Integrating Template Information into HMM-based ASR

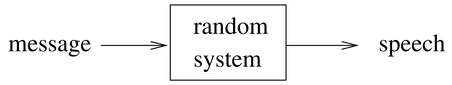
Guillermo Aradilla, Jithendra Vepa and Hervé Bourlard
IDIAP Research Institute, CH-1920 Martigny, Switzerland
{aradilla,vepa,bourlard}@idiap.ch

Summary

In this work, we present a new approach for automatic speech recognition (ASR) which tries to take advantages from the two main approaches applied to this field currently: templates and hidden Markov models (HMMs).

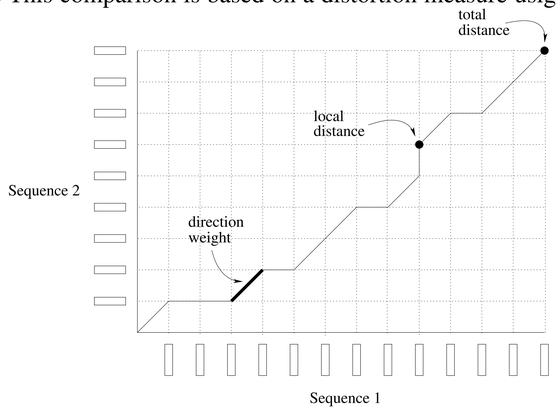
Automatic Speech Recognition

- An ASR system extracts a sequence of feature vectors from speech. This sequence can be considered as a trajectory in the feature space.
- The main task of ASR is decoding trajectories to obtain the underlying message.
- These trajectories present a large variability due to both internal and external factors such as gender, pitch, accent or interlocutor.



Template Approach

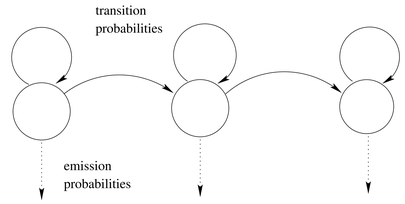
- Test utterances are compared with reference templates.
- This comparison is based on a distortion measure usig a DTW technique.



- Every template represents a particular way of pronuntiation for a specific word, hence, explicit representation of the speech variability is carried out.
- The larger the number of templates, the better accuracy the ASR system will achieve since more variability will be represented.

Stochastic Approach

- A HMM is built for every linguistic unit.



- HMMs are easily scalable. Moreover, efficient algorithms exist for training and decoding.
- All the necessary information from the training data is summarized in some model parameters.
- Assumptions about the model may not satisfy speech properties.

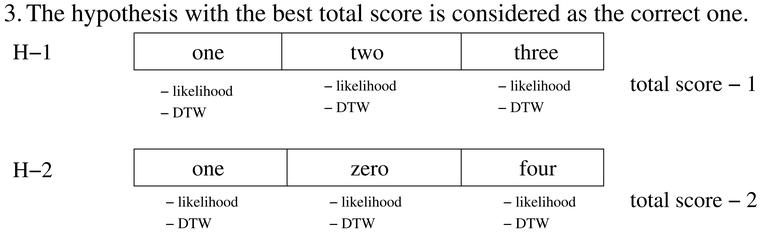
Integraging Template Information

- Template and stochastic approaches are somehow complementary: HMMs generalizes speech trajectories whereas templates carry out an explicit representation of them.
- On the other hand, they have some similarities since both of them use time warping for dealing with temporal distortion. Also, templates can be considered as a kind of simple HMM.

Experiment Description

The dataset used for this experiment is OGI Digits. As a feature vector, we have used static and delta features (26 dimensions).

1. N-best hypothesis are obtained by a conventional HMM/GMM system.
2. DTW-based measure is computed for each word of each hypothesis. The K best measures are chosen and an average measure is obtained for each word. This measure can be combined with the likelihood obtained by the HMM system.
3. The hypothesis with the best total score is considered as the correct one.



- Using this approach, we take profit from HMMs by obtaining the best hypothesis efficienty.
- Template matching computation is reduced considerably since only those templates representing the hypothesized word are considered.

Results

- Influence of the number of templates on the accuracy using only DTW-based measure:

Experiments	Validation	Test
Baseline	4.0%	4.3%
Max. Accuracy	2.5%	2.6%
1000 templates	3.9%	4.1%
6000 templates	3.7%	3.8%

- Combination of the DTW-based measure and HMM-based likelihood:

Experiments	Validation	Test
1000 templates - Combination	3.6%	3.9%
6000 templates - Combination	3.5%	3.6%

- High correlation factor between DTW-based measure and HMM-based likelihood ($\rho = 0.96$).
- Pitch information has been used for clustering the templates. Results have not changed significantly but computation time has been halved.

Conclusions & Future Work

- Experiments have showed that template matching using HMM-based hypothesis improves the system accuracy. This is due to the better representation of speech trajectories with templates.
- Combination between likelihood and DTW-measure improves the results. HMMs and templates carry some complementary information.
- A mathematical framework should be established for a better understanding.
- More meta-information can be used for improving accuracy.