Deterministic Sequence Recognition

Notes on Speech and Audio Processing

Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan ROC

Introduction

- We have now established feature representation for local short-term spectrum. This representation is associated with linguistic categories such as phones or other sub-word units.
- In ASR, we have a sequence of features associated with an unknown class.
- Therefore, we need a framework to handle recognition based on feature sequences.

General Setting

- Suppose we have a sequence of feature vectors $X = (x_1, \ldots, x_N)$ and we wish to associate X with a second sequence $Q = (q_1, \ldots, q_N)$, where each q corresponds to a linguistic (or quasi-linguistic) unit.
- Suppose in addition that we have K reference sequences, $X_k^{\text{ref}} = (x_{k,1}^{\text{ref}}, \ldots, x_{k,N_k}^{\text{ref}}), k = 1, \ldots, K$.
- Each reference sequence X^{ref} is called a template. $q_i = j$ means x_i is aligned to x_j^{ref}. Q is chosen such that some error function is minimized.

Linear Time Warp

- The simplest case is that the template has the same length as the test sequence. Then we can use

$$D(X^{\text{ref}}, X) = \sum_{i=1}^{N} d(x_i^{\text{ref}}, x_i),$$

  where N is the length of sequence and the global distance is the sum of local distances.
- To choose the best template, simply choose the one with the least global distance.
- If the lengths are different, we can downsample or interpolate the template sequence.

Dynamic Time Warping

- Linear time warping does not properly compensate for the speaking rate: often the vowels are elongated while the consonants are roughly constant.
- It is thus desired to deal with this variety by dynamic time warping, where we allow for different warping factor at different segments of speech.
- An alignment is a correspondence between the vectors in the test sequence and the template. We want to find the alignment such that the total distortion is minimized.

Distortion for a Reference Template

- We define the distortion between the observed and reference templates to be the minimum over all alignments between the templates.
- Let the smallest cumulative distortion between $x_{1:i}$ and $x^{\text{ref}}_{1:j}$ is stored in D_{ij}. then

$$D(i,j) = d(i,j) + \min_{p(i,j)} [D(p(i,j)) + T((i,j), p(i,j))],$$

  where T is a transition cost.
- The minimum in the last column of D is the distortion

$$D = \min_j D(N, j).$$

Further Comments for DTW

- The optimal path yields the template score. The optimum template score yields the optimal answer.
- One may apply global and local constraints to reduce the search space.
- One may use clustering to reduce the number of templates, e.g. for speaker-independent systems.
- One may use probabilistic distance (or other distance measures) rather than the Euclidean distance.
- End-pointing is often mandatory for template-based system.

Connected Word Recognition

- Effects of allowing connected words
  - pronunciation may be altered due to context
  - number of hypotheses increases drastically
  - need to deal with both segmentation and recognition
- One can still use DTW for this problem!
  - The basic idea is to use a large matrix consisting of all the word templates.
  - Backtrack information can be stored by two lists per frame: $T(j)$, the lowest-cost template ending at j, and $F(j)$, the end frame of the previous template.

Segmental Approaches

- We mainly illustrate dynamic-time warping using words as templates. But the same idea can be applied to subword units as well.
- Subword units have been used in statistical systems, with essentially the same dynamic-programming search algorithm for best hypothesis.