Speech Processing

Simon King
University of Edinburgh

additional class slides for 2020-21
Module 6

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
COCHLEA
Cochlea

Mel scale
The auditory system is like a bank of bandpass filters: a “filterbank”
Each filter’s output is a useful feature for doing Automatic Speech Recognition.
COCHLEA → MEL SCALE → FILTERBANK → FEATURES → FEATURE VECTOR
Filterbank features for one frame are speech are stored in a single vector.
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
COCHLEA \rightarrow MEL SCALE \rightarrow FILTERBANK \rightarrow FEATURES \rightarrow FEATURE VECTOR \rightarrow SEQUENCE OF FEATURE VECTORS \rightarrow SEQUENCE
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
Cochlea

Mel scale

Filterbank

Features

Feature vector

Sequence of feature vectors

Exemplar

Sequence
“three”
Cochlea -> Mel scale -> Filterbank -> Feature vectors -> Sequence of feature vectors -> Exemplar -> Distance
“three”
global distance = \sum \text{local distances}
Pattern matching by Dynamic Time Warping

<table>
<thead>
<tr>
<th>template</th>
<th>1, 1</th>
<th>2, 2</th>
<th>3, 3</th>
<th>4, 3</th>
<th>5, ?</th>
<th>6, ?</th>
<th>7, ?</th>
</tr>
</thead>
<tbody>
<tr>
<td>unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dynamic Time Warping is a form of Dynamic Programming

- Understanding Dynamic Programming, as an algorithm

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

- Devising a suitable data structure for that problem

Getting harder

Really quite difficult

My brain hurts
“three”

“???”
Distance

local

Sequence of feature vectors

 GRID

global

local

Sequence of feature vectors
Dynamic programming (DTW)
<table>
<thead>
<tr>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
<th>Column 5</th>
</tr>
</thead>
</table>

**Dynamic Programming (DTW)**

- 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**

• 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**!