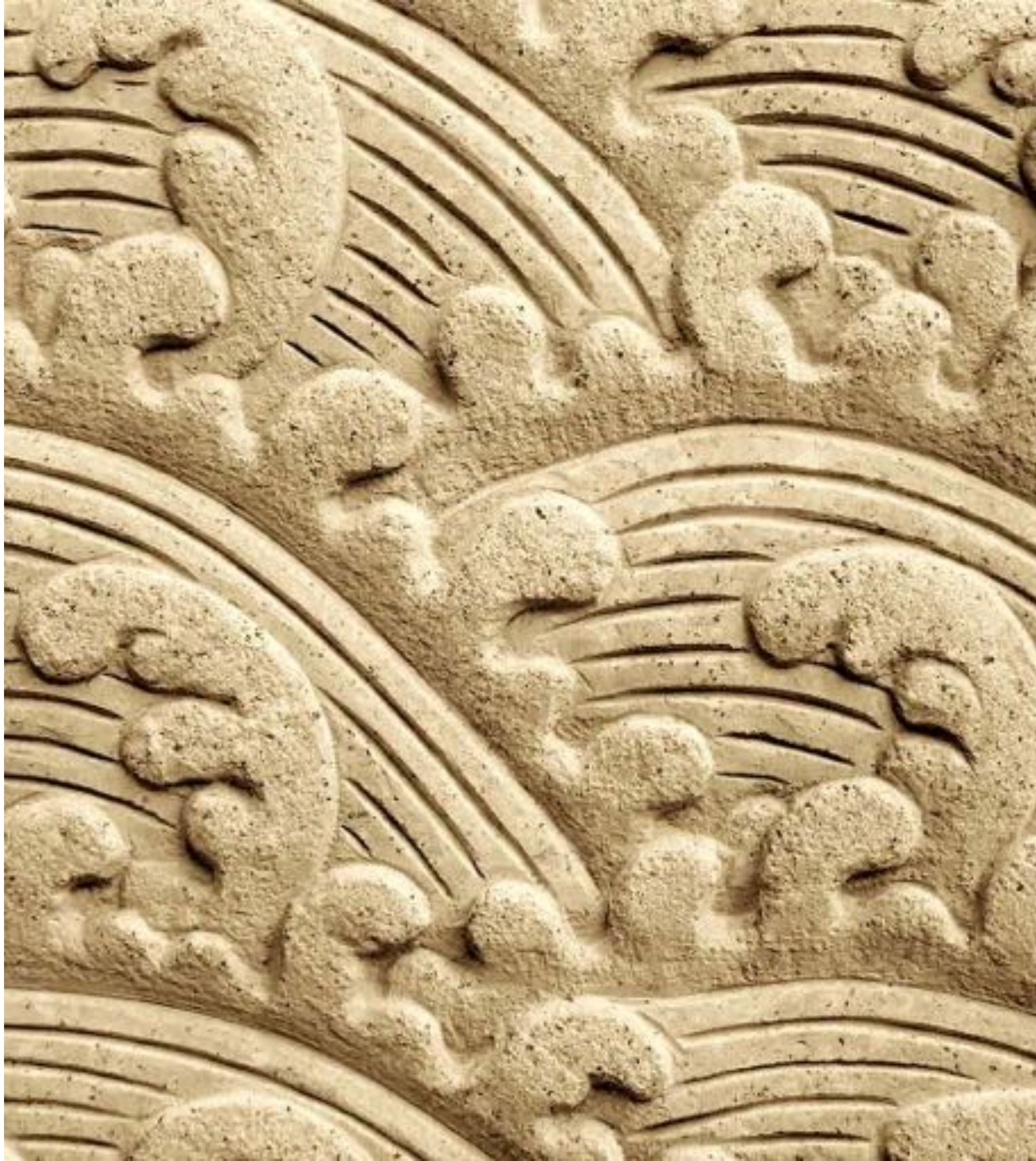
# Speech Processing

Simon King
University of Edinburgh

2022-23



# Module 7

## Pattern matching

# Orientation

- We're on a journey towards HMMs
  - Pattern **matching**
  - Extracting **features** from speech
  - Probabilistic **generative** modelling

*What we are learning along the way*

Dynamic programming

(in the form of Dynamic Time Warping)

The interaction between

- choice of model
- choice of features

Dynamic programming

(in the form of the Viterbi algorithm)

# What you should already know

- Why the **waveform** is not good for pattern recognition
- Concept of a **feature vector**
- Let's start as simple as possible: whole word templates
  - But we already have to deal with sequences of **different lengths**

Source and filter are combined

But we only want the **filter**

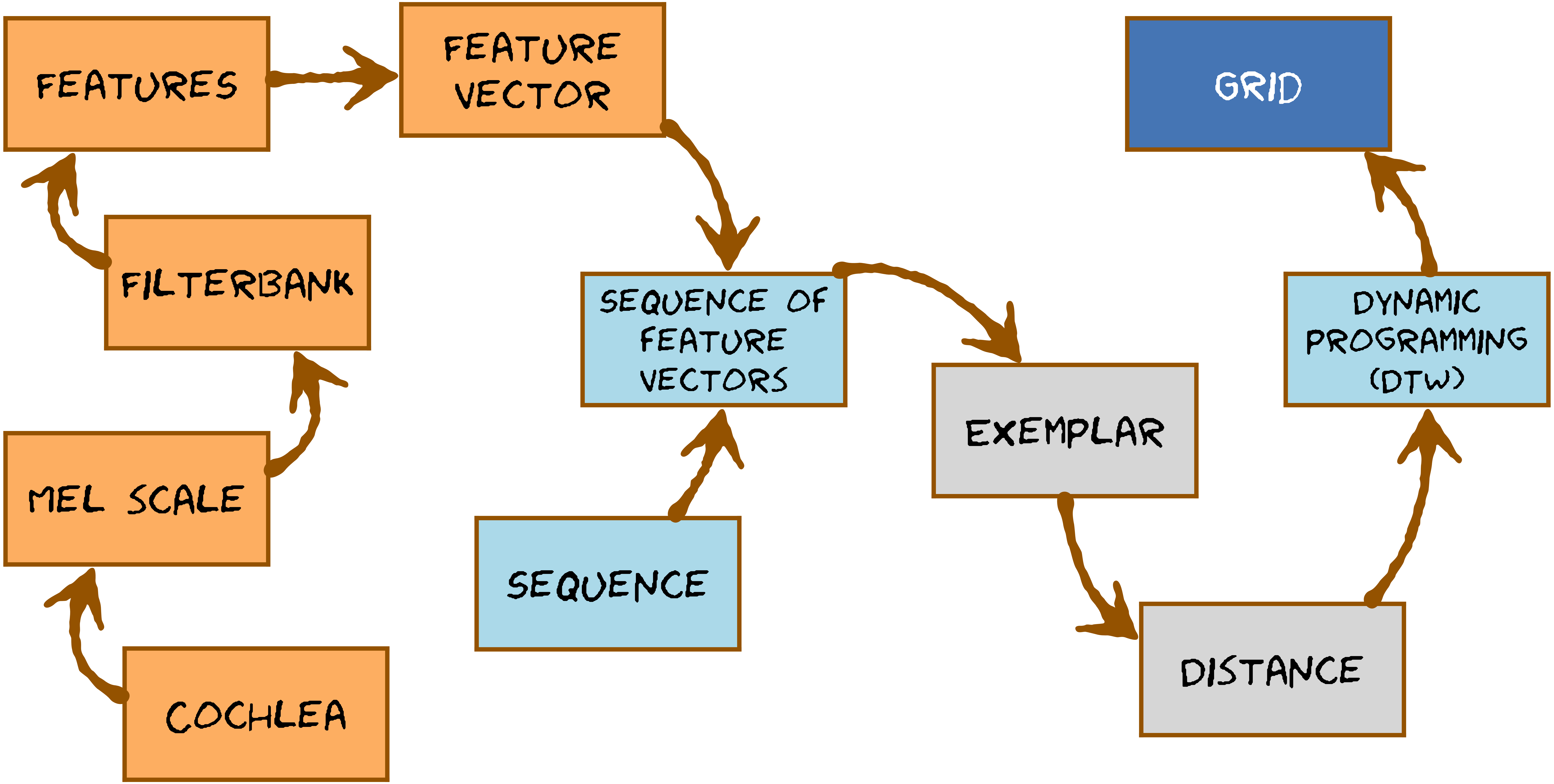
Speech waveforms change over time

Use short-term analysis

Extract features from frames of speech

Finding an alignment between two sequences

- linear time warping
- non-linear ('dynamic') time warping



FEATURES

FEATURE VECTOR

GRID

FILTERBANK

SEQUENCE OF FEATURE VECTORS

DYNAMIC PROGRAMMING (DTW)

EXEMPLAR

MEL SCALE

SEQUENCE

DISTANCE

COCHLEA

COCHLEA

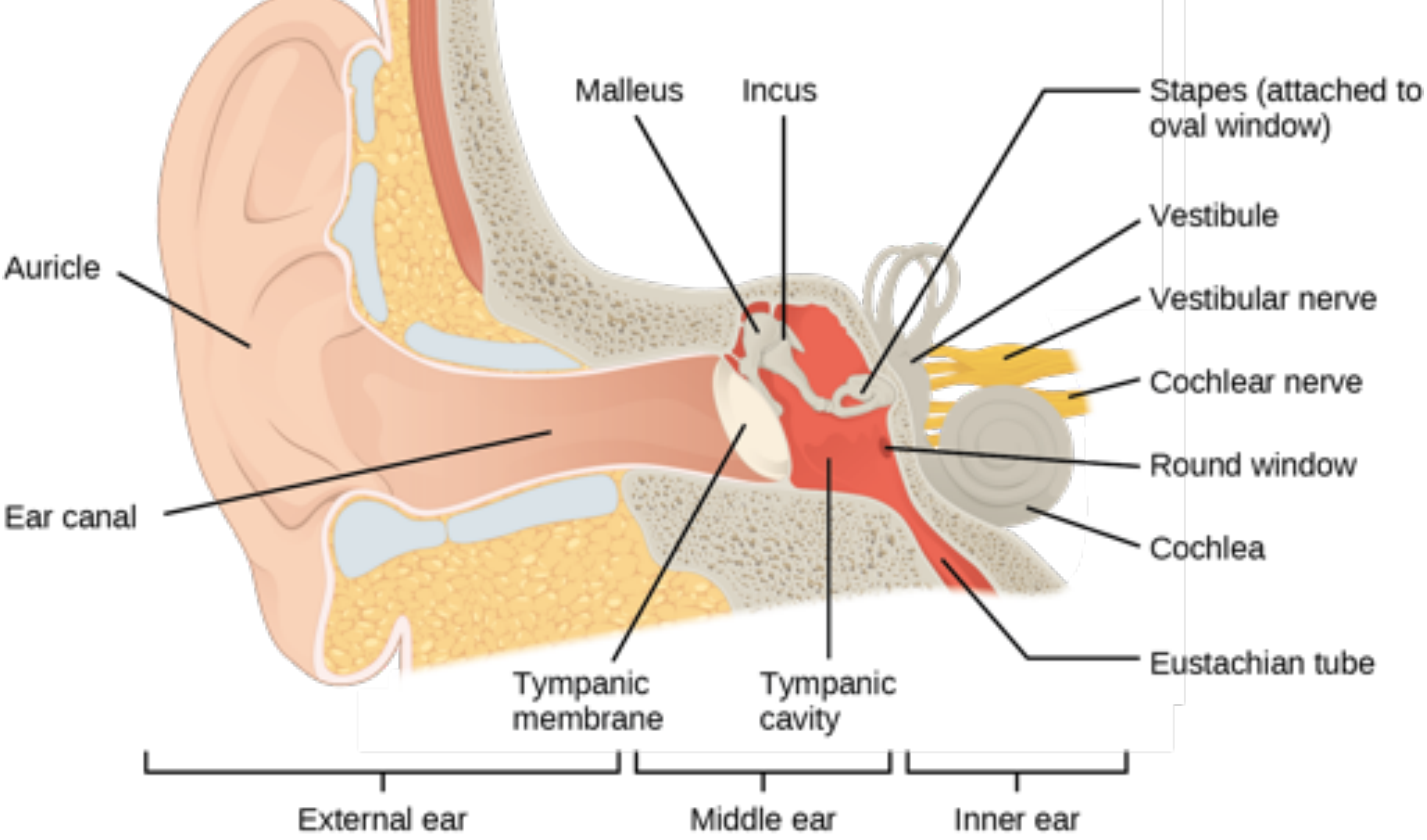


Image credit: OpenStax CC-BY-4.0 - Access for free at https://openstax.org/books/biology-2e/pages/1-introduction

