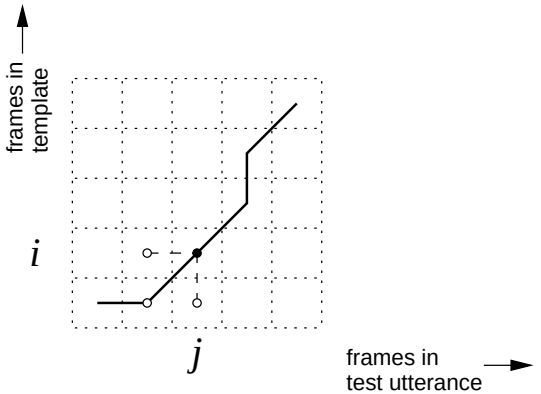
UniS Centre for Vision Speech & Signal Processing
University of Surrey, Guildford GU2 7XH.

Dynamic Time Warping

by Dr Philip Jackson

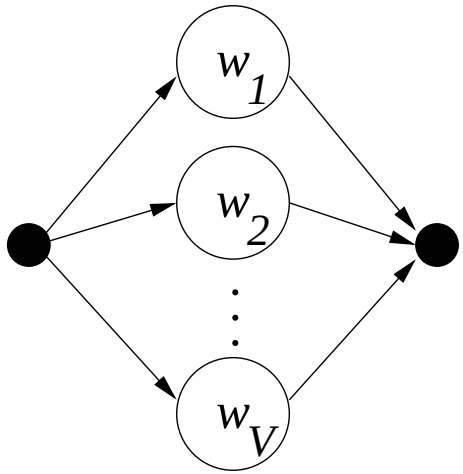
- Pattern matching
- Distance measures
- Dynamic programming
  - DTW algorithm
  - DTW with traceback
  - DTW with penalties
- Training & application



