5.0 Acoustic Modeling

References: 1. 2.2, 3.4.1, 4.5, 9.1~ 9.4 of Huang

2. " Predicting Unseen Triphones with Senones",

IEEE Trans. on Speech & Audio Processing, Nov 1996

Unit Selection for HMMs

- **Possible Candidates**
  - phrases, words, syllables, phonemes.....
- **Phoneme**
  - the minimum units of speech sound in a language which can serve to distinguish one word from the other

    e.g. bat / pat , bad / bed
  - phone : a phoneme's acoustic realization

    the same phoneme may have many different realizations

    e.g. sat / meter
- **Coarticulation and Context Dependency**
  - context: right/left neighboring units
  - coarticulation: sound production changed because of the neighboring units
  - right-context-dependent (RCD)/left-context-dependent (LCD)/ both
  - intraword/interword context dependency
- **For Mandarin Chinese**
  - character/syllable mapping relation
  - syllable: Initial (聲母) / Final (韻母) / tone (聲調)

tea	it	ㄅㄢ
two	at	ㄅㄨ
target		ㄅㄧ

Unit Selection Principles

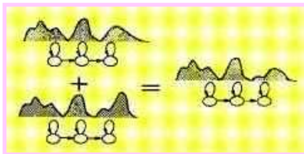
- **Primary Considerations**
  - accuracy: accurately representing the acoustic realizations
  - trainability: feasible to obtain enough data to estimate the model parameters
  - generalizability: any new word can be derived from a predefined unit inventory
- **Examples**
  - words: accurate if enough data available, trainable for small vocabulary, NOT generalizable
  - phoneme : trainable, generalizable

    difficult to be accurate due to context dependency
  - syllable: 50 in Japanese, 1300 in Mandarin Chinese, over 30000 in English
- **Triphone**
  - a phoneme model taking into consideration both left and right neighboring phonemes

    $(60)^3 \rightarrow 216{,}000$
  - very good generalizability, balance between accuracy/ trainability by parameter-sharing techniques

Sharing of Parameters and Training Data for Triphones

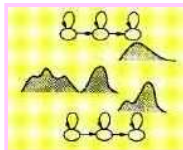
- **Sharing at Model Level**



Generalized Triphone

- clustering similar triphones and merging them together

- **Sharing at State Level**



Shared Distribution Model (SDM)

- those states with quite different distributions do not have to be merged

Some Fundamentals in Information Theory

- **Quantity of Information Carried by an Event (or a Random Variable)**
  - Assume an information source: output a random variable m_j at time j

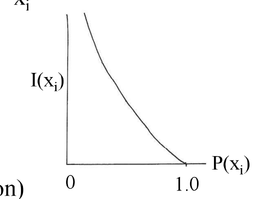
S → $U = m_1 m_2 m_3 m_4 \ldots m_j$: the j^{th} event, a random variable

$m_j \in \{x_1, x_2, \ldots x_M\}$, M different possible kinds of outcomes

$P(x_i)$= Prob $[m_j = x_i]$, $\sum_{i=1}^{M} P(x_i) = 1$, $P(x_i) \geq 0$, i= 1,2,.....M

  - Define $I(x_i)$= quantity of information carried by the event $m_j = x_i$

    Desired properties:
    1. $I(x_i) \geq 0$
    2. $\lim_{P(x_i) \to 1} I(x_i) = 0$
    3. $I(x_i) > I(x_j)$, if $P(x_i) < P(x_j)$
    4. Information quantities are additive



  - $I(x_i) = \log\left[\frac{1}{p(x_i)}\right] = -\log[P(x_i)] = -\log_2[P(x_i)]$ bits (of information)
  - H(S) = entropy of the source = average quantity of information out of the source each time

$$= \sum_{i=1}^{M} P(x_i)\, I(x_i) = -\sum_{i=1}^{M} P(x_i) \{\log[P(x_i)]\} = E\,[I(x_i)]$$