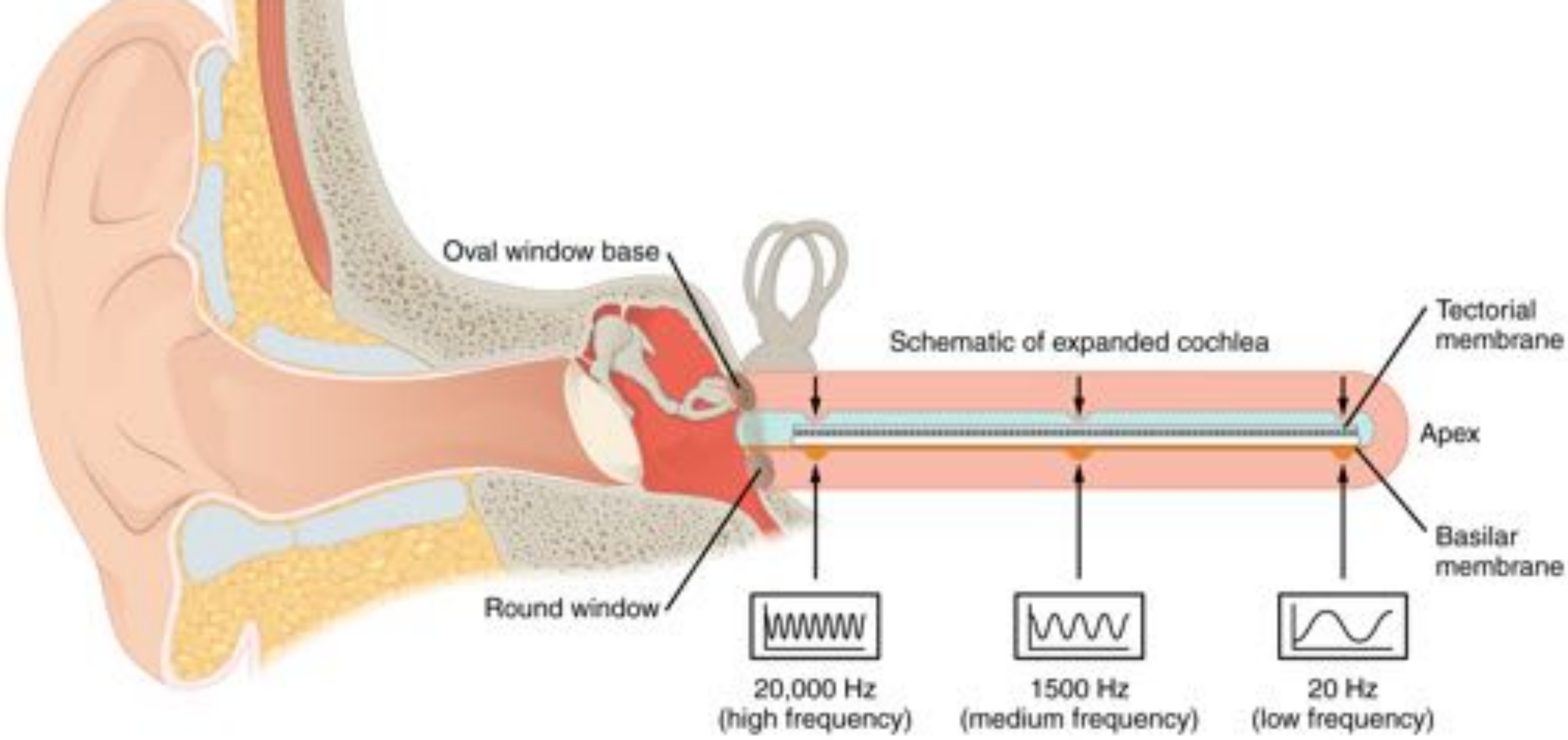
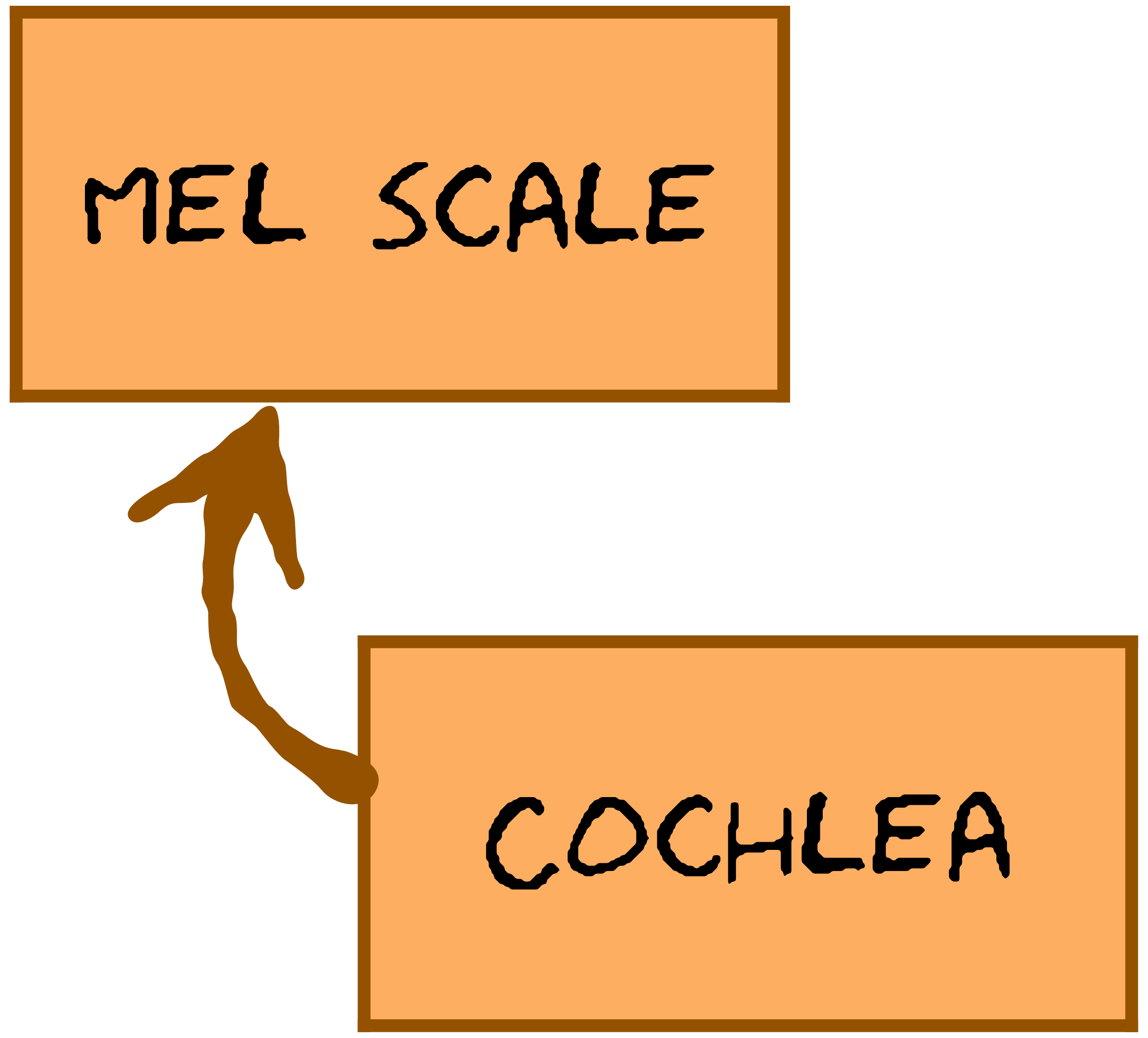
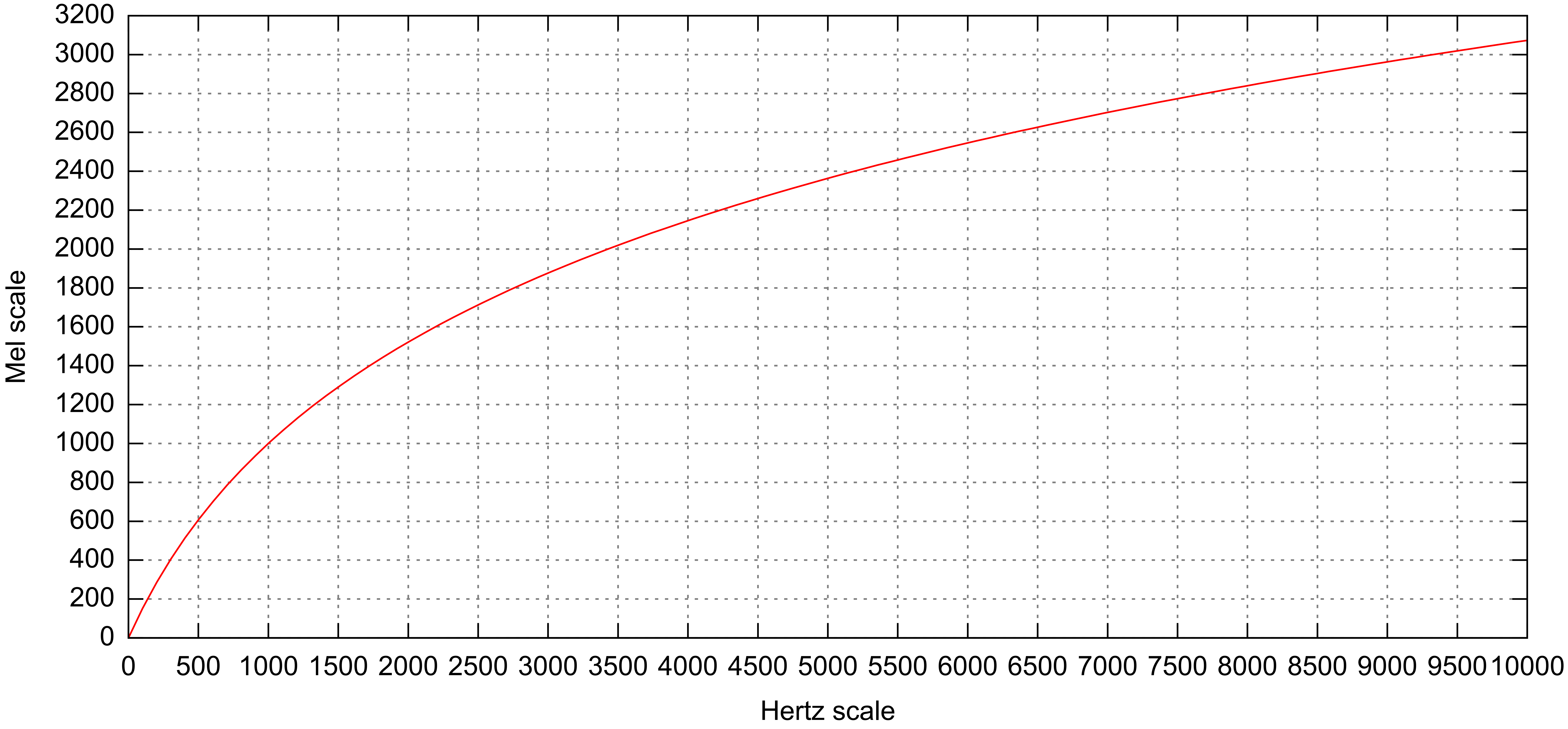


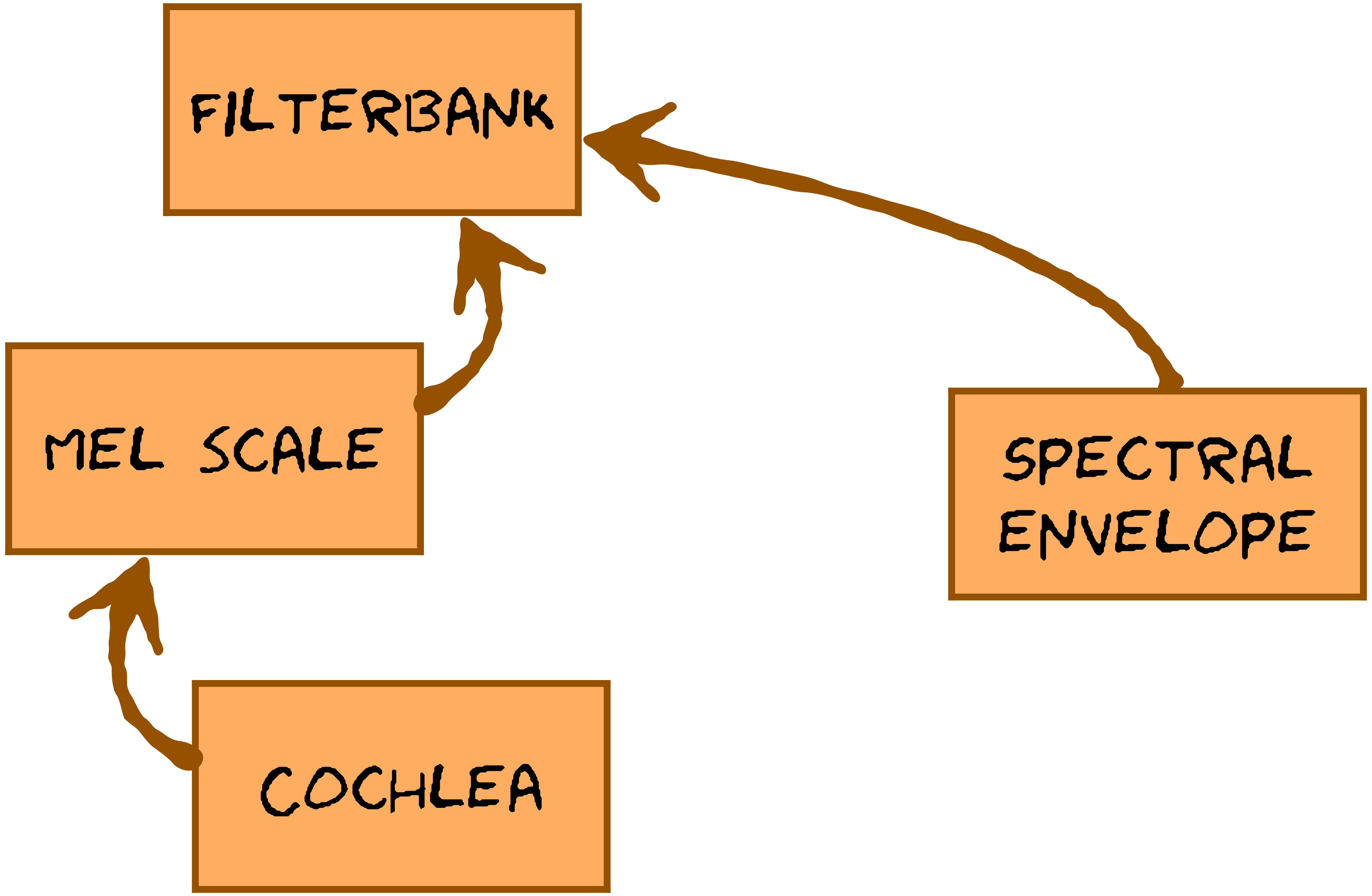
Image credit: OpenStax CC-BY-4.0 - Access for free at https://openstax.org/books/biology-2e/pages/1-introduction



MEL SCALE

COCHLEA


3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
Mel scale
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
9000
9500
10000
Hertz scale



