



TRƯỜNG ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Công nghệ Nhận dạng và Tổng hợp Tiếng nói

TRỊNH VĂN LOAN, ĐẠI HỌC BÁCH KHOA HÀ NỘI

ONE LOVE. ONE FUTURE.

1



TRƯỜNG ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

NHẬN DẠNG TIẾNG NÓI

ONE LOVE. ONE FUTURE.

2

Hidden Markov Model (HMM): A Brief Overview

History


- A Markov Model is a set of mathematical procedures developed by Russian mathematician Andrei Andreyevich Markov (1856-1922) who originally analyzed the alternation of vowels and consonants due to his passion for poetry.
- In the late 1960s and early 1970s Leonard E. Baum and his colleagues studied, developed and extended the Markov techniques by creating new models such as the Hidden Markov Model (HMM)
- Introduced to speech processing by Baker (CMU) and Jelinek (IBM) in the 1970s (discrete HMMs)
- Then extended to continuous HMMs by Bell Labs

Assumptions

- Speech signal can be characterized as a parametric random (stochastic) process
- Parameters can be estimated in a precise, well-defined manner

Three fundamental problems

- Evaluation of probability (likelihood) of a sequence of observations given a specific HMM
- Determination of a best sequence of model states
- Adjustment of model parameters so as to best account for observed signals (or discrimination purposes)

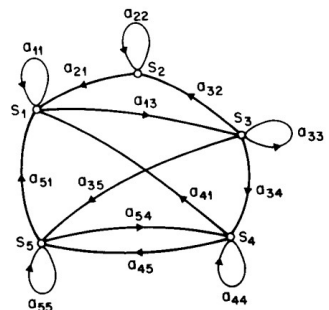



TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

3

Các quá trình Markov rời rạc về thời gian

- Xét một hệ thống có thể được biểu diễn tại bất kì thời điểm nào bởi một trạng thái trong tập hợp N trạng thái phân biệt, được đánh số $\{1, 2, \dots, N\}$. ($N=5$)
- Tại những thời điểm rời rạc cách đều, hệ thống thực hiện sự thay đổi trạng thái (có thể giữ nguyên cùng trạng thái) tùy theo một tập hợp các xác suất tương ứng với trạng thái.
- Các thời điểm thay đổi trạng thái $t = 1, 2, \dots$
- Trạng thái tại thời điểm t là q_t .

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

4

Các quá trình Markov rời rạc về thời gian

- Trường hợp đặc biệt của **chuỗi Markov rời rạc về thời gian bậc 1**: sự phụ thuộc xác suất được coi là chỉ liên quan tới trạng thái liền trước:

$$P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \dots] = P[q_t = S_j | q_{t-1} = S_i]$$

- Hơn thế, chỉ quan tâm đến những quá trình mà về phía của công thức trên là độc lập với thời gian, các xác suất chuyển trạng thái a_{ij} có dạng như sau:

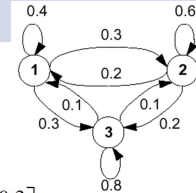
$$a_{ij} = P[q_t = S_j | q_{t-1} = S_i] \quad 1 \leq i, j \leq N$$

$$a_{ij} \geq 0, \forall i, j \quad \sum_{i=1}^N a_{ij} = 1$$

Các quá trình Markov rời rạc về thời gian

Xét một mô hình Markov 3 trạng thái đơn giản trong ví dụ về thời tiết sau:

- Giả sử rằng mỗi ngày thời tiết có thể quan sát được và thuộc một trong các trạng thái :
 - Trạng thái 1 : Trời mưa
 - Trạng thái 2 : Trời nhiều mây
 - Trạng thái 3 : Trời nắng



Ma trận chuyển trạng thái có thể như sau:

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

Các quá trình Markov rời rạc về thời gian

- Có thể đặt ra một số câu hỏi thú vị về thời tiết, chẳng hạn

- Tính xác suất để thời tiết trong 8 ngày nối tiếp nhau là: "Nắng - Nắng - Nắng - Mưa - Mưa - Nắng - Mây - Nắng"

Định nghĩa dãy quan sát O:

O = (Nắng, Nắng, Nắng, Mưa, Mưa, Nắng, Mây, Nắng)

= (3 , 3 , 3 , 1 , 1 , 3 , 2 , 3)

Ngày 1 , 2 , 3 , 4 , 5 , 6 , 7 , 8

P(O|Model): xác suất của dãy quan sát O tương ứng với mô hình đã cho

P(O|Model) = P[3,3,3,1,1,3,2,3|Model]

= P[3].P[3|3].P[1|3].P[1|1].P[3|1].P[2|3].P[3|2]

= $\pi_3(a_{33})^2 \cdot a_{31} \cdot a_{11} \cdot a_{13} \cdot a_{32} \cdot a_{23} = 1 \cdot (0,8) \cdot (0,8) \cdot (0,1) \cdot (0,4) \cdot (0,3) \cdot (0,1) \cdot (0,2)$

= $1,536 \cdot 10^{-4}$

Xác suất trạng thái khởi đầu $\pi_i = P[q_1 = S_i]$ với $1 \leq i \leq N$

Các quá trình Markov rời rạc về thời gian

- Giả sử rằng hệ thống đang ở trong 1 trạng thái đã biết, hãy tính xác suất để hệ thống vẫn giữ nguyên trạng thái đó.

Xác suất này có thể tính như là xác suất của dãy quan sát sau:

O = (S_i, S_i, S_i, ..., S_i, S_j ≠ S_i)

Ngày 1 2 3 ... d d + 1

P(O|Model, q₁ = S_i) =

$\pi_i(a_{ii})^{d-1} (1 - a_{ii})$

= $(a_{ii})^{d-1} (1 - a_{ii})$

= $p_i(d)$

Đại lượng $p_i(d)$ là hàm phân bố xác suất cho khoảng thời gian d giữ trạng thái S_i.

- Kỳ vọng số ngày d_i để giữ trạng thái S_i

$$\bar{d}_i = \sum_{d=1}^{\infty} d p_i(d) = \sum_{d=1}^{\infty} d (a_{ii})^{d-1} (1 - a_{ii}) = \frac{1}{1 - a_{ii}}$$

- Số lượng mong đợi những ngày nối tiếp nhau có nắng là

$$\frac{1}{1 - 0,8} = 5$$

có mây là

$$\frac{1}{1 - 0,6} = 2,5$$

có mưa là

$$\frac{1}{1 - 0,4} = 1,67$$

Mở rộng sang mô hình Markov ẩn

- Phần trên đã xét những mô hình Markov trong đó mỗi trạng thái tương ứng với một sự kiện quan sát được
- Mô hình này có nhiều hạn chế trong việc ứng dụng để giải quyết các vấn đề phức tạp
- Phần này sẽ mở rộng quan niệm về mô hình Markov sang mô hình Markov ẩn trong đó sự kiện quan sát được là một hàm xác suất của trạng thái.
- Mô hình Markov ẩn là một quá trình ngẫu nhiên được nhúng hai lần, trong đó quá trình ngẫu nhiên chính không quan sát được một cách trực tiếp (nó là ẩn) mà chỉ có thể được quan sát thông qua một tập hợp các quá trình ngẫu nhiên khác

Mở rộng sang mô hình Markov ẩn

VÍ DỤ

- Giả sử bạn ở trong một căn phòng với một màn chắn và bạn không thể nhìn thấy điều gì xảy ra.
- Phía bên kia của màn chắn có một người đang thực hiện thí nghiệm gieo đồng xu (sử dụng một hay nhiều đồng xu).
- Tại bất kì thời điểm nào, người này sẽ không nói với bạn là anh ta chọn đồng xu nào để gieo, anh ta chỉ cho bạn biết kết quả của mỗi lần gieo đồng xu là sấp hay ngửa mà thôi. Theo cách đó, một chuỗi các thực nghiệm gieo đồng xu ẩn sẽ được thực hiện với dây quan sát bao gồm một loạt các trường hợp sấp và ngửa. Một dãy quan sát điển hình sẽ là:

$$\begin{aligned} O &= (o_1 \ o_2 \ o_3 \ \dots \ o_T) \\ &= (H \ H \ T \ T \ T \ H \ T \ T \ H \ \dots \ H) \end{aligned}$$

Mở rộng sang mô hình Markov ẩn

- Làm thế nào có thể xây dựng được một mô hình Markov ẩn để giải thích cho dãy quan sát nhận được.
- Những trạng thái trong mô hình sẽ tương ứng với những trường hợp nào và cần có bao nhiêu trạng thái? Chẳng hạn có 1 số lựa chọn:
 - Giả sử 2 đồng xu được gieo. Có 2 trạng thái ứng với đồng xu nào được gieo. Mỗi trạng thái được đặc trưng bởi phân bố xác suất gieo mặt sấp hay mặt ngửa và chuyển biến giữa các trạng thái được đặc trưng bởi ma trận xác suất chuyển trạng thái
 - Có 3 đồng xu được gieo

Các thành phần của một mô hình Markov ẩn

- Mô hình hai đồng xu có 4 tham số chưa biết (a_{11}, a_{22}, P_1, P_2), mô hình ba đồng xu có tới 9 tham số chưa biết ($a_{11}, a_{12}, a_{21}, a_{22}, a_{31}, a_{32}, P_1, P_2, P_3$).
- Một mô hình Markov ẩn được mô tả bởi các thành phần sau đây:
 - N : Số trạng thái của mô hình (trong mô hình tung đồng xu thì mỗi trạng thái tương ứng với một đồng xu). Chỉ quan tâm đến những kiểu chuyển thích hợp các trạng thái trong ứng dụng nhận dạng tiếng nói. