Doctoral Course
in
Speech and Speaker Recognition

Part 1

Teachers: Mats Blomberg, Kjell Elenius

March - June 2007

Introduction

- Course objective
  - deeper insight into basic and specific methods and algorithms
  - *understanding* - not exact details of equations
  - no derivation of theorems and algorithms
  - Not covered
    - Phonetics, linguistics
  - Signal processing relevant parts (short time spectral analysis)
  - theory of probabilistics and pattern recognition overviewed
  - merit 7.5p in GSLT (5p in old system))
- Recommended background
  - GSLT or TMH course in “Speech technology” or equivalent

Recommended Background

- Basic mathematics, statistics and programming
- Acoustic phonetics
- Speech analysis
  - Short Time Spectral Analysis
  - MFCC
- Recognition
  - Dynamic programming and DTW
  - Fundamentals of hidden Markov models
  - Viterbi decoding
  - Phoneme-based speech recognition methods

Literature

- Spoken Language Processing
  - A Guide to Theory, Algorithm and System Development
  - X. Huang, A. Acero and H-W Hon
  - Contains theoretically heavy parts and many equations but it is not necessary to follow all derivations. The verbose explanations of their functions are easier to follow.
- Separate papers
  - Speaker recognition
  - Finite State Transducers
  - Bayesian Networks
  - Articulatory inspired approaches
  - ...

Course organization

- 3 lecture days
  - March 29-30, May 11
- Practical and computational exercises
- Write term paper + review + presentation
- Closing seminar day
  - June 8
  - Students’ presentation of individual term papers

Course overview

- Day #1 March 29
  - Probability, Statistics and Information Theory (pp 73-131: 59 pages)
  - Pattern Recognition (pp 133-197: 65 pages)
  - Speech Signal Representations (pp 275-336 62 pages)
  - Hidden Markov Models (pp 377-413: 37 pages)
  - HTK tutorial & practical exercise
- Day #2 March 30
  - Acoustic Modeling (pp 415-475: 61 pages)
  - Environmental Robustness (pp 477-544: 68 pages)
  - Computational problems exercise
- Day #3 May 11
  - Language Modeling (pp 545-590: 46 pages)
  - Basic and Large-Vocabulary Search Algorithms (pp 591-685: 94 pages)
  - Applications and User Interfaces (pp 919-956: 38 pages)
  - Speaker recognition
- Day #4 June 8
  - Presentations of term papers & Solutions to exercises

Term paper

- Choose subject from a list or suggest one yourself
- Review each others reports
- Suggested topics
  - Further experiments on the practical exercise corpus
  - Phoneme recognition experiments on larger corpus (e.g. TIMIT or WAXHOLM)
  - Language models for speech recognition
  - Limitations in standard HMM and ways to reduce them
  - Pronunciation variation and their importance for speech recognition
  - New search methods
  - Techniques for robust recognition of speech
  - Speaker recognition topics: impersonation, forensics, channel and score normalisation
  - Own work and experiments after discussion with the teacher

Course Book

- The authors work for Microsoft Research
- Topics
  - Fundamental theory
    - Speech & Language, Statistics, Pattern Recognition, Information Theory
  - Speech processing
  - Speech recognition
  - Text-to-Speech
  - Spoken Language systems
- Historical Perspective and Further Reading in each chapter
- Important algorithms described in step-by-step
- Examples from Microsoft's own research

Book organization 1(2)

- Ch 1 Introduction

- Part I: Fundamental theory
  - Ch 2 Spoken Language Structure
  - Ch 3 Probability, Statistics and Information Theory
  - Ch 4 Pattern Recognition
- Part II: Speech Processing
  - Ch 5 Digital Signal Processing
  - Ch 6 Speech Signal Representation
  - Ch 7 Speech Coding

Book organization 2(2)

- Part III: Speech Recognition
  - Ch 8 Hidden Markov Models
  - Ch 9 Acoustic Modeling
  - Ch 10 Environmental Robustness
  - Ch 11 Language Modeling
  - Ch 12 Basic Search Algorithms
  - 13 Large-Vocabulary Search Algorithms
- Part IV: Text-to-Speech Systems
  - Ch 14 Text and Phonetic Analyses
  - Ch 15 Prosody
  - Ch 16 Speech Synthesis
- Part V: Spoken Language systems
  - Ch 17 Spoken Language Understanding
  - Ch 18 Applications and User Interfaces