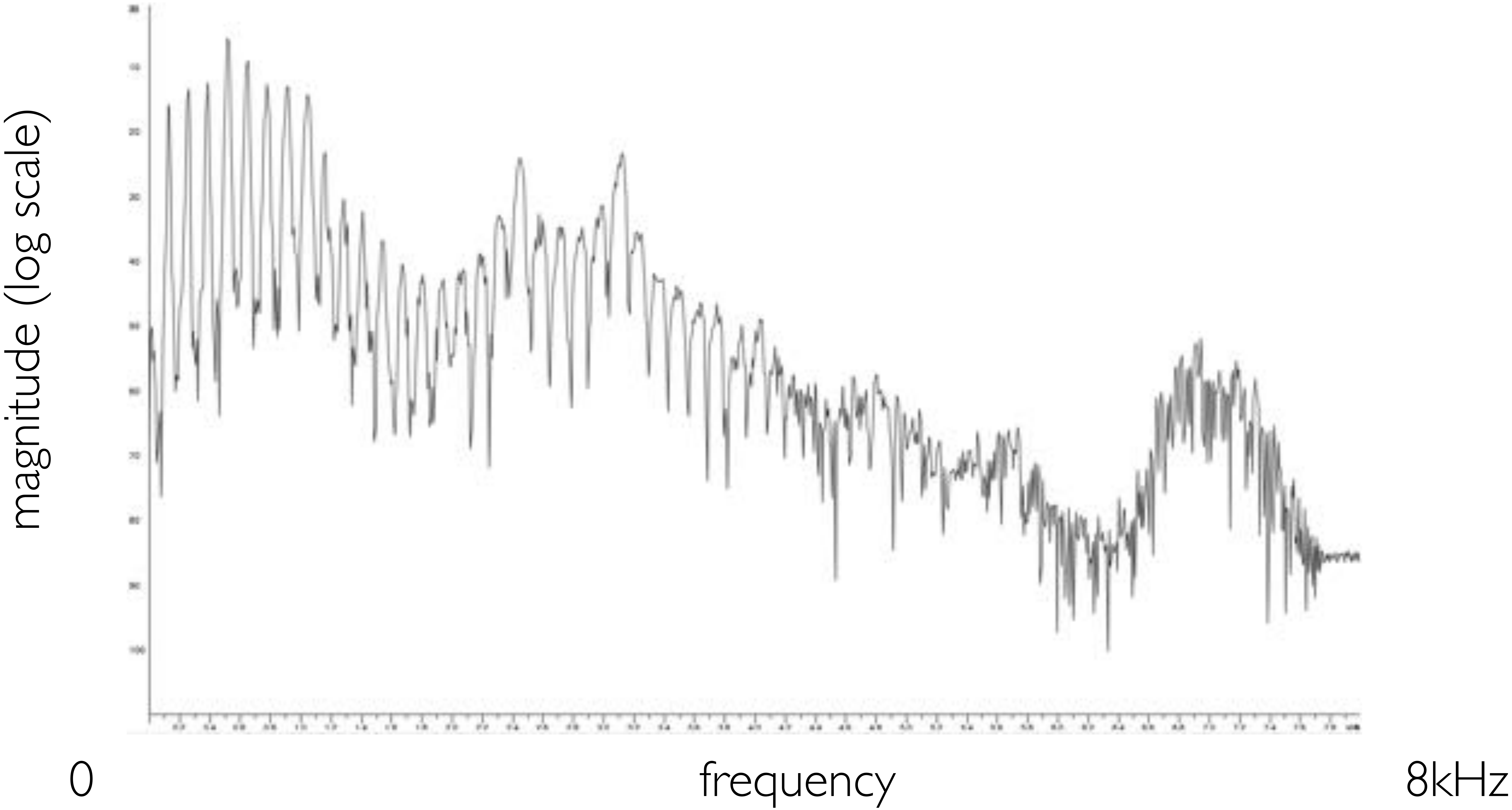
FILTERBANK

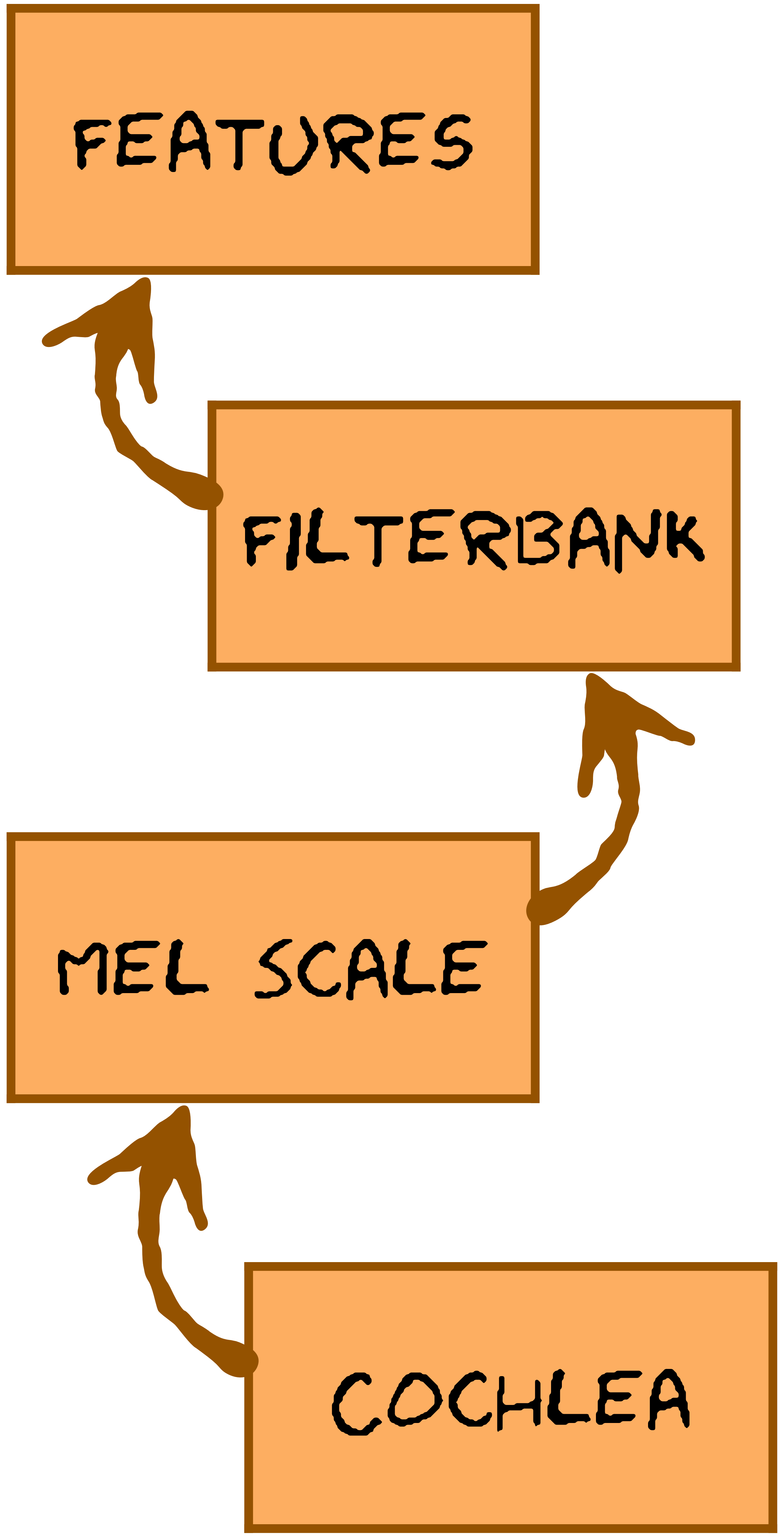
MEL SCALE

SPECTRAL ENVELOPE

COCHLEA

# The auditory system is like a bank of bandpass filters: a “filterbank”





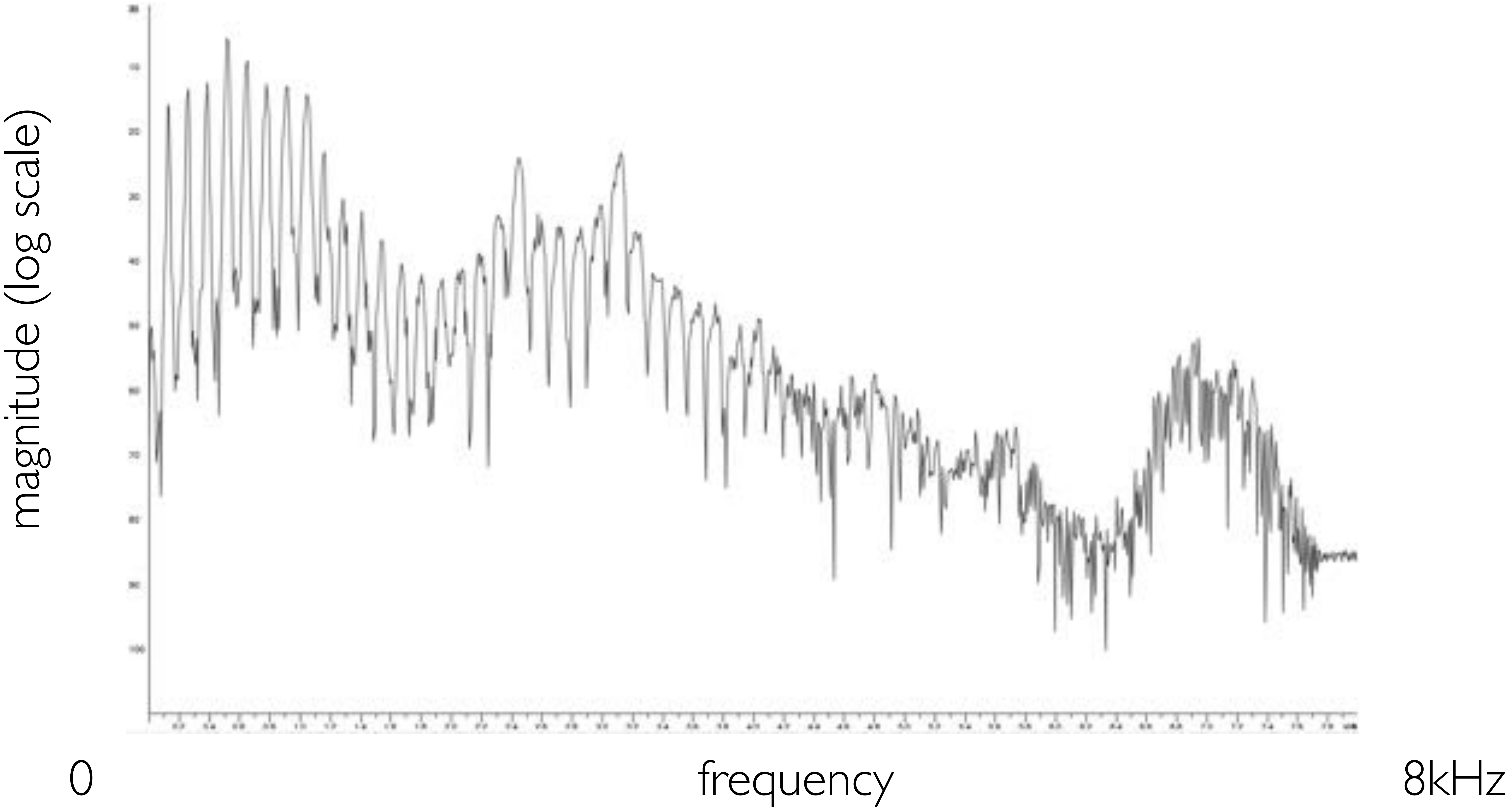
FEATURES

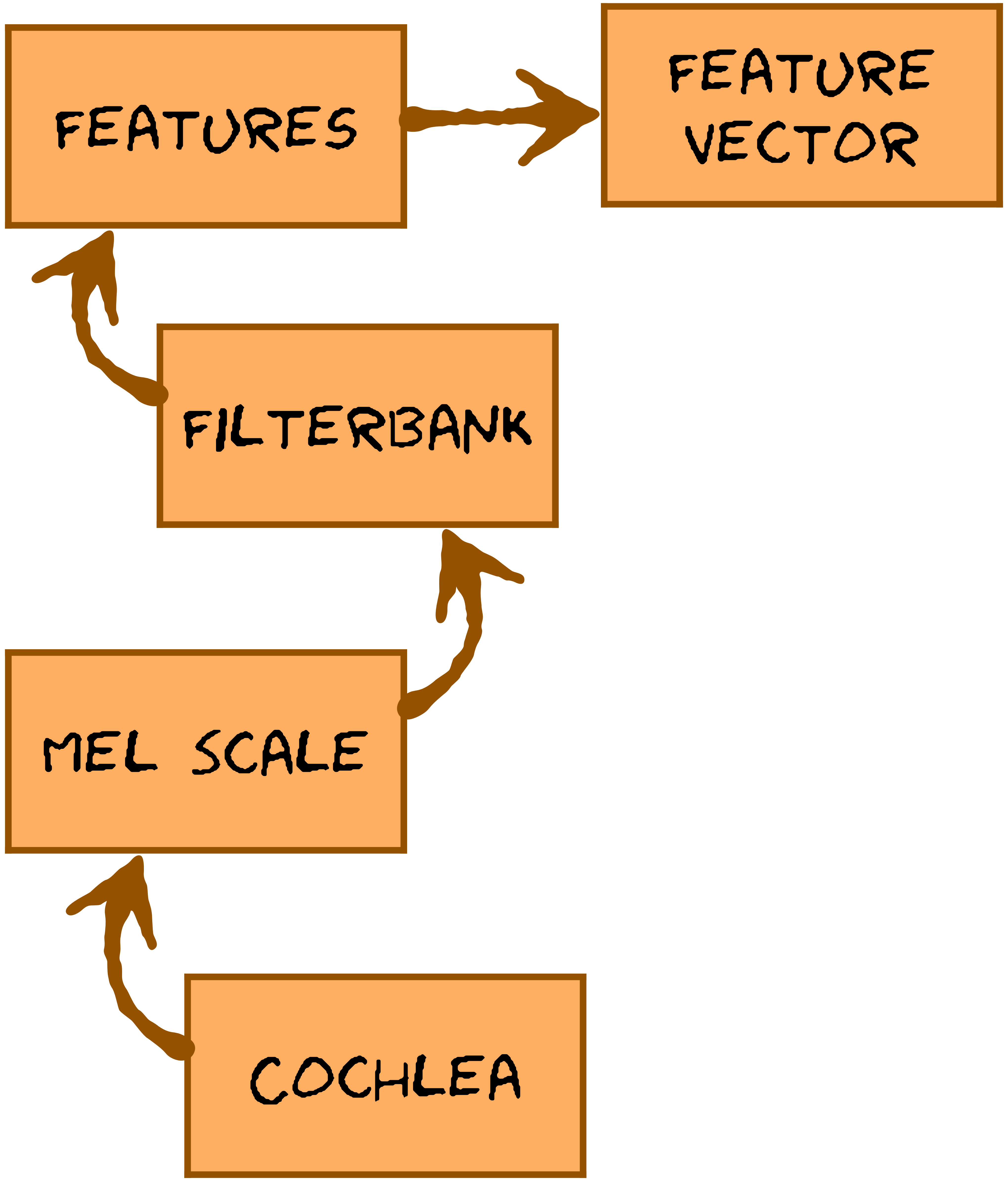
FILTERBANK

MEL SCALE

COCHLEA

# Each filter’s output is a useful feature for doing Automatic Speech Recognition





FEATURES

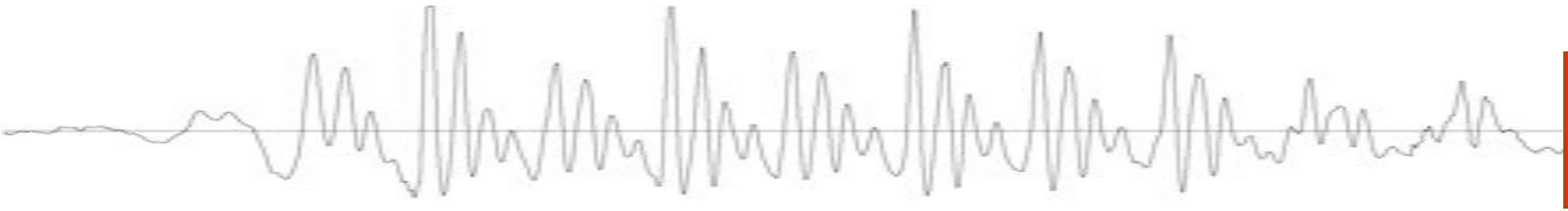
FEATURE VECTOR

FILTERBANK

MEL SCALE

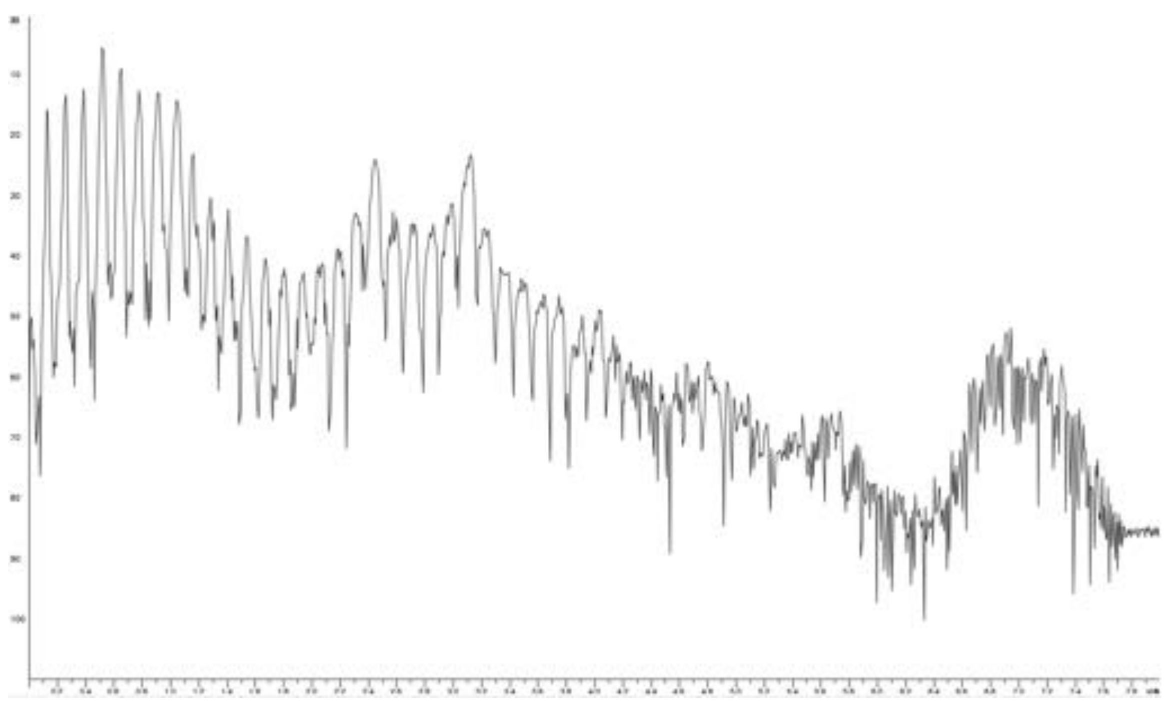
COCHLEA

# Filterbank features for one frame are speech are stored in a single vector

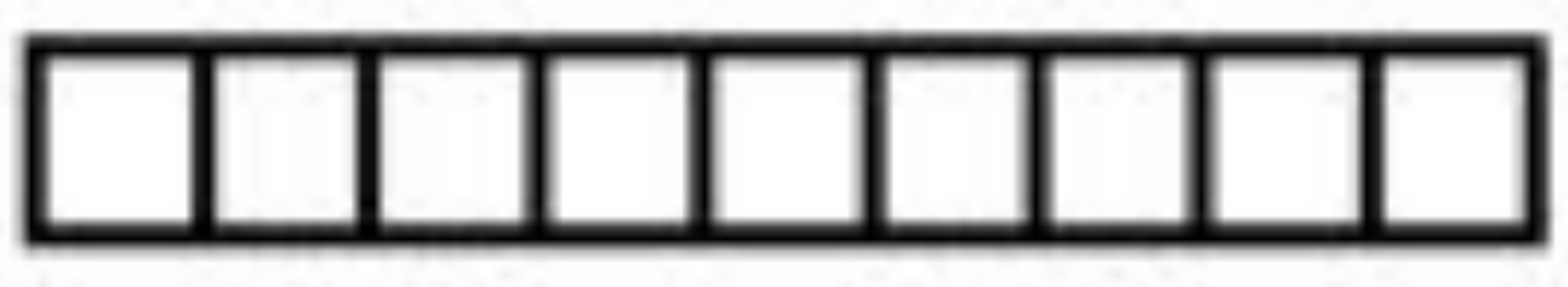