http://www.ee.surrey.ac.uk/Personal/P.Jackson/ce.adsp/

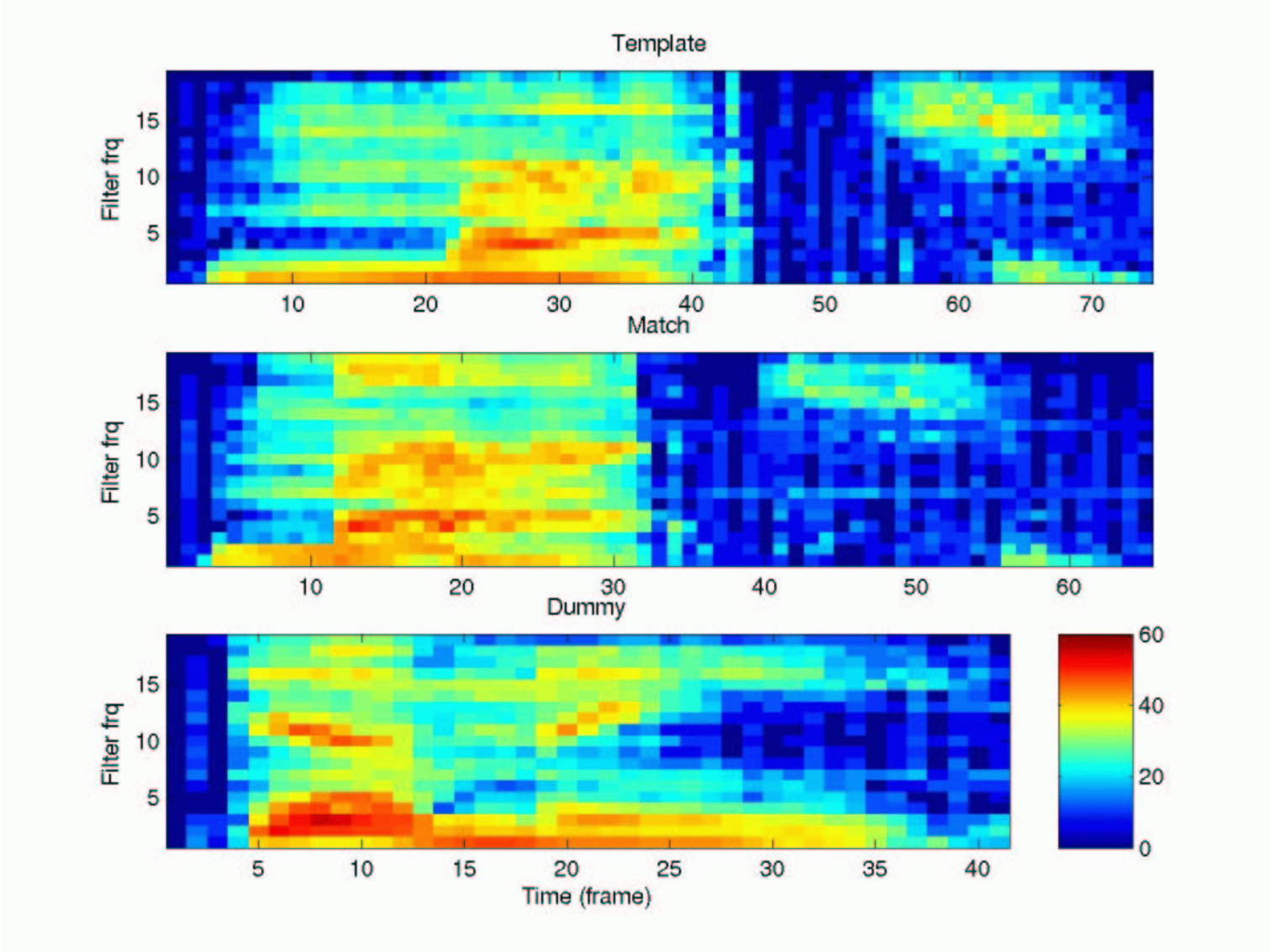
Distance-from-template principle

- Template
  - a typical example of the word or utterance to be recognized
- Distance
  - a measure of how well a new test utterance matches the reference template



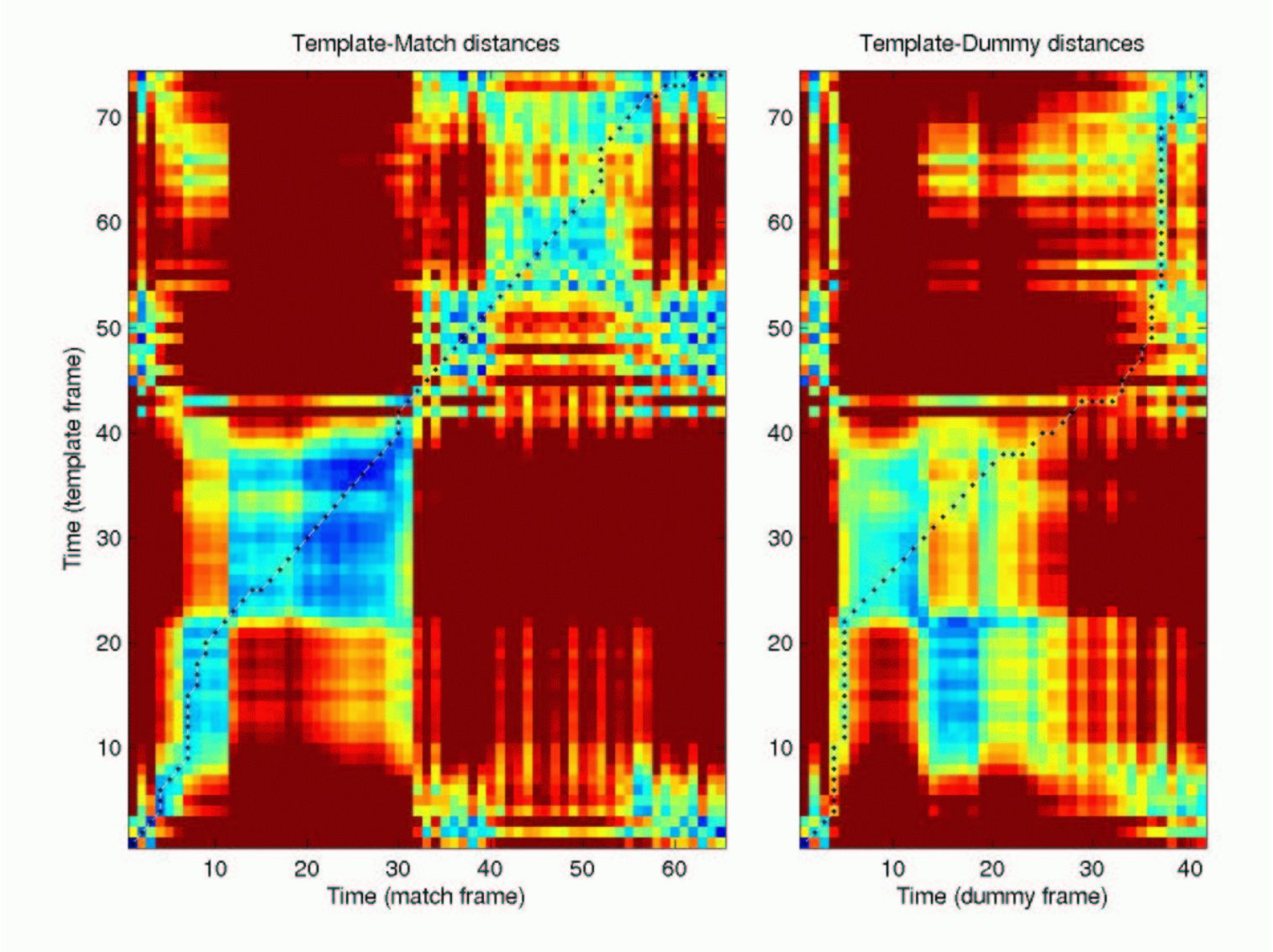
Grammar for an isolated word recognition task.