Ch 3. Probability, Statistics and Information Theory

- Conditional Probability and Bayes' Rule
- Covariance and Correlation
- Gaussian Distributions
- Bayesian Estimation and MAP Estimation
- Entropy
- Conditional Entropy
- Mutual Information and Channel Coding

Conditional Probability and Bayes' Rule

- Bayes' rule - the common basis for all pattern recognition

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{N_{AB}/N_S}{N_B/N_S}$$

$$P(AB) = P(A\,|\,B)P(B) = P(B\,|\,A)P(A)$$

$$P(A|B) = \frac{P(B\,|\,A)P(A)}{P(B)}$$

$$P(A_i|B) = \frac{P(A_iB)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{n} P(B|A_k)P(A_k)}$$

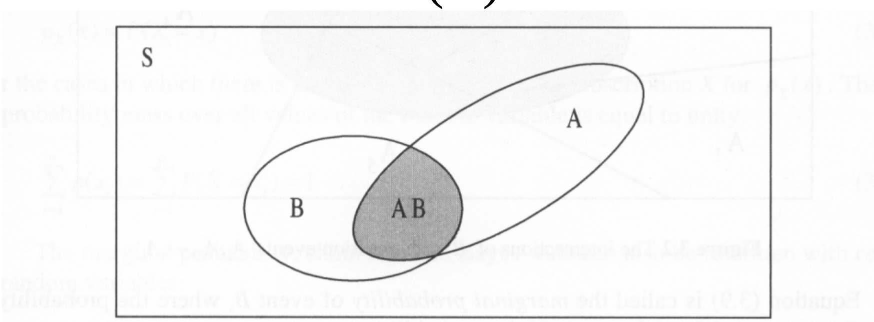

Figure 3.1 The intersection AB represents where the joint event A and B occurs concurrently.

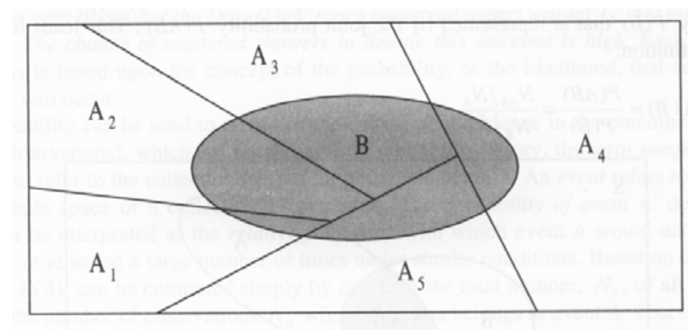

Figure 3.2 The intersections of B with partition events $A_1, A_2, \cdots, A_n$.

Bayes' rule in ASR

$$P(Word \mid Acoustics) = \frac{P(Acoustics \mid Word) \times P(Word)}{P(Acoustics)}$$

P(Word | Acoustics) is the *a posteriori probability* for a word sequence given the acoustic information

P(Acoustics | Word) is the *probability* that the word sequence generates the acoustic information and is calculated from the training data

P(Word) is given by the language model and is the *a priori probability* for the word sequence

P(Acoustics) may be seen as *constant* since it is independent of the word sequence and may be ignored

A combination of acoustic and language knowledge!

Mean, Covariance and Correlation

- Mean $$\mu_x = E(X) = \sum_x x f(x)$$
- Variance $$Var(X) = \sigma_x^2 = E[(X-\mu_x)^2] = \frac{\sum (x_i - \mu_x)^2}{n-1}$$
- Covariance $$Cov(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right]$$
- Correlation $$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$
- Multidimensional (Mean and variance vectors, covariance matrix)

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{bmatrix} \qquad \mathbf{\Sigma_X} = Cov(\mathbf{X}) = \begin{bmatrix} Cov(X_1, X_1) & \cdots & Cov(X_1, X_n) \\ \vdots & & \vdots \\ Cov(X_n, X_1) & \cdots & Cov(X_n, X_n) \end{bmatrix}$$

Gaussian Distributions

- One-dimensional

$$f(x \mid \mu, \sigma^2) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right]$$

- Multivariate n-dimensional

$$f(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = N(x; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