TIME DOMAIN

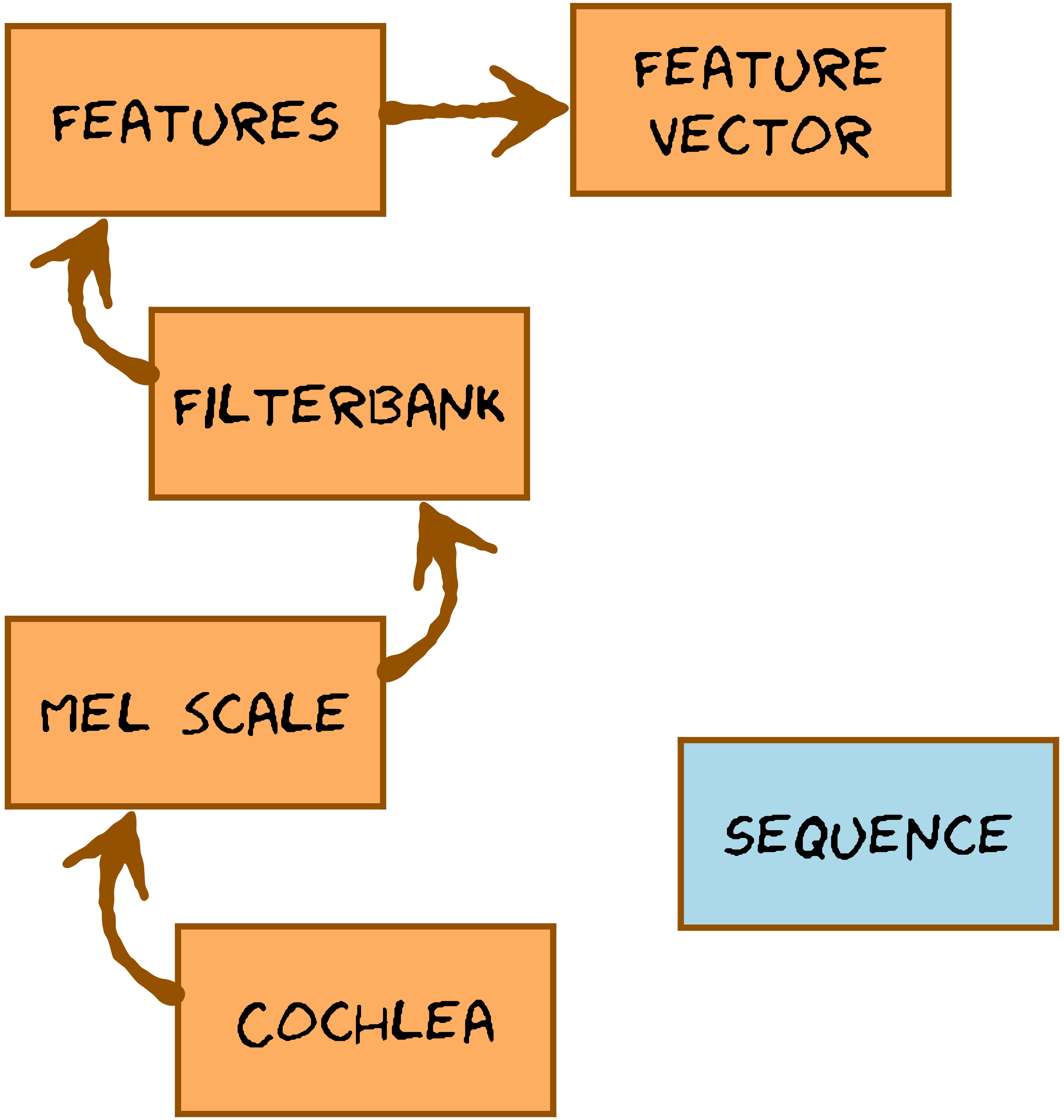
FEATURES



FREQUENCY DOMAIN



FEATURE VECTOR



FEATURES

FEATURE VECTOR

FILTERBANK

MEL SCALE

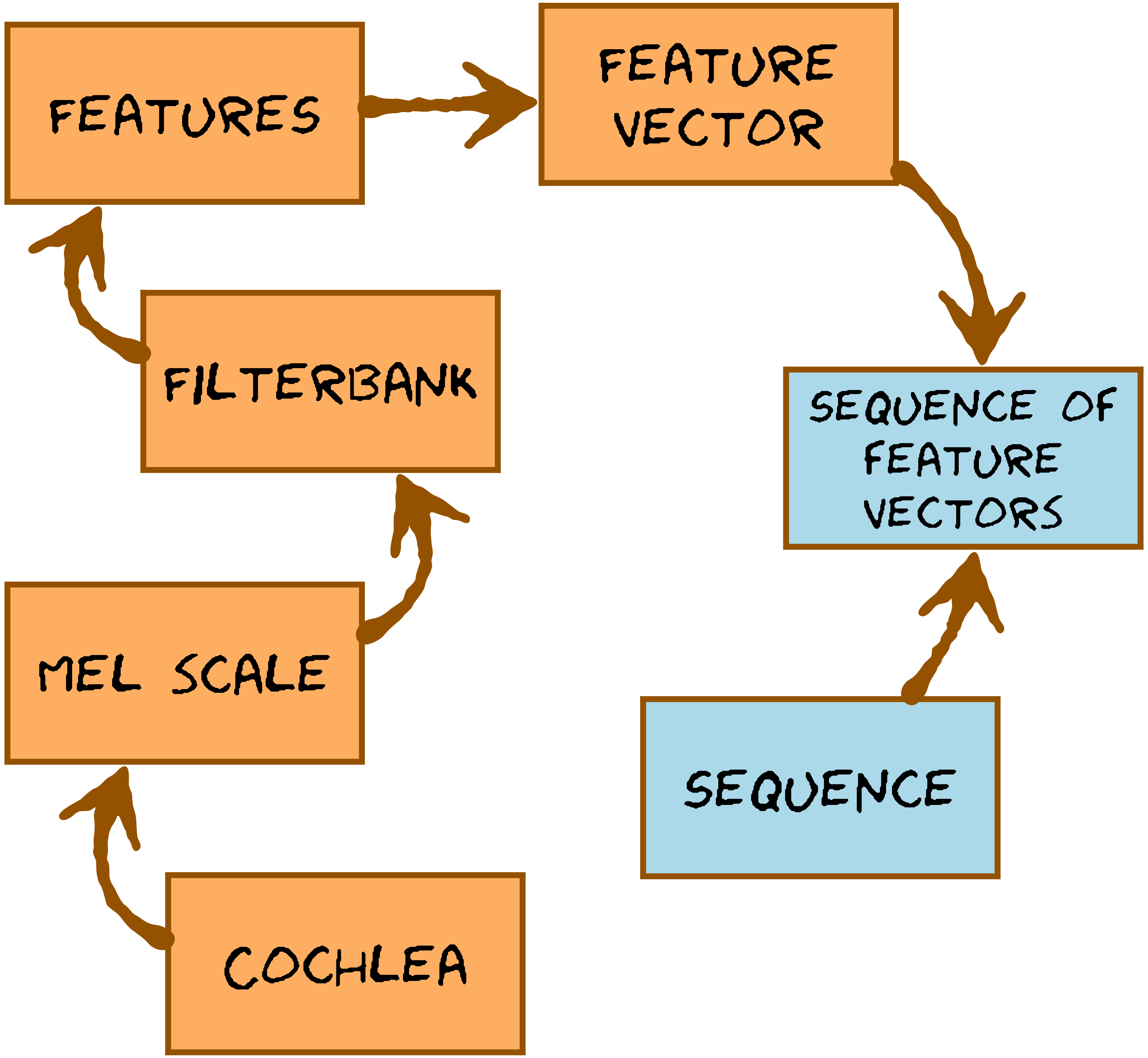
COCHLEA

SEQUENCE

SEQUENCE

# **Sequences** are *everywhere* in language

- We've already seen
  - a waveform is a sequence of **samples**
  - a waveform can be analysed as a sequence of overlapping analysis **frames**
  - a sentence is a sequence of **words**
  - a spoken word is a sequence of **phones**
  - a written word is a sequence of **letters**
- Now we have
  - from each frame we extract a feature vector
  - so a waveform becomes a **sequence of feature vectors**