= the average quantity of information carried by each random variable

Fundamentals in Information Theory

M=2, $\{x_1, x_2\} = \{0, 1\}$

S → U = 1 1 0 1 0 0 1 0 1 0 1 1 0 0 1 … …

P (0) = P (1) = ½

U = 1 1 1 1 1 1 1 1 1 … …

P (1) = 1 , P (0) = 0

U = 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 … …

P (1) ≈ 1 , P (0) ≈ 0

M=4, $\{x_1, x_2, x_3, x_4\} = \{00, 01, 10, 11\}$

S → U = 0 1 0 0 1 0 1 1 0 1 … …

Some Fundamentals in Information Theory

- **Examples**
  - M = 2, $\{x_1, x_2\} = \{0,1\}$, P(0)= P(1)= $\frac{1}{2}$

    I(0) = I(1) = 1 bit (of information), H(S)= 1 bit (of information)

    U = 0 1 1 0 1 1 0 1 0 0 1 0 1 0 1 1 0 … …

    ↑

    This binary digit carries exactly 1 bit of information
  - M =4, $\{x_1, x_2, x_3, x_4\}=\{00, 01, 10, 11\}$, $P(x_1)= P(x_2)= P(x_3)= P(x_4)= \frac{1}{4}$

    $I(x_1)= I(x_2)= I(x_3)= I(x_4)=$ 2 bits (of information),

    H(S)= 2 bits (of information)

    U = 0 1 0 0 0 1 1 1 1 0 1 0 1 1 … …

    ↑

    This symbol (represented by two binary digits) carries exactly 2 bits of information
  - M = 2, $\{x_1, x_2\} = \{0,1\}$, P(0)= $\frac{1}{4}$, P(1)= $\frac{3}{4}$

    I(0)= 2 bits (of information), I(1)= 0.42 bits (of information)

    H(S)= 0.81 bits (of information)

    U = 1 1 1 0 1 1 1 1 1 0 0 1 1 1 1 1 0 … …

    ↑ ↑

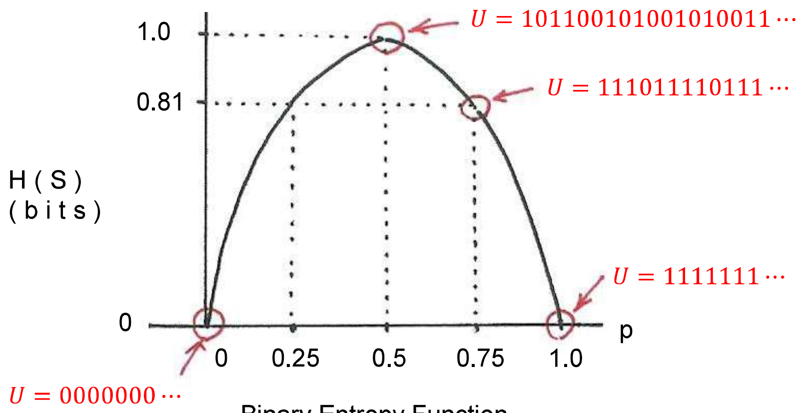
    This binary digit carries 0.42 bit of information

    This binary digit carries 2 bits of information

Fundamentals in Information Theory

M=2, $\{x_1, x_2\} = \{0, 1\}$, P (1) = p , P (0) = 1 - p

H (S) = - [p log p + (1-p) log (1-p)]



Binary Entropy Function

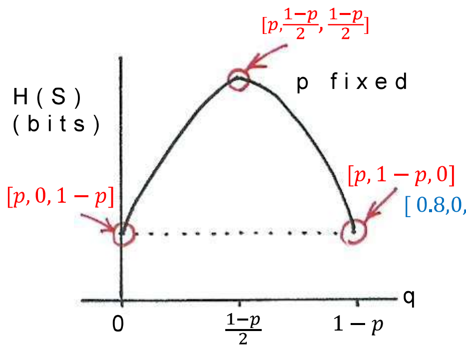
Fundamentals in Information Theory

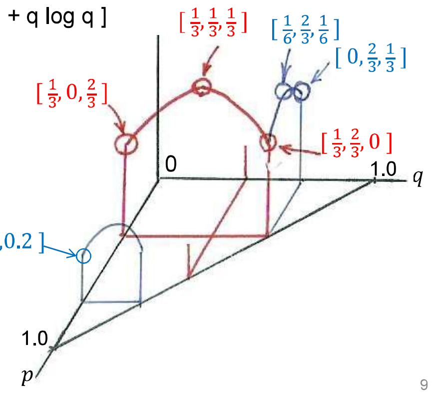
M=3, $\{x_1, x_2, x_3\} = \{0, 1, 2\}$

P(0) = p, P(1) = q, P(2) = 1-p-q

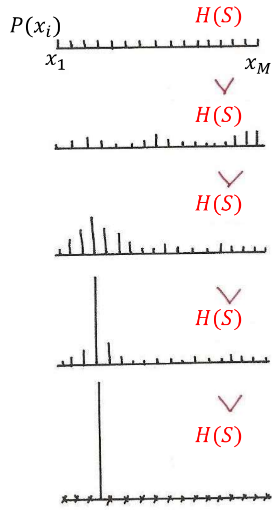
[p, q, 1-p-q]