Utterance features



Template and test word features: "match", "match" & "dummy".

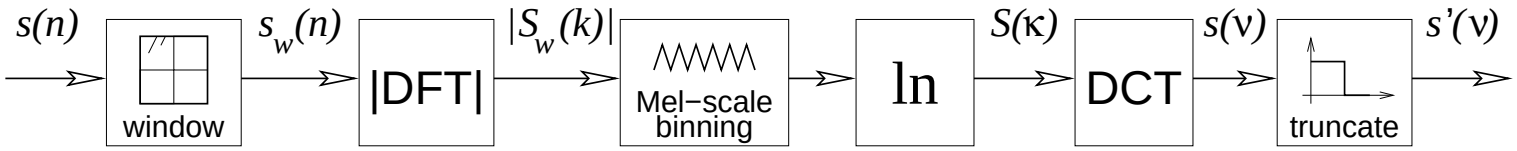
Inter-utterance distances



Computed Euclidean feature distances for template and test words.

Distance measures

- Features
  - Filterbank
  - Linear predictive coding (LPC)
  - Cepstral analysis
  - Mel-frequency cepstrum
  - Perceptual linear prediction



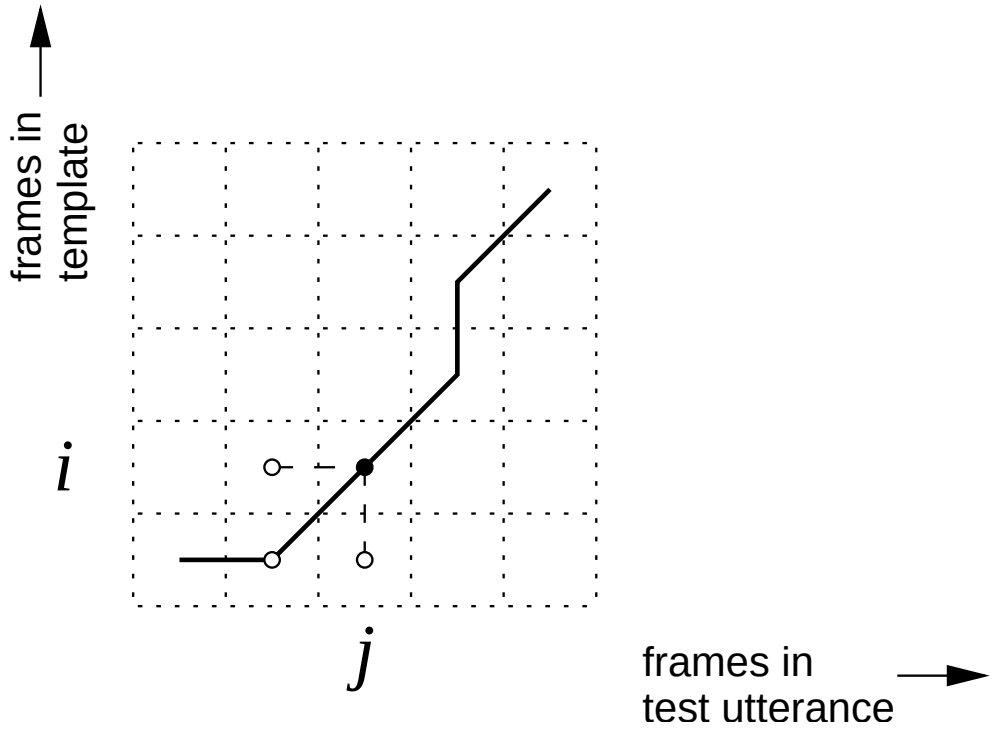
Process for computing Mel-frequency cepstral coefficients (MFCCs).