FEATURES

FEATURE VECTOR

FILTERBANK

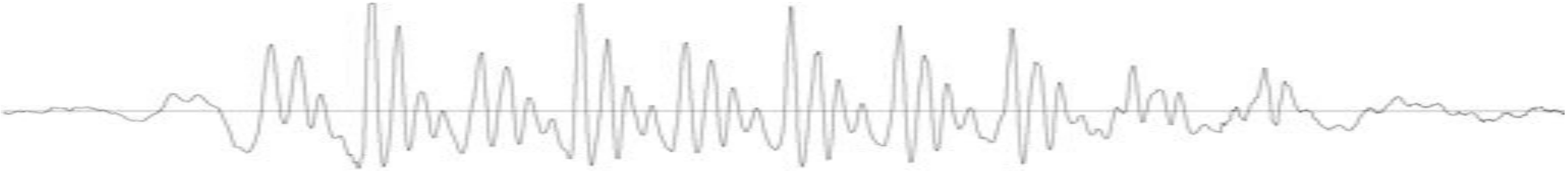
SEQUENCE OF FEATURE VECTORS

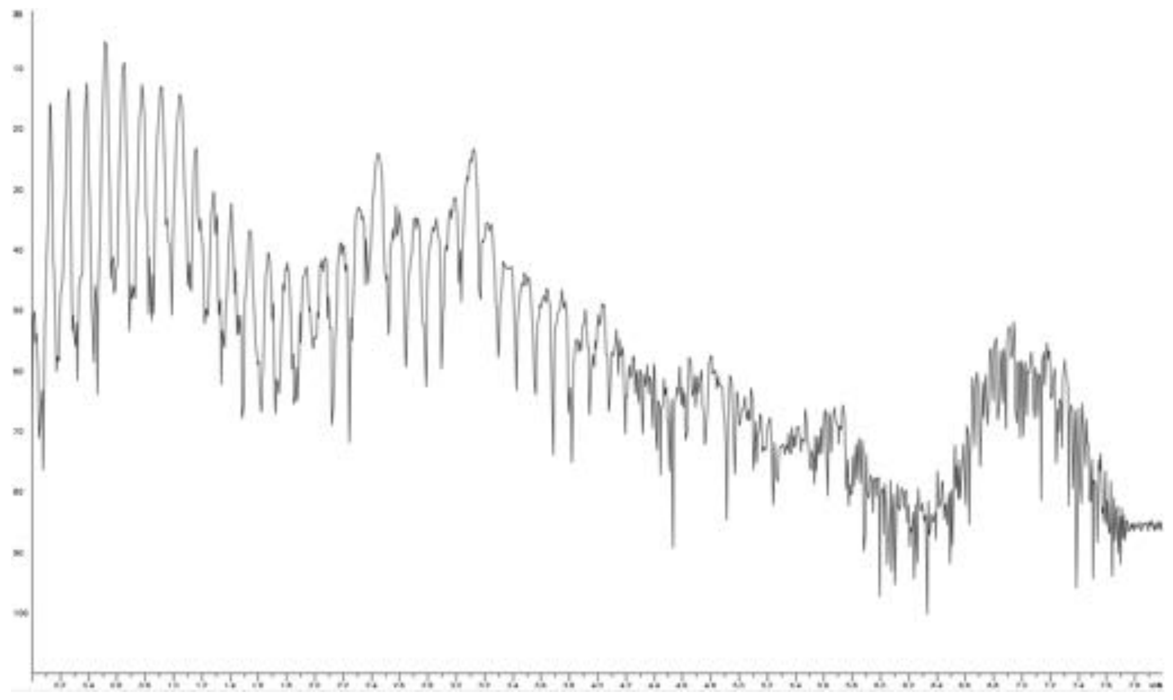
MEL SCALE

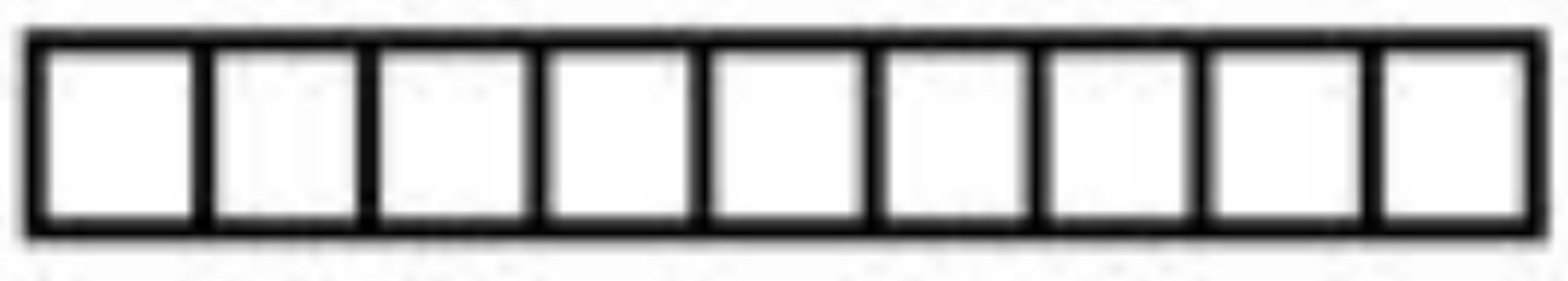
SEQUENCE

COCHLEA

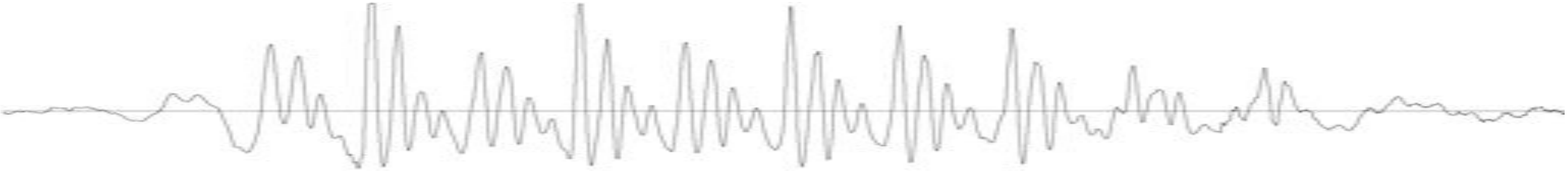
# Filterbank features for one frame are speech are stored in a single vector

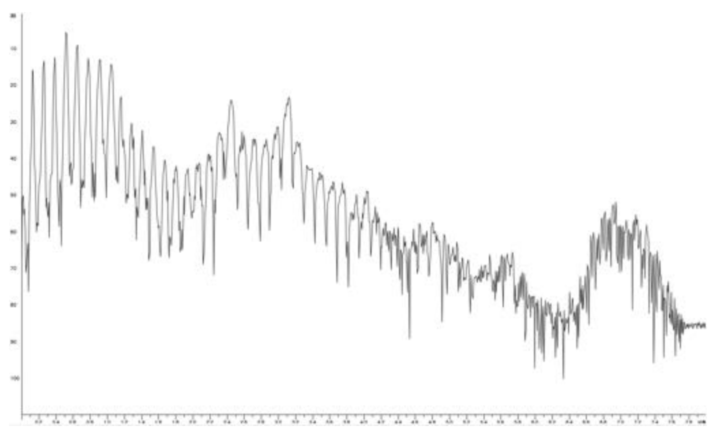






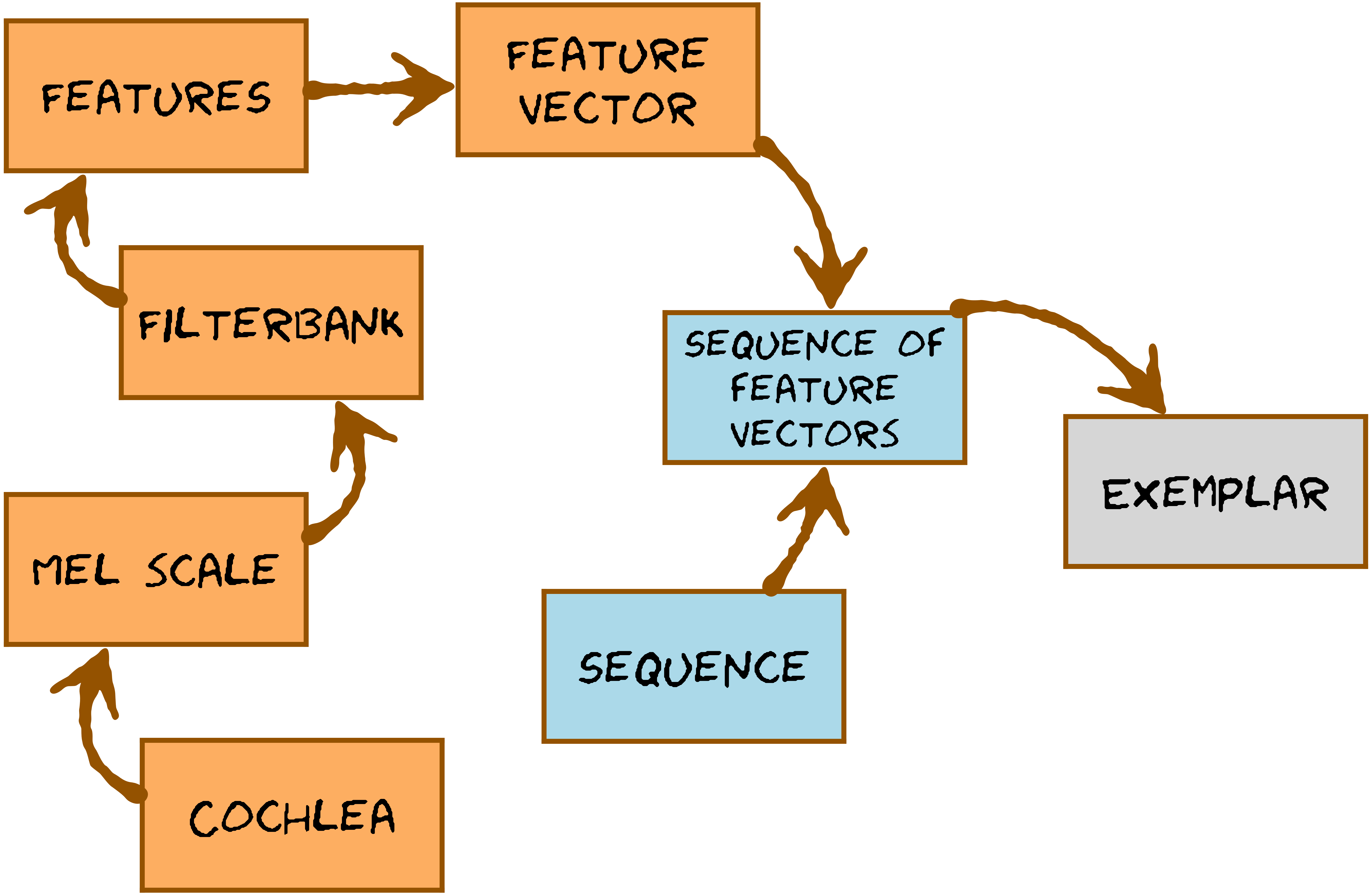
# Filterbank features for automatic speech recognition





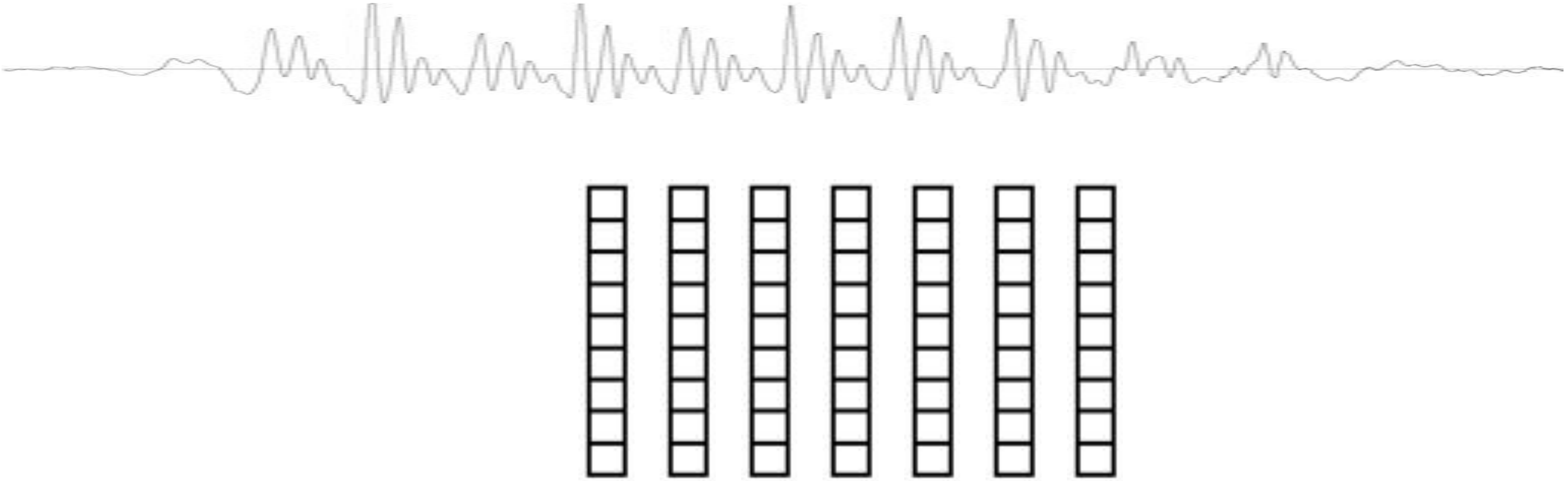
SEQUENCE OF FEATURE VECTORS

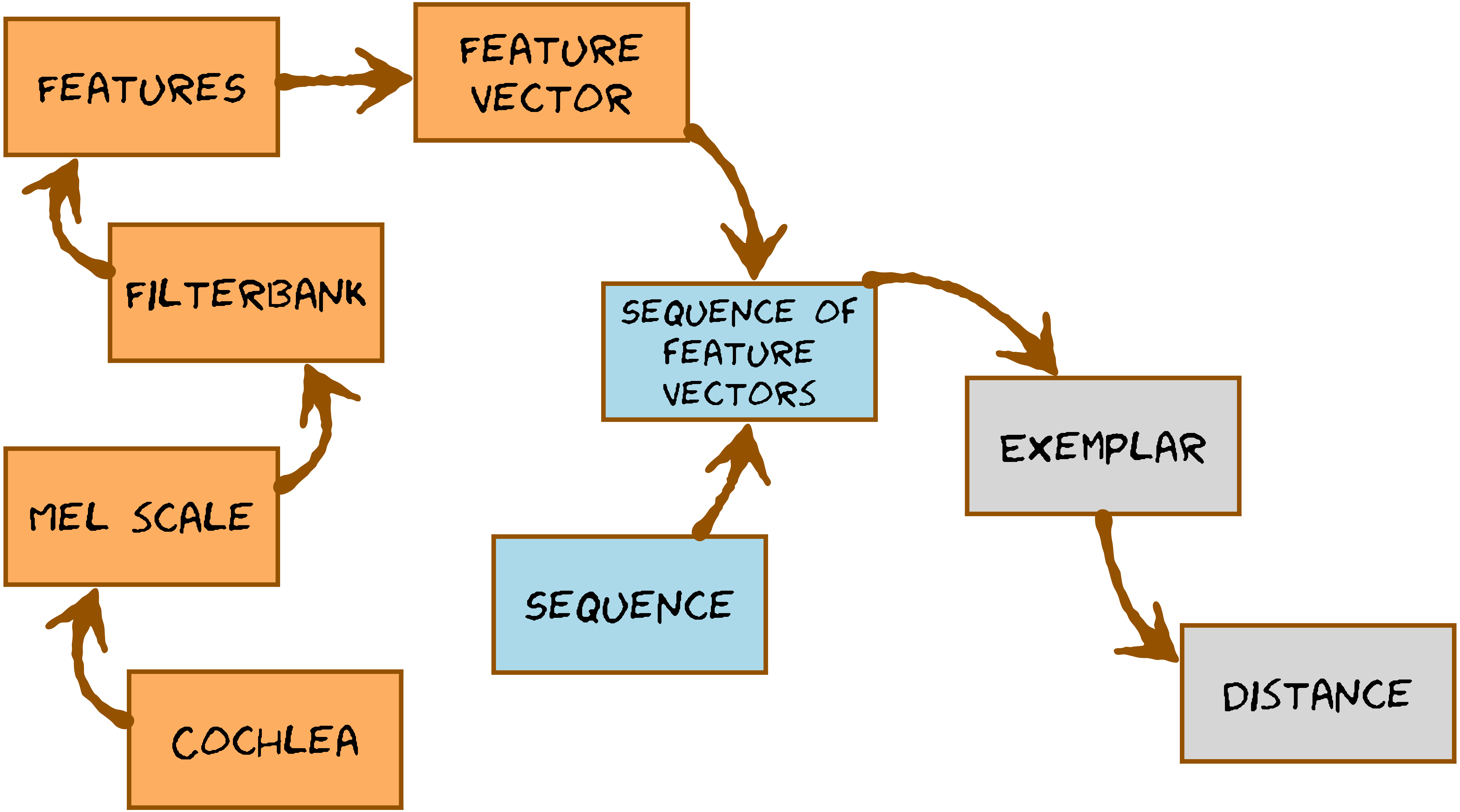
# Filterbank features for automatic speech recognition