H (S) = - [p log p + (1-p-q) log (1-p-q) + q log q]





Fundamentals in Information Theory



所帶Information 量最大

亂度最大，最random

不確定性最大

It can be shown

$0 \le H(S) \le \log M$, M: number of different symbols

equality when
$P(x_j)$= 1, some j
$P(x_k)$=0, k≠ j

equality when
$P(x_i)=\frac{1}{M}$, all i

一個 distribution
集中或分散的程度

H(S)：Entropy

確定性最大，最不random

純度最高

Some Fundamentals in Information Theory

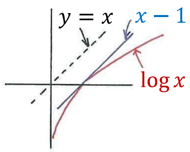
- **Jensen's Inequality**

  $$-\sum_{i=1}^{M} p(x_i)\log[p(x_i)] \le -\sum_{i=1}^{M} p(x_i)\log[q(x_i)]$$

  $q(x_i)$: another probability distribution, $q(x_i) \ge 0$, $\sum_{i=1}^{M} q(x_i) = 1$

  equality when $p(x_i)= q(x_i)$, all i

  – proof: log x≤ x-1, equality when x=1

  $$\sum_i p(x_i)\log\left[\frac{q(x_i)}{p(x_i)}\right] \le \sum_i p(x_i)\left[\frac{q(x_i)}{p(x_i)} - 1\right] = 0$$

  – replacing $p(x_i)$ by $q(x_i)$, the entropy is increased
  using an incorrectly estimated distribution giving higher degree of uncertainty

- **Kullback-Leibler(KL) Distance (KL Divergence)**

  $$D[p(x)\| q(x)] = \sum_i p(x_i)\log\left[\frac{p(x_i)}{q(x_i)}\right] \ge 0$$

  – difference in quantity of information (or extra degree of uncertainty) when p(x) replaced by q(x), a measure of distance between two probability distributions, asymmetric

  – Cross-Entropy (Relative Entropy)

- **Continuous Distribution Versions**

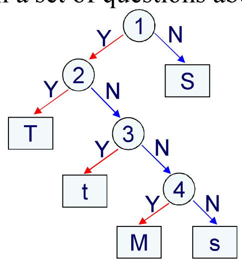
Classification and Regression Trees (CART)

- **An Efficient Approach of Representing/Predicting the Structure of A Set of Data — trained by a set of training data**
- **A Simple Example**
  – dividing a group of people into 5 height classes without knowing the heights:

  Tall(T), Medium-tall(t), Medium(M), Medium-short(s),Short(S)

  – several observable data available for each person: age, gender, occupation....(but not the height)

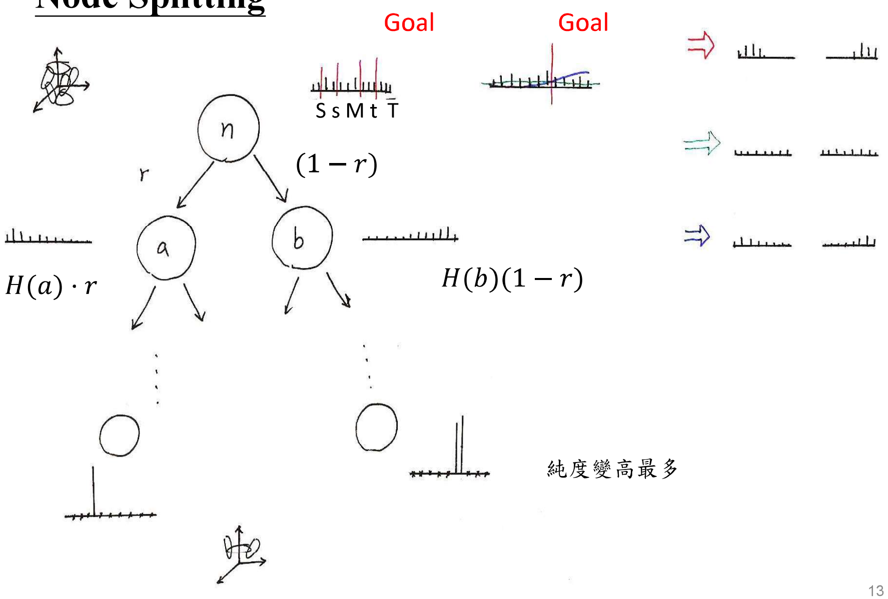
  – based on a set of questions about the available data

