INFONET Seminar Application Group 05/25/2013

Exemplar-Based Processing for Speech Recognition

Tara N. Bhuvana Ramabhadra.
IEEE SIGNAL PROCESSING MAGAZINE

Presenter Pavel Ni

Gwangju Institute of Science and Technology

Introduction

Automatic Speech Recognition is the translation of spoken words in to text.
(Voice dial, Apple Siri, Google One Voice, Samsung S voice)

Recognition and classification of speech requires modeling of speech production and uncertainty in it. Vocal tract complexity noise corruption, and vocal tract variations amongst different people arise uncertainty.

The goal of modeling is to establish a generalization from the set of observed data such that accurate classification can be made about unseen data i.e. speaker independent speech recognition.

Introduction

Construction of the model leads to two categories of approaches for modeling the observed data:

- Global-data model uses all available training data to build a model before the test sample seen
  - allow for a generalization of the observed data only if distribution estimated by the model provides a reasonable description of unseen data.
  - with limited training data unable of representing the fine detail in distribution of the data. Therefore require large amount of training data.

- exemplar based modeling since the model is build from a few relevant training examples for each test sample.
  - Building an instance of the model based only on the relevant and informative exemplars. Doesn't need large data however for each query builds local model.

Speech recognition problem

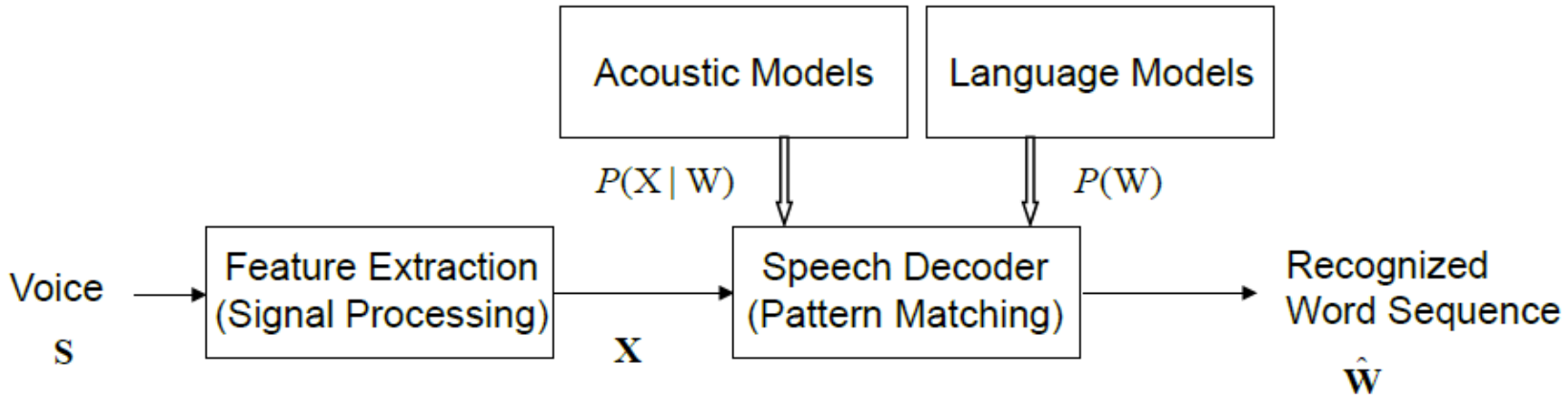

Figure 1. Block diagram of speech recognition system

Where X is a set of observations, W sequence of words.

Acoustic model: physics of sound speech, models of vocal tract
Language model: syntax and semantics

$$\hat{W} = \arg\max_{W} P(W \mid X) = \arg\max_{W} \frac{P(X \mid W)P(W)}{P(X)}$$

In speech recognition we need to find sequence of words $\hat{W}$. The common solu tion is to find the word sequence that maximizes the posterior probability $P(W \mid X)$

System overview

Features - such as power, pitch, and vocal tract configuration fr om the speech signal.
Frames – fixed length features

3 stages of recognition:
- Exemplar selection
- Instance modeling
- Frame decoding