FEATURES
FEATURE VECTOR
FILTERBANK
SEQUENCE OF FEATURE VECTORS
EXEMPLAR
MEL SCALE
SEQUENCE
COCHLEA

# “three”

EXEMPLAR

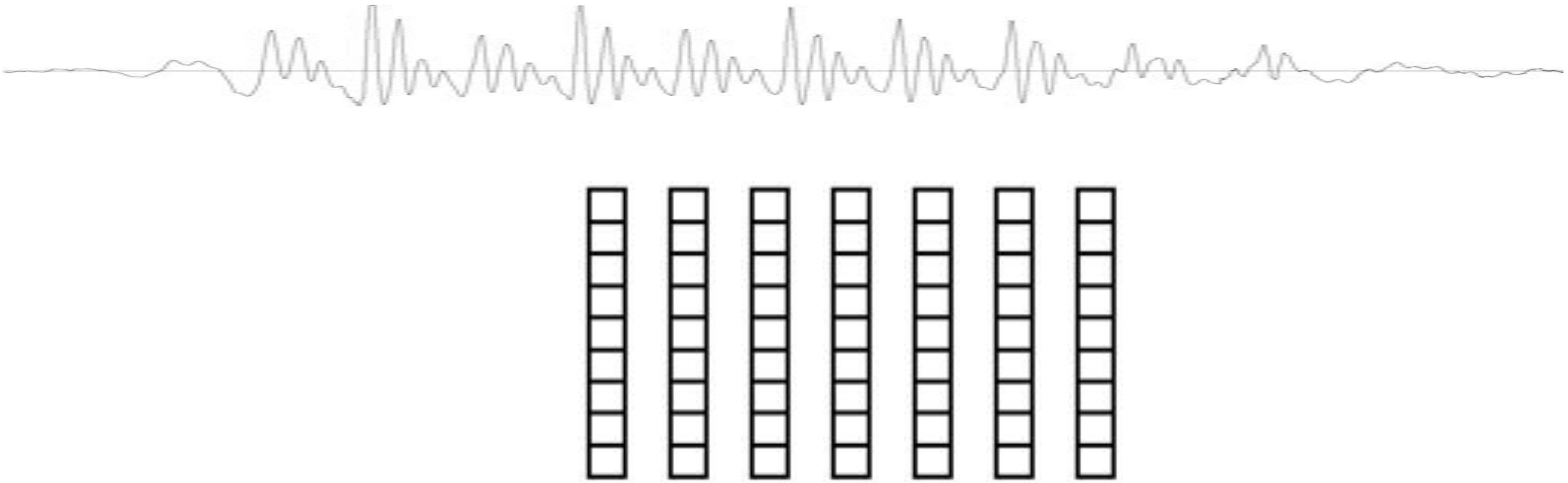




FEATURES
FEATURE VECTOR
FILTERBANK
SEQUENCE OF FEATURE VECTORS
MEL SCALE
EXEMPLAR
SEQUENCE
COCHLEA
DISTANCE

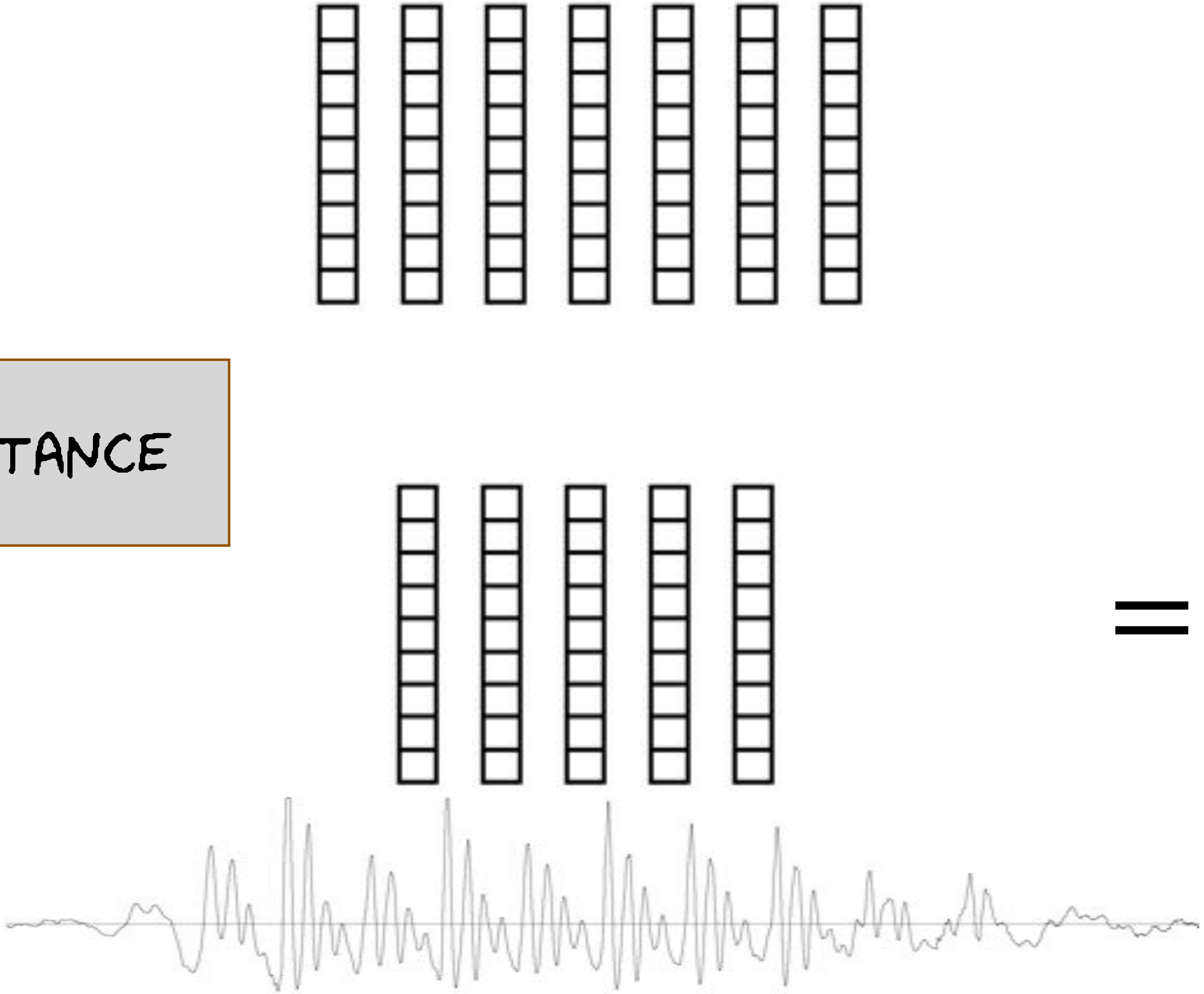
# “three”

EXEMPLAR



EXEMPLAR

“three”



DISTANCE

“???”

global distance

$$= \sum \text{local distances}$$

EXEMPLAR

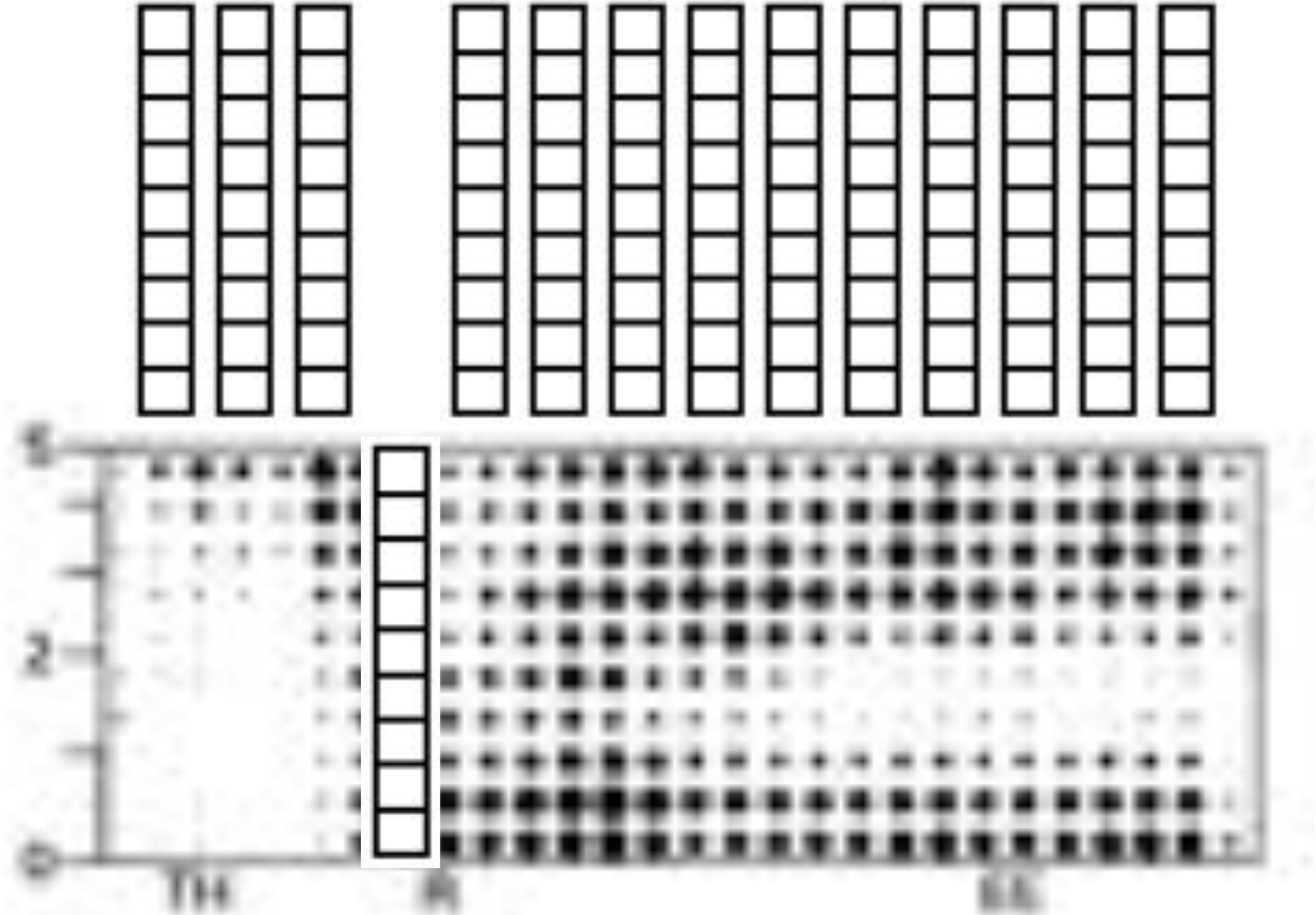
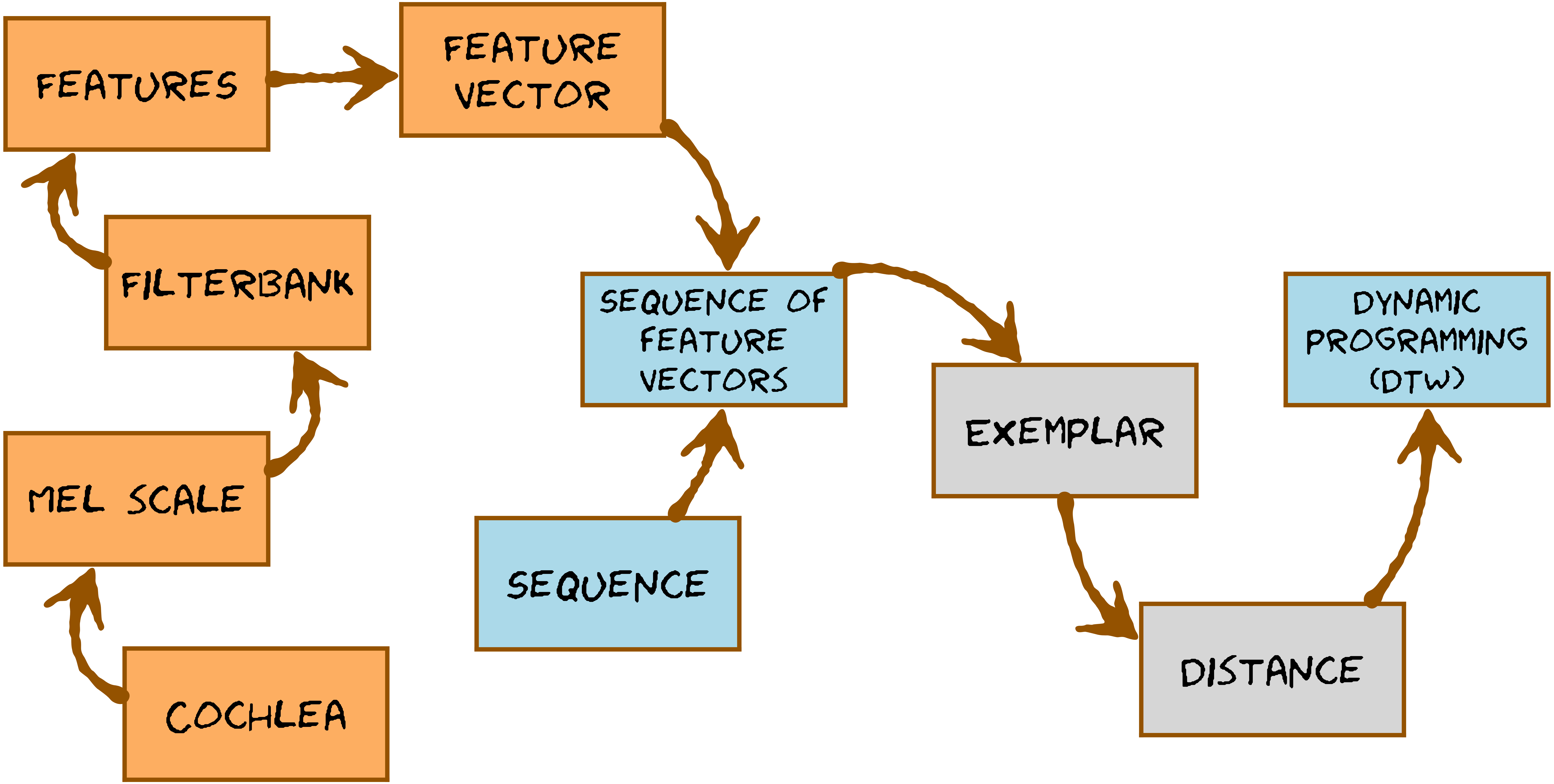
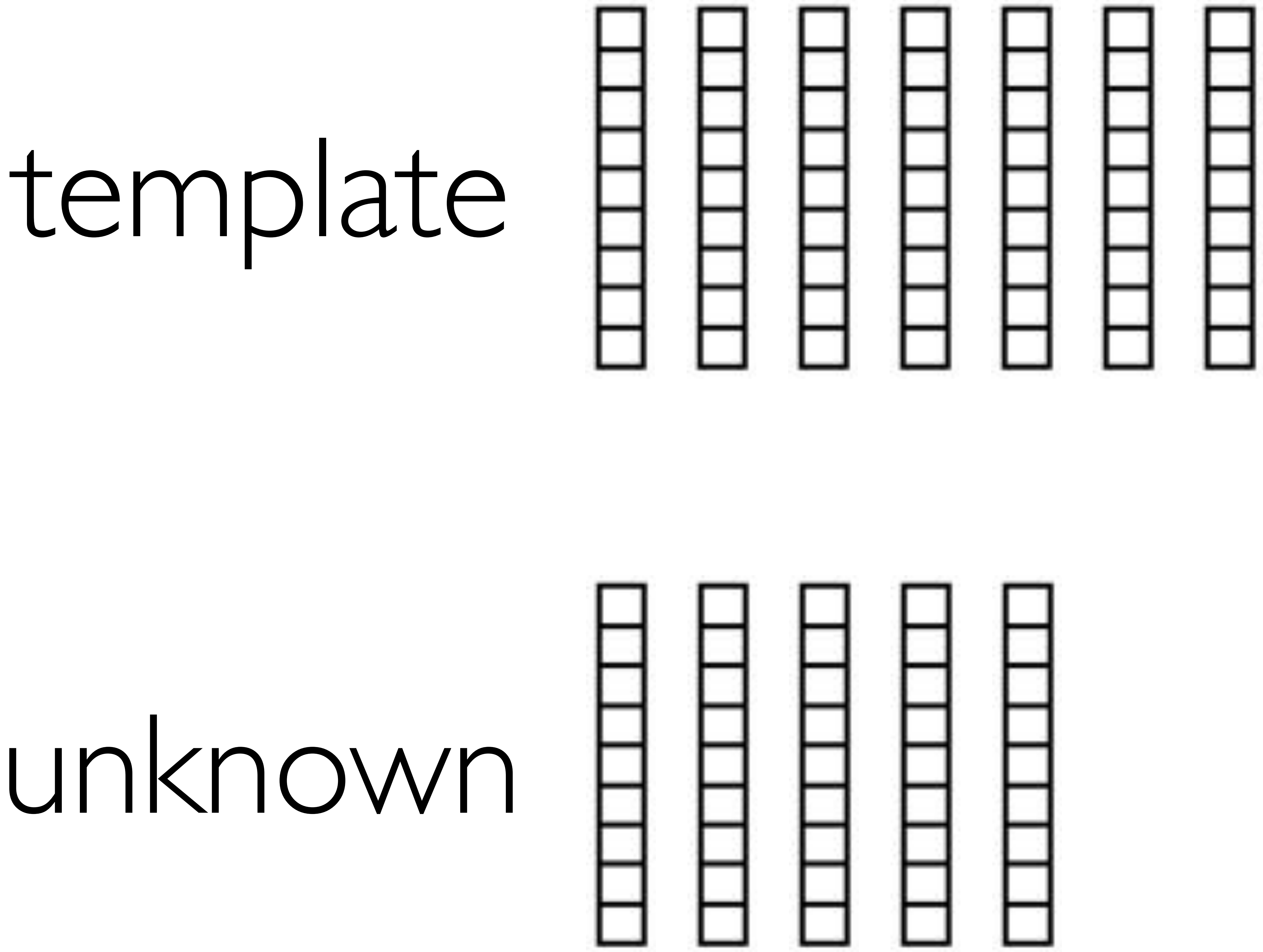