  1. Age > 12 ?
  2. Occupation= professional basketball player ?
  3. Milk Consumption > 5 quarts per week ?
  4. gender = male ?

  – question: how to design the tree to make it most efficient?

Node Splitting



Splitting Criteria for the Decision Tree

- **Assume a Node n is to be split into nodes a and b**
  - weighted entropy

    $\overline{H}_n = \left(-\sum_i p(c_i|n)\log\left[p(c_i|n)\right]\right)p(n)$

    $p(c_i|n)$: percentage of data samples for class i at node n

    p(n): prior probability of n, percentage of samples at node n out of total number of samples
  - entropy reduction for the split for a question q

    $\Delta\overline{H}_n(q) = \overline{H}_n - \left[\overline{H}_a + \overline{H}_b\right]$
  - choosing the best question for the split at each node

    $q^* = \underset{q}{\arg\max}\left[\Delta\overline{H}_n(q)\right]$
- **It can be shown**

  $\Delta\overline{H}_n = \overline{H}_n - (\overline{H}_a + \overline{H}_b)$

  $= D\left[a(x)\|n(x)\right]p(a) + D\left[b(x)\|n(x)\right]p(b)$

  a(x): distribution in node a, b(x) distribution in node b

  n(x): distribution in node n , $D[\bullet\|\bullet]$: KL divergence
  - weighting by number of samples also taking into considerations the reliability of the statistics
- **Entropy of the Tree T**

  $\overline{H}(T) = \sum_{\text{terminal } n} \overline{H}_n$
  - the tree-growing (splitting) process repeatedly reduces $\overline{H}(T)$

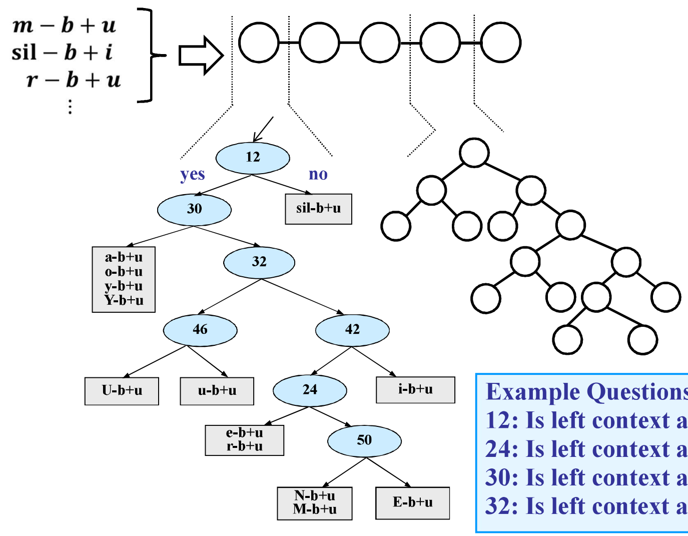
Training Triphone Models with Decision Trees

- **Construct a tree for each state of each base phoneme (including all possible context dependency)**
  - e.g. 50 phonemes, 5 states each HMM

    5*50=250 trees
- **Develop a set of questions from phonetic knowledge**
- **Grow the tree starting from the root node with all available training data**
- **Some stop criteria determine the final structure of the trees**
  - e.g. minimum entropy reduction, minimum number of samples in each leaf node
- **For any unseen triphone, traversal across the tree by answering the questions leading to the most appropriate state distribution**
- **The Gaussian mixture distribution for each state of a phoneme model for contexts with similar linguistic properties are "tied" together, sharing the same training data and parameters**
- **The classification is both data-driven and linguistic-knowledge-driven**
- **Further approaches such as tree pruning and composite questions (e.g.** $q_i\overline{q}_j + q_k$ **)**