- Metrics

Distance measures

- Features
  - Filterbank
  - Linear predictive coding (LPC)
  - Cepstral analysis
  - Mel-frequency cepstrum
  - Perceptual linear prediction
- Metrics
  - Euclidean; level-differences and normalisation
  - Malhalanobis
  - Itakura

Allowing for timescale variations



Time alignment of two instances of the same word (Holmes & Holmes 2001, p.116). Open circles mark permitted predecessors to the closed circle at (i, j).

Dynamic programming for time alignment

Cummulative distance along the best path upto frame i in template and jth test frame is:

$$D(i,j) = \sum_{u,v=1,1}^{i,j}\Big|_{\text{along best path}} d(u,v), \qquad (1)$$

where $d(u,v)$ is distance between features from uth frame of template and those from vth frame of test utterance.

If we only allow transitions from current and previous states, we have

$$\begin{aligned} D(i,j) &= \min\left[D(i-1,j), D(i-1,j-1), D(i,j-1)\right] \\ &\quad + d(i,j). \end{aligned} \qquad (2)$$

Summary of DTW algorithm

Consider N-frame template and T-frame test utterance:

1. Initially,

$$D(i,1) = \begin{cases} d(i,1) & \text{for } i = 1 \\ d(i,1) + D(i-1,1) & \text{for } i = 2,\ldots,N \end{cases} \tag{3}$$

2. For $t = 2,\ldots,T$,

$$D(i,t) = \begin{cases} d(i,t) + D(i,t-1) & \text{for } i = 1 \\ d(i,t) + \min\,[\,D(i-1,t), & \\ \qquad\qquad D(i-1,t-1), & \\ \qquad\qquad D(i,t-1)\,] & \text{for } i = 2,\ldots,N \end{cases} \tag{4}$$

3. Finally,

$$\Delta = D(N, T).$$

Thus, the cost of each potential path can by computed efficiently by recursion.

DTW summary with traceback

1. Initially,

$$D(i,1) = \begin{cases} d(i,1) & \text{for } i = 1 \\ d(i,1) + D(i-1,1) & \text{for } i = 2,\ldots,N \end{cases}$$

$$\phi(i,1) = \begin{cases} [0,0] & \text{for } i = 1 \\ [i-1,1] & \text{for } i = 2,\ldots,N \end{cases}$$

2. For $t = 2,\ldots,T$,

$$D(i,t) = \begin{cases} d(i,t) + D(i,t-1) & \text{for } i = 1 \\ d(i,t) + \min\,[\,D(i-1,t), & \\ \qquad\qquad D(i-1,t-1), & \\ \qquad\qquad D(i,t-1)\,] & \text{for } i = 2,\ldots,N \end{cases}$$

$$\phi(i,t) = \begin{cases} [i,t-1] & \text{for } i = 1 \\ \arg\min\,[\,D(i-1,t), & \\ \qquad D(i-1,t-1), & \\ \qquad D(i,t-1)\,] & \text{for } i = 2,\ldots,N \end{cases}$$