Image credit: Figure 8.1 from Holmes & Holmes


FEATURES
FEATURE VECTOR
FILTERBANK
SEQUENCE OF FEATURE VECTORS
EXEMPLAR
DYNAMIC PROGRAMMING (DTW)
MEL SCALE
SEQUENCE
DISTANCE
COCHLEA

# Pattern matching by Dynamic Time Warping



template

unknown

| | | |
|---|---|---|
| 1,1 | 1,1 | 1,1 |
| 2,2 | 2,2 | 2,1 |
| 3,3 | 3,2 | 3,2 |
| **4,3** | **4,3** | **4,3** |
| 5,? | 5,? | 5,? |
| 6,? | 6,? | 6,? |
| 7,? | 7,? | 7,? |

# Dynamic Time Warping is a form of Dynamic Programming

- Understanding Dynamic Programming, as an algorithm

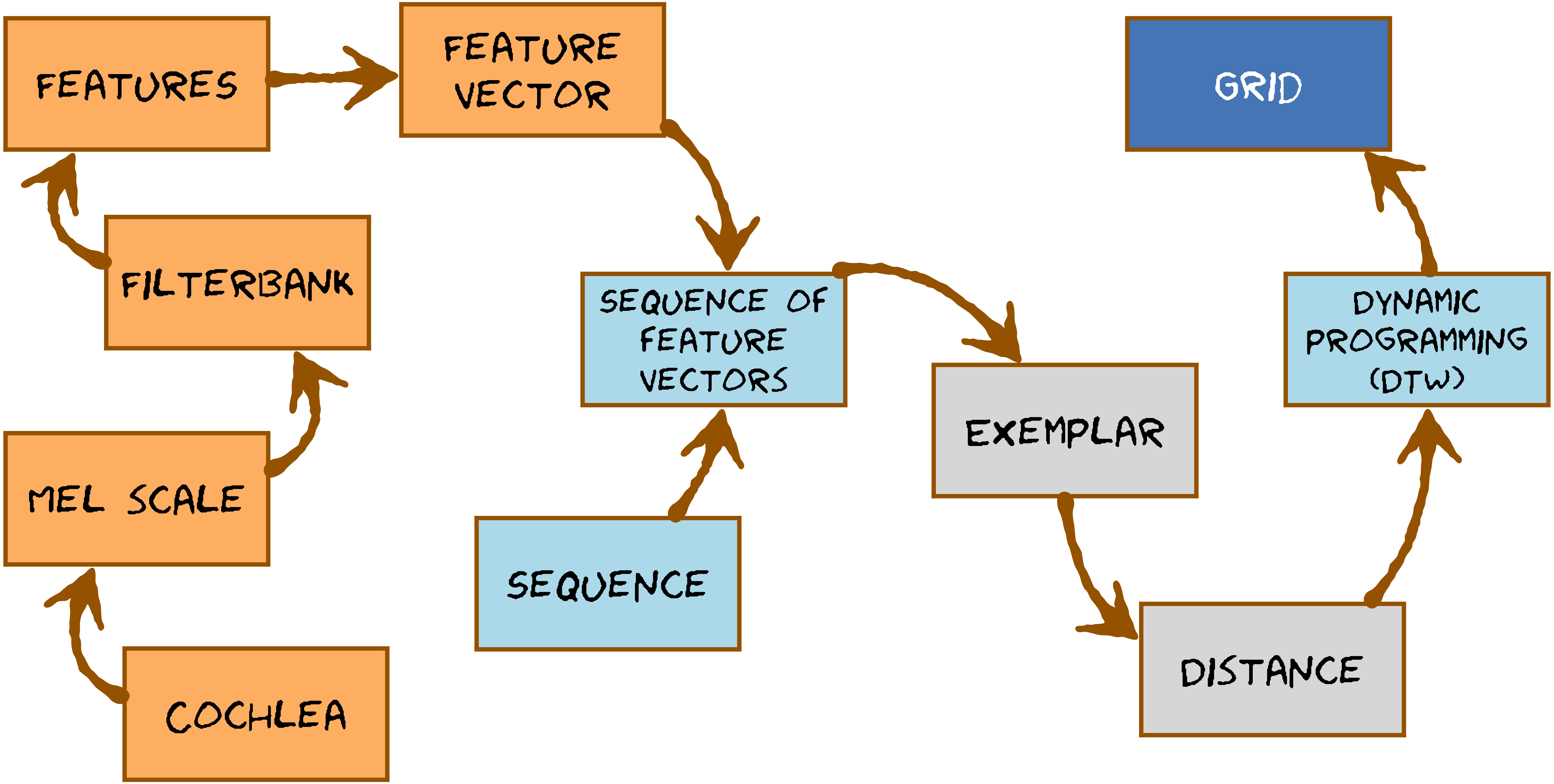
**Getting harder**

- Being able to see that Dynamic Programming can be applied to a particular problem

**Really quite difficult**

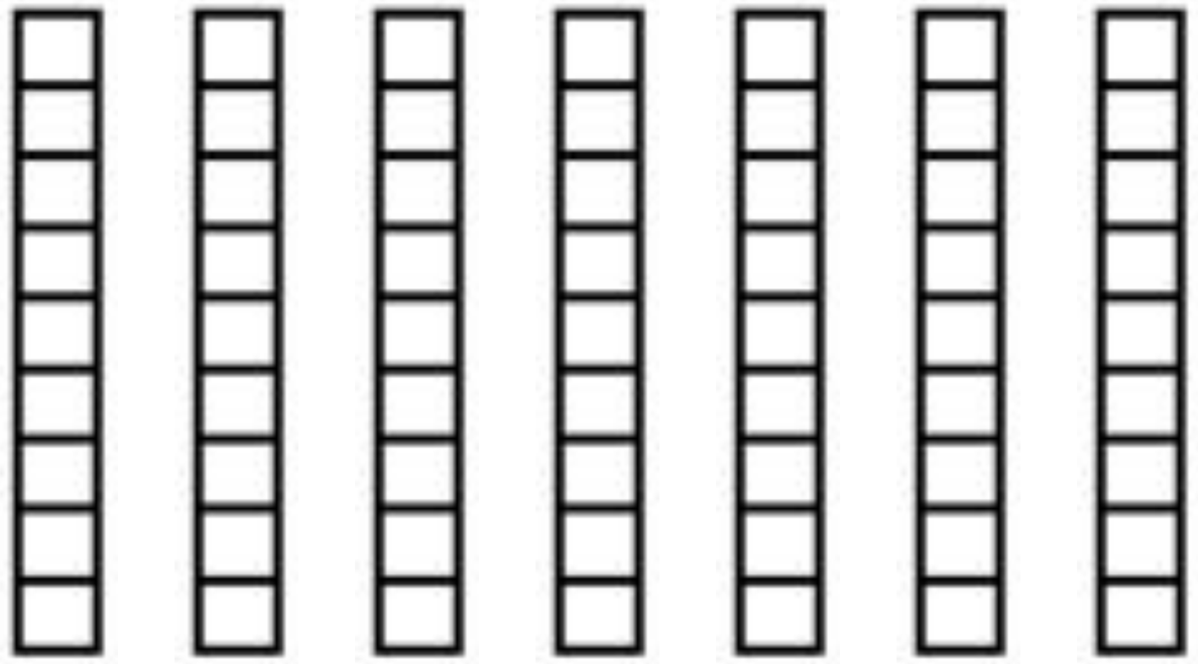
- Devising a suitable data structure for that problem

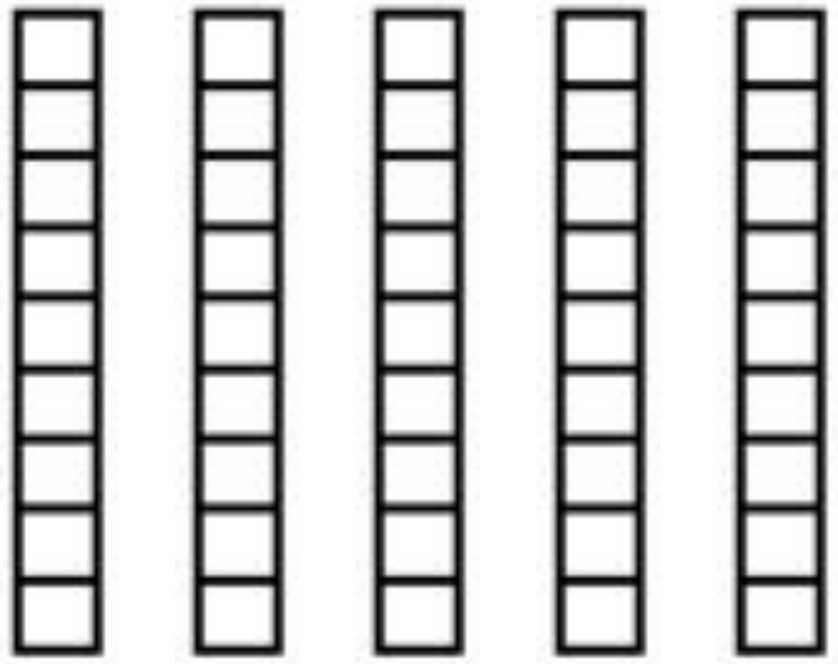
**My brain hurts**



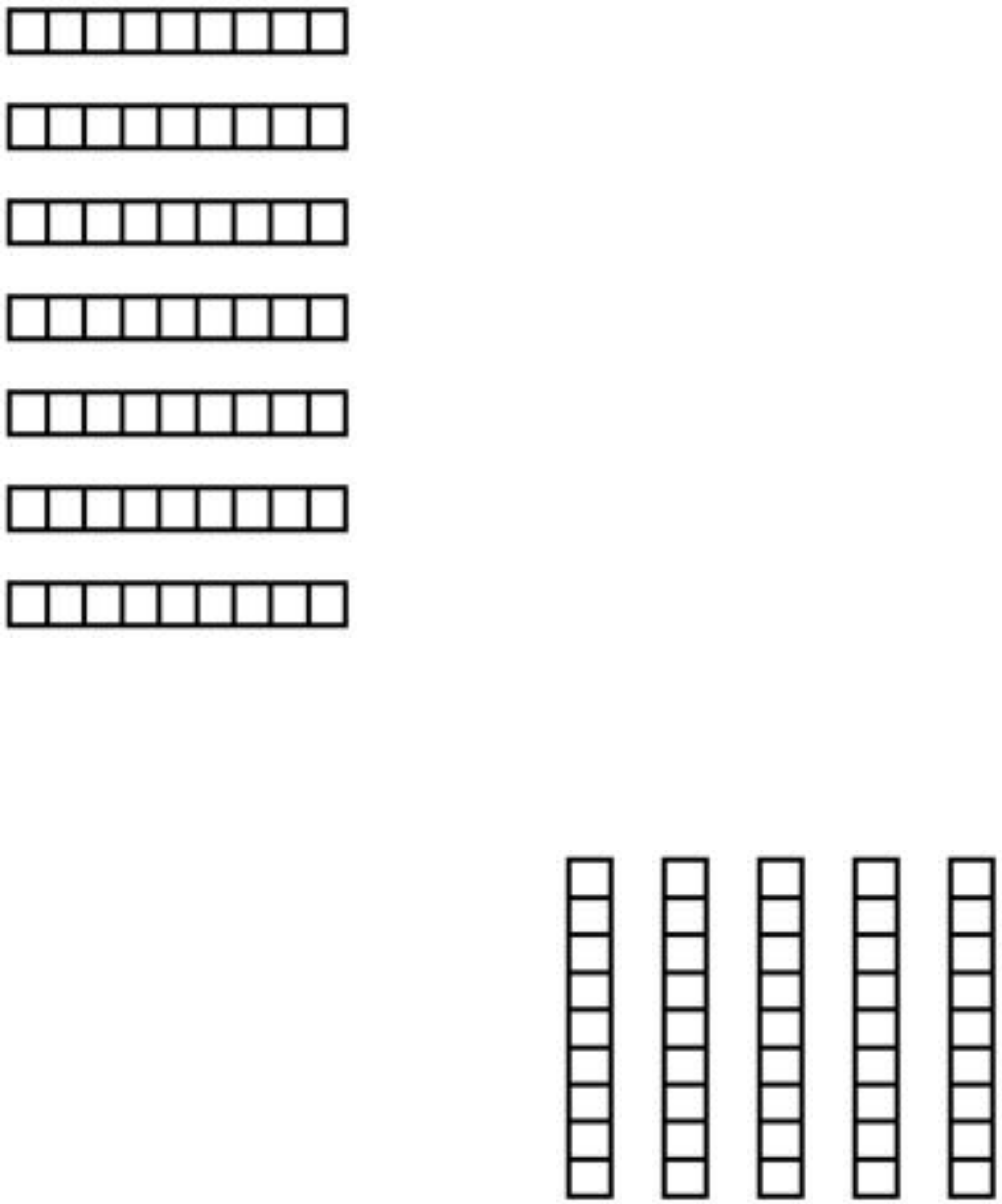
FEATURES
FEATURE VECTOR
GRID
FILTERBANK
SEQUENCE OF FEATURE VECTORS
DYNAMIC PROGRAMMING (DTW)
EXEMPLAR
MEL SCALE
SEQUENCE
DISTANCE
COCHLEA