Training Tri-phone Models with Decision Trees

- **An Example: "(_ −) b (+_)"**

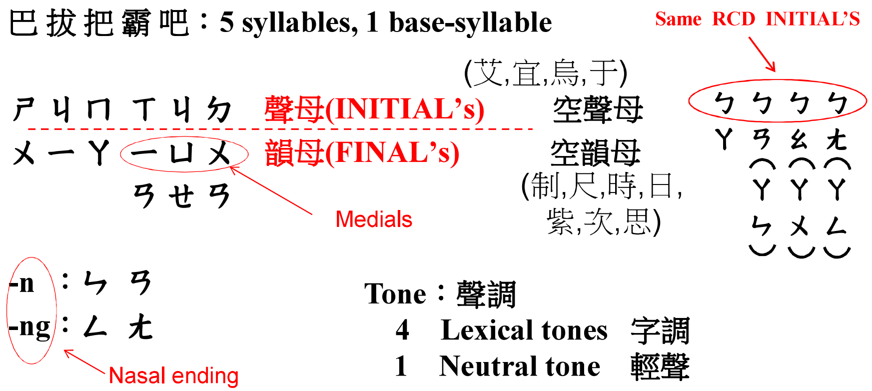


Example Questions:
12: Is left context a vowel?
24: Is left context a back-vowel?
30: Is left context a low-vowel?
32: Is left context a rounded-vowel?

Phonetic Structure of Mandarin Syllables

<table>
<tr><td colspan="4">Syllables (1,345)</td></tr>
<tr><td colspan="4">Base-syllables (408)</td><td rowspan="5">Tones (4+1)</td></tr>
<tr><td rowspan="2">INITIAL's (21)</td><td colspan="3">FINAL's (37)</td></tr>
<tr><td>Medials (3)</td><td>Nucleus (9)</td><td>Ending (2)</td></tr>
<tr><td>Consonants (21)</td><td colspan="3">Vowels plus Nasals (12)</td></tr>
<tr><td colspan="4">Phonemes (31)</td></tr>
</table>

Phonetic Structure of Mandarin Syllables

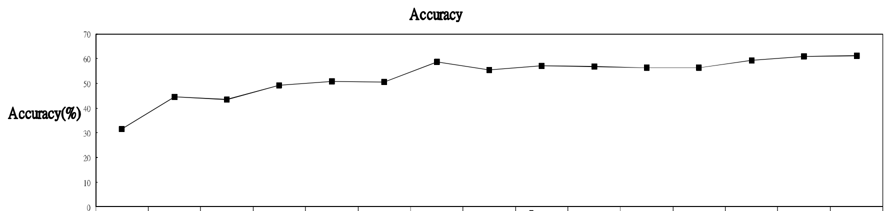


Subsyllabic Units Considering Mandarin Syllable Structures

- **Considering Phonetic Structure of Mandarin Syllables**
  - INITIAL / FINAL's
  - Phone(me)-like-units / phonemes
- **Different Degrees of Context Dependency**
  - intra-syllable only
  - intra-syllable plus inter-syllable
  - right context dependent only
  - both right and left context dependent
- **Examples :**
  - 113 right-context-dependent (RCD) INITIAL's extended from 22 INITIAL's plus 37 context independent FINAL's: 150 intrasyllable RCD INITIAL/FINAL's
  - 33 phone(me)-like-units extended to 145 intra-syllable right-context-dependent phone(me)-like-units, or 481 with both intra/inter-syllable context dependency
  - At least 4,600 triphones with intra/inter-syllable context dependency

Comparison of Acoustic Models Based on Different Sets of Units

- **Typical Example Results**



CI Phone	CI IF	LCD Phone	RCD Phone	RCD IF	Demi-Phone	Inter-RCD Phone	Inter-RCD IF	Inter-Demi-Phone	Triphone. backoff	Triphone. ml.cb	Triphone. ml.rb	Triphone. sl.cb	Triphone. sl.rb	Triphone. sl.rb.nq
31.47	44.56	43.43	49.19	50.74	50.39	58.56	55.46	57.04	56.8	56.1	56.21	59.34	60.77	61.22

- **INITIAL/FIANL (IF) better than phone for small training set**
- **Context Dependent (CD) better than Context Independent (CI)**
- **Right CD (RCD) better than Left CD (LCD)**
- **Inter-syllable Modeling is Better**
- **Triphone is better**
- **Approaches in Training Triphone Models are Important**
- **Quinphone (2 context units on both sides considered) are even better**