3.2 Estimation theory

- The basis for training a speech recogniser
- Estimate parameters of a probability distribution function
  - Minimum/Least Mean Squared Error Estimation
    - Minimize the difference between the distribution of the data and the model
  - Maximum Likelihood Estimation
    - Find the distribution with the maximum likelihood of generating the data
  - Bayesian Estimation and MAP Estimation
    - Assumes that we have a prior distribution that is modified by the new data

Minimum Mean / Least Squared Error Estimation

- Modify a model of the distribution to approximate the data with minimum error
- Find a function that predicts the value of Y from having observed X
- Estimation is made on joint observations of X and Y
- Minimize: $E(Y-\hat{Y})^2 = E(Y-g(X))^2$
- Minimum Mean Squared Error (MMSE) when the joint distribution is known
- Least Squared Error (LSE) when the distribution is unknown, only observation pairs (Ex. curve fitting)
- MMSE and LSE becomes equivalent with infinite number of samples

Maximum Likelihood Estimation (MLE)

- The most widely used parametric estimation method
- Find the distribution that maximizes the likelihood of generating the observed data

$$\mathbf{\Phi}_{MLE} = \arg\max_{\Phi} p(\mathbf{x} \mid \mathbf{\Phi})$$

- Corresponds to intuition
  - Max likelihood is normally achieved when the model has the same distribution as the observed data
- Example: univariate Gaussian pdf

$$\mu_{MLE} = \frac{1}{n}\sum_{k=1}^{n} x_n = E(x) \qquad \sigma^2_{MLE} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \mu_{MLE})^2 = E\left[(x_k - \mu_{MLE})^2\right]$$

Bayesian and MAP Estimation

- Assumes that we have a prior distribution that is modified by the new data
- Use Bayes' rule to find the new posterior distribution $\mathbf{\Phi}$

$$\mathbf{\Phi}_{MAP} = \arg\max_{\Phi} p(\mathbf{x} \mid \mathbf{\Phi})\, p(\mathbf{\Phi})$$

- Univariate Gaussian Mean: $\rho = \frac{\sigma^2 \mu + n\nu^2 \bar{x}_n}{\sigma^2 + n\nu^2}$ Var: $\tau^2 = \frac{\sigma^2 \nu^2}{\sigma^2 + n\nu^2}$
- MAP: Maximum A Posteriori probability is a Bayesian Estimator
- MAP becomes MLE with uniform prior distribution or infinite number of training data
- Valuable for limited training data and for adaptation

3.3 Significance testing

- For practical methods, see Chapter 4
- How certain are the achieved results?
  - The true result is within an interval around the measured value with a certain probability
  - Confidence level and interval
  - Rule of thumb in speech recognition (Doddington, 198x)
    - To assure that the true error rate is within the measured value ± 30% with a probability of 0.9, requires at least 30 errors to have been made
- Is algorithm A better than B?
  - Matched-Pairs Test
    - Compare results on the same test data,
    - Sign Test
    - Magnitude difference Test
    - McNemar Test (Ch 4)

Entropy and Perplexity

- The information in seeing event x_i with probability $P(x_i)$ is defined as:

$$I(x_i) = {}^2\log \frac{1}{P(x_i)}$$

- Entropy is the average information over all possible x values:

$$H(X) = E[I(X)] = \sum_S P(x_i) I(x_i) = \sum_S P(x_i)\, {}^2\log \frac{1}{P(x_i)} = -\sum_S P(x_i)\, {}^2\log(P(x_i))$$

- Perplexity $PP(X) = 2^{H(X)}$
  - The equivalent size of an imaginary list with equi-probable words
  - Perplexity for English letters: 2.39, English words: 130
- Conditional Entropy
  - Input X is distorted by a noisy channel into output Y
  - What is the uncertainty of X after observing Y?
  - Example: Confusion matrix

$$H(X|Y) = -\sum_X \sum_Y P(X = x_i, Y = y_j)\, {}^2\log P(X = x_i | Y = y_j)$$

  - If only diagonal values, the conditional entropy is 0

3.4.4 Mutual Information and Channel Coding

- Mutual Information I(X;Y)
  - How much does Y tell us about X?
  - The difference between the entropy of X and the conditional entropy of X given Y

$$I(X;Y) = H(X) - H(X \mid Y) = \ldots = E\left[\log \frac{P(X,Y)}{P(X)P(Y)}\right]$$

- $H(X|Y)$
  - "the amount of uncertainty remaining about *X* after *Y* is known”
  - represents the noise in the channel
- If X and Y are independent: I = 0