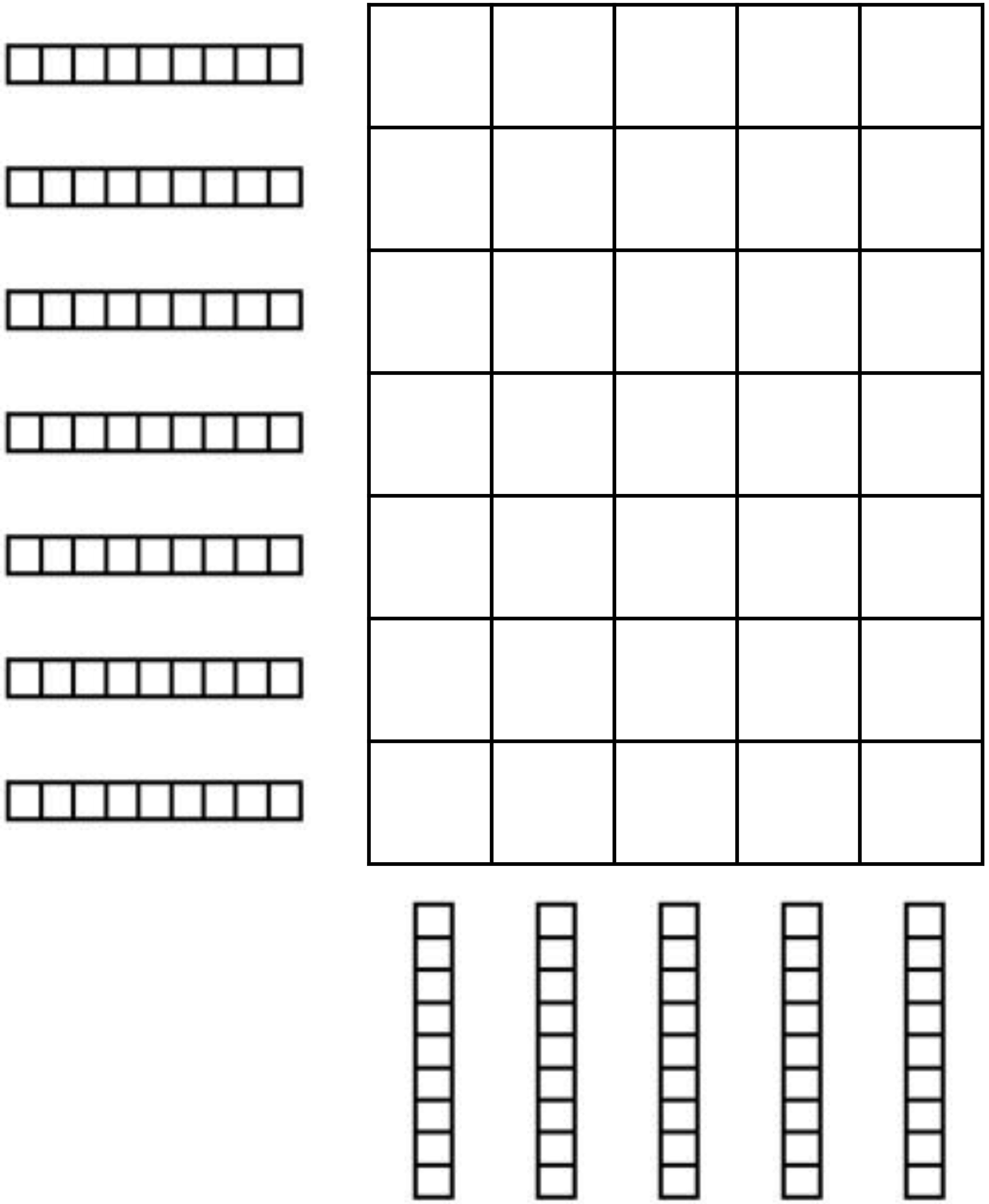
‘‘three’’





‘‘???’’





GRID

SEQUENCE OF FEATURE VECTORS

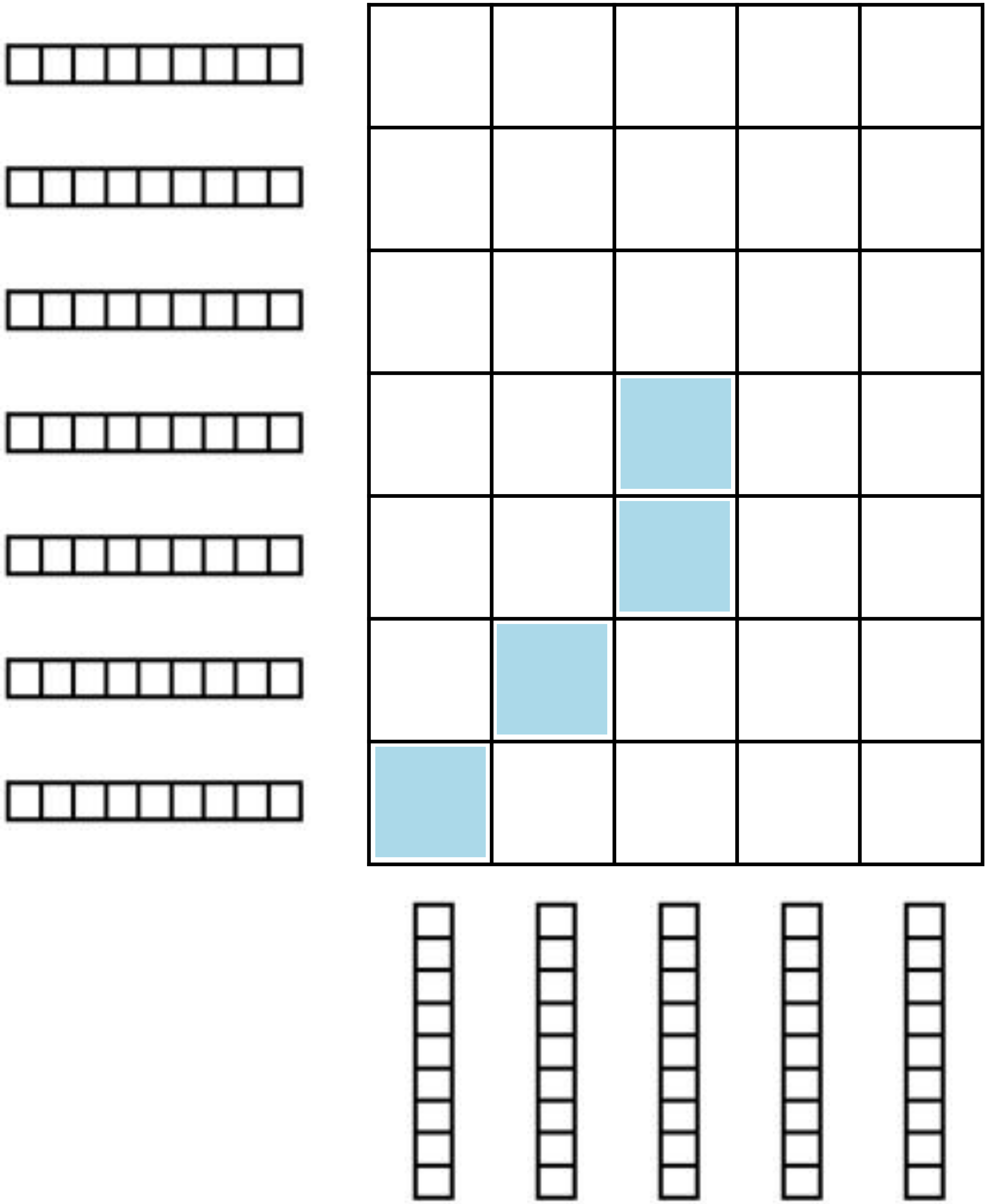
global

DISTANCE

local

SEQUENCE OF FEATURE VECTORS

DYNAMIC PROGRAMMING (DTW)



1 , 1

2 , 2

3 , 3

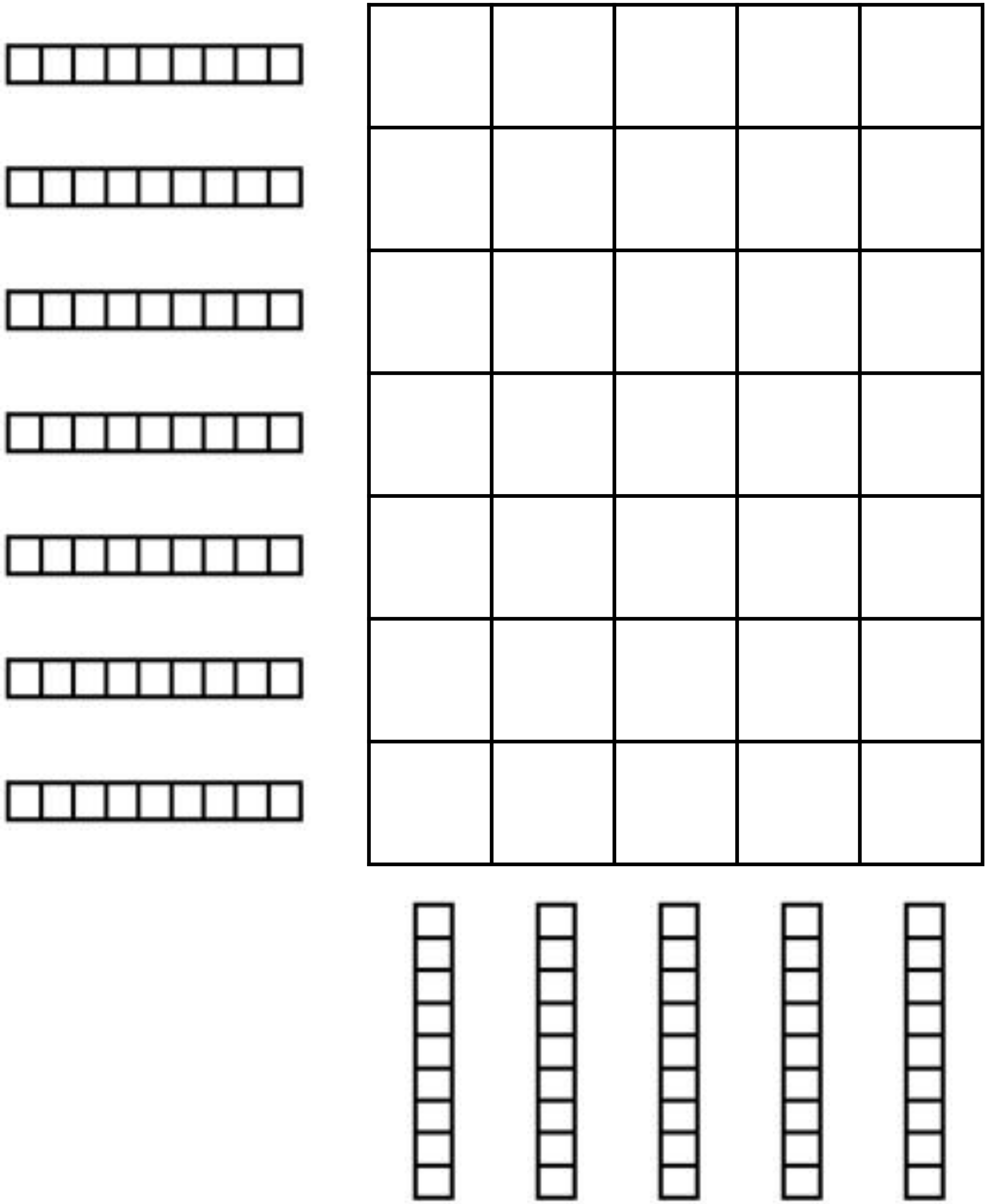
**4 , 3**

5 , ?

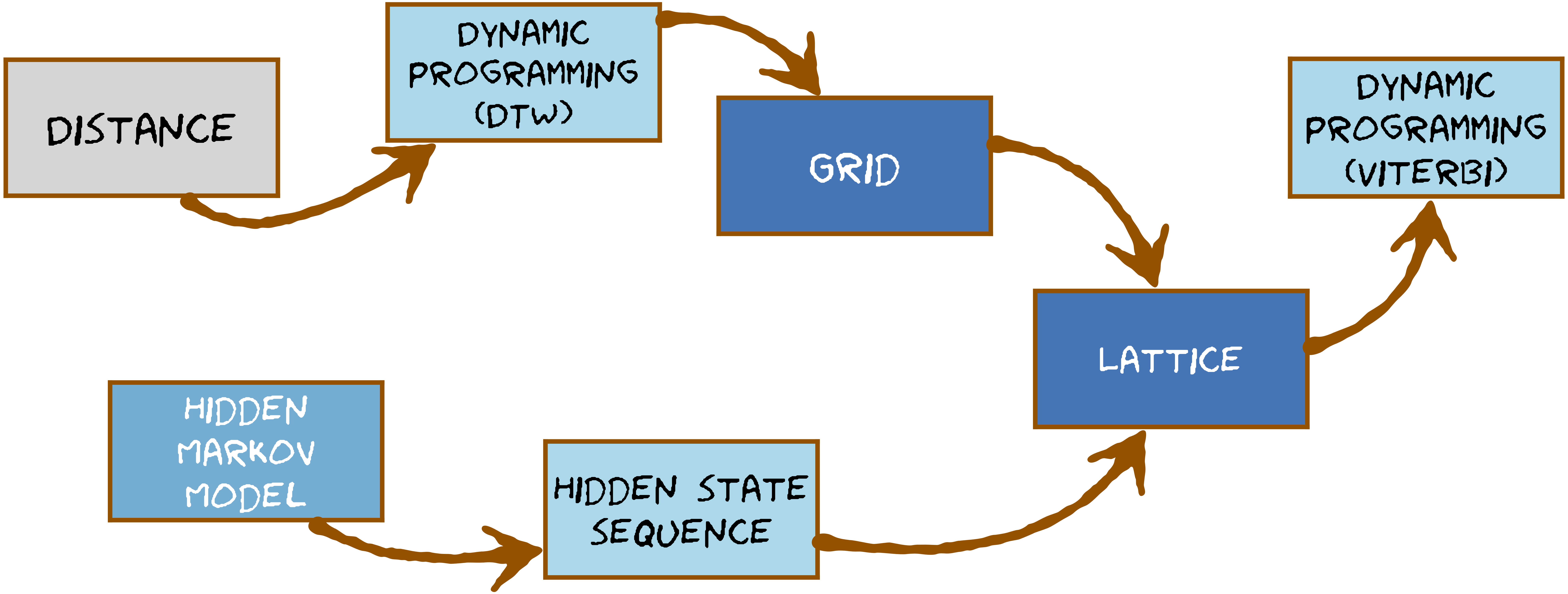
6 , ?

7 , ?

DYNAMIC PROGRAMMING (DTW)



# What you can learn next


DISTANCE

DYNAMIC PROGRAMMING (DTW)

GRID

LATTICE

DYNAMIC PROGRAMMING (VITERBI)

HIDDEN MARKOV MODEL

HIDDEN STATE SEQUENCE

# What next?

- DTW, and especially the local distance measure doesn't account for **variability**
  - so we'll replace it with a **probabilistic model**

HMMs in Module 9

- That model will use Gaussian probability density functions
  - to make these simpler, we will first try to remove covariance from our **features**
  - time for some **feature engineering** !

MFCCs in Module 8