3. Finally,
   $\Delta = D(N, T)$
   $z_K = [N, T]$,
   where K is the number of nodes on the optimal path.

4. Traceback for $k = K - 1, \ldots, 1$,
   $z_k = \phi(z_{k+1})$, and
   $Z = \{z_1, \ldots, z_K\}$.

Q. How could we refine the search strategy so as to encourage linear alignment, meanwhile allowing some warping?

Abbridged DTW with distortion penalty

1. Initially,

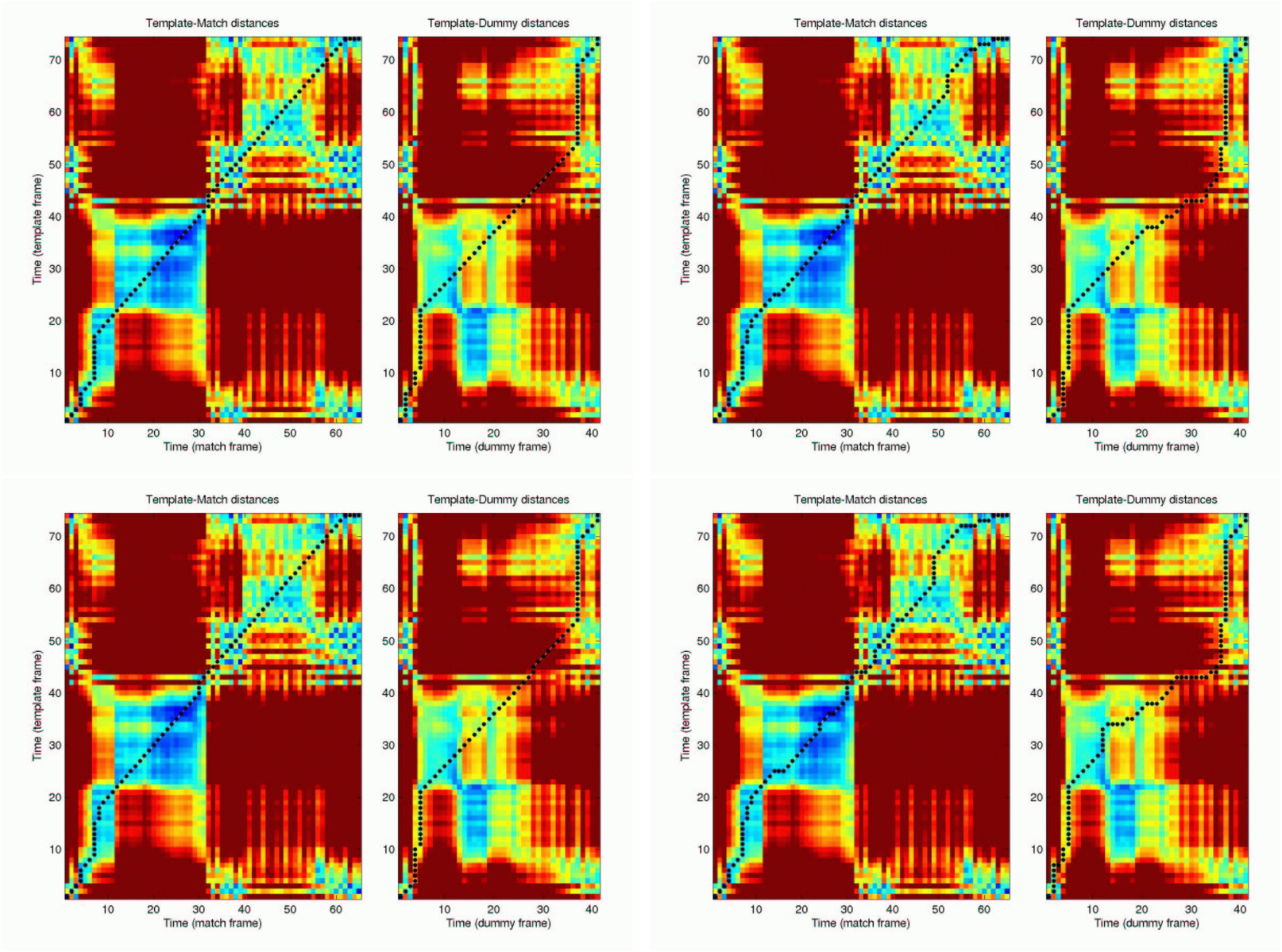
$$D(i,1) = \begin{cases} d(i,1) & \text{for } i = 1 \\ d(i,1) + D(i-1,1) + d_V & \text{for } i = 2,..,N \end{cases} \tag{5}$$

2. For $t = 2, \ldots, T$,

$$D(i,t) = \begin{cases} d(i,t) + D(i,t-1) + d_H & \text{for } i = 1 \\ \min[& \\ \quad d(i,t) + D(i-1,t) + d_V, & \\ \quad 2d(i,t) + D(i-1,t-1), & \\ \quad d(i,t) + D(i,t-1) + d_H\,] & \text{for } i = 2,..,N \end{cases} \tag{6}$$

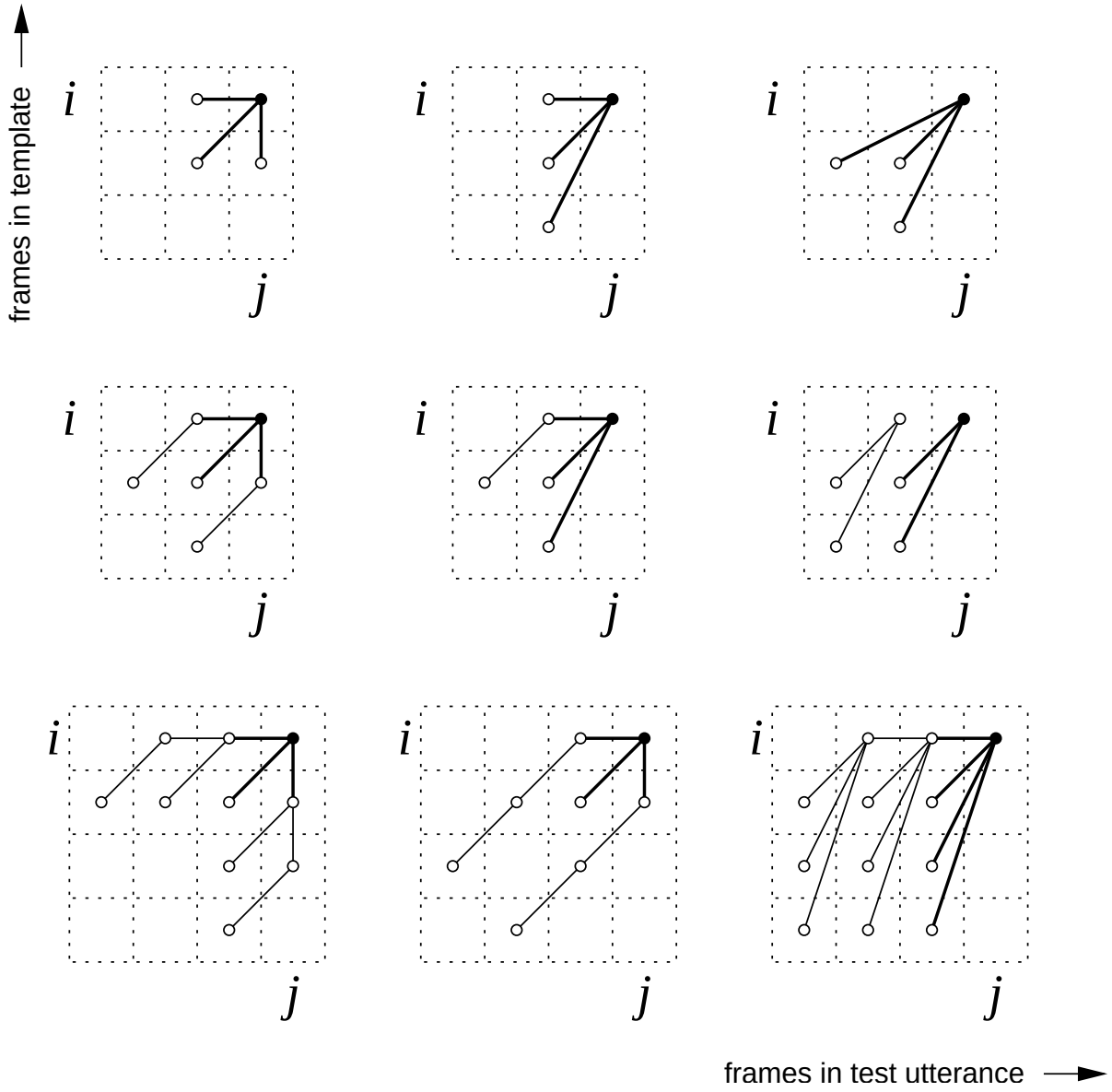
where d_V and d_H are costs associated with vertical and horizontal transitions, respectively.

Distortion penalty examples



Path with various distortion penalties (clockwise from top left): none, standard, low and high.

Alternative sets of predecessor nodes

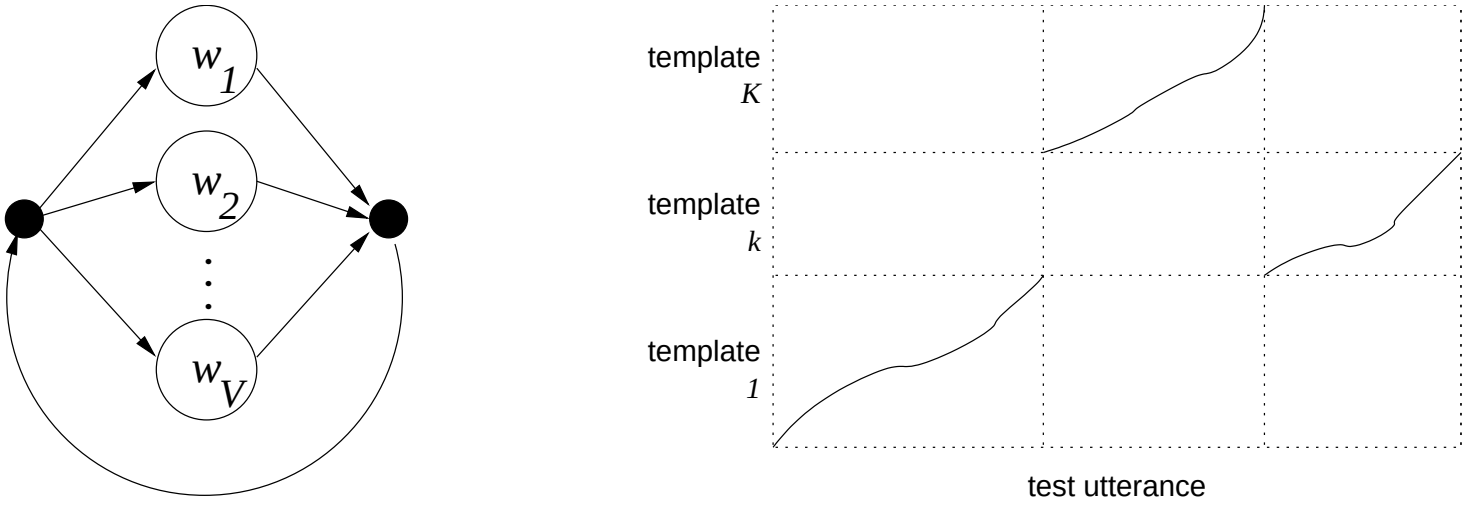


Permissible preceding nodes under various transition constraints.

Pruning and End points

- Reducing the search space
  - Gross partitioning
  - Score pruning
- End-point issues
  - Detection
  - Allowing for errors

Connected word recognition



Left: grammar for a connected word recognition task.
Right: trellis diagram showing connected templates.

Distance metric now extends across word boundaries:

$$D(i, 1, m) = d(i, 1, m) + \min_k \left[D(i-1, L(k), k) \right], \qquad (7)$$

where m is the current template, k is the previous one, and $L(k)$ is k's length.

Additional templates

- Silence template
- Wildcard template

Training a DTW recognizer

- Enrollment
  - training session with new user
  - recordings used to provide templates
- Reliable templates
  - time aligning examples
  - clustering features
  - end-point detection
  - word-boundary segmentation

Summary of Dynamic Time Warping

- Distance measures
  - features: filterbank, MFCC, PLP
  - metrics: Euclidean, Malhalanobis, Itakura
- Isolated Word Recognition
  - time alignment
  - traceback
  - distortion penalties
  - pruning
  - end points
- Connected Word Recognition
  - silence template
  - wildcard template
- Training a DTW recognizer
  - enrollment recordings
  - reliable templates