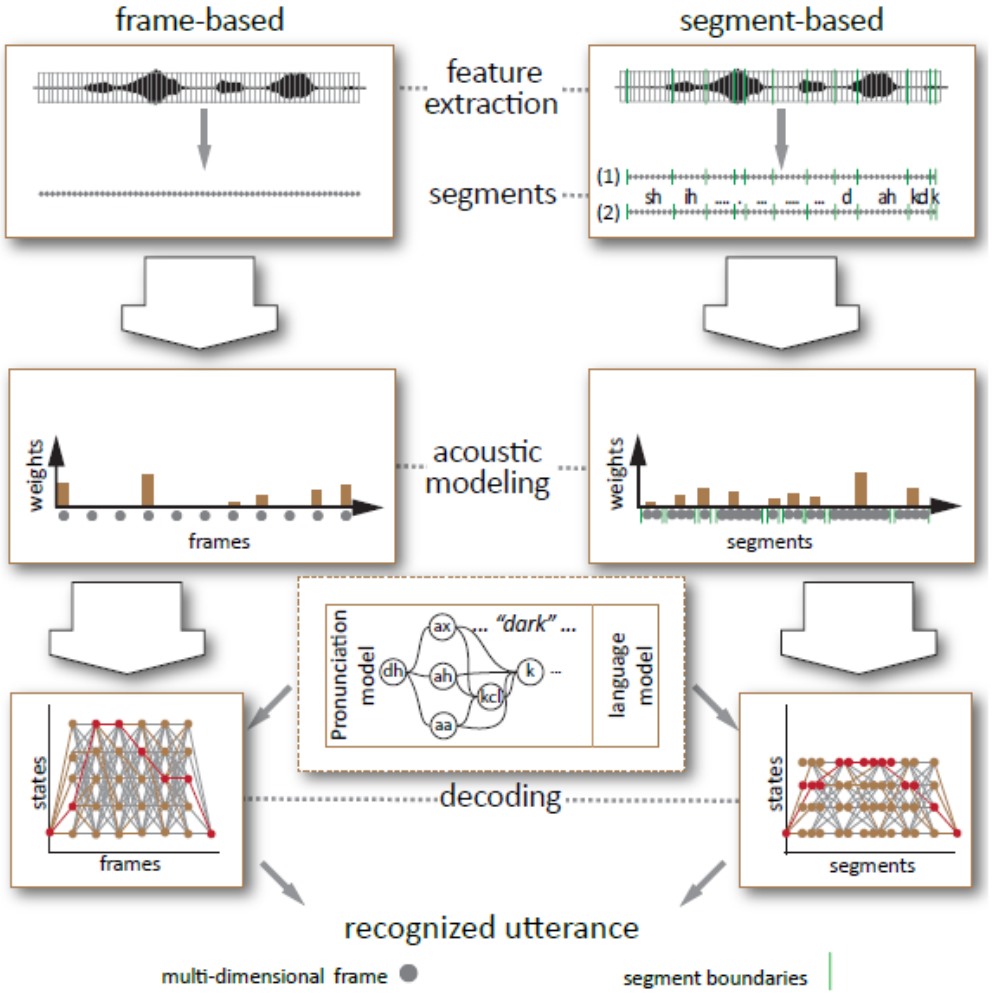

Figure 2. Frame and segment based recognition

System overview

- Exemplar selection

Identifies instances from the training data which are more relevant to test inst ances. Training set selectively downsampled to simplify the search and then appropriate exemplars are taken from this reduced set.

- Instance modeling

Set of training instances which most relevant for test instance is used to mode l the test instance. Weight distribution using k-NN (near neighbors) or sparse r epresentation (SR or compressive sensing)

- Frame decoding

Recognizing an entire sentence.

k-Nearest Neighbors

k-NN algorithm classifies test point based on *k* closest neighbors in the training set.
Exemplar selection: exemplars for individual frames, fixed length sequences of multiple frames

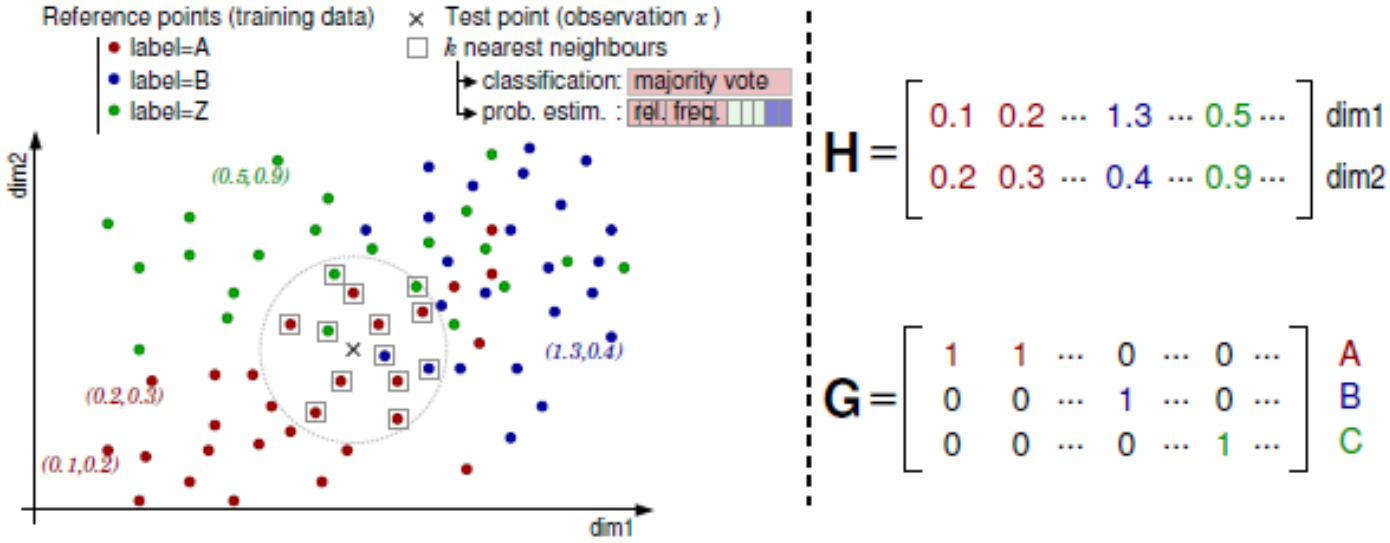

Figure 3. An example of k-NN in a 2-D feature space. In the left panel the distance between an observed feature vector *x* and a set of exemplars are shown, and in the right panel the mathematical description of the exemplars and the exemplar-label association, H and G respectively, is visualized. The columns of H and G correspond to the exemplars in the left panel.

k-Nearest Neighbors

In order to extract useful information for speech recognition or classification exemplars are associated with label. (phone classes and word labels)
Matrix G is a binary matrix that associates each exemplar in H with class labels.

Classification made by maximum likelihood.

Sparse Representation classification

Test vector y, dictionary H with exemplars h_i from training set.
Where H is a $m \times N$ matrix where m is the dimension of each feature vector x and N is the total number of all training examples from all class es. (m<<N)

Then vector y can be written as linear combination of all training examples.

$$y = H\beta$$

β should be sparse with non-zero elements for the elements in H which belong to the same class.

Results: Sparse representation reconstruction features was tested on TIMIT database. Phonetic Error Rate 18.6%. 0.8% improved over a state-of-art GMM/HMM

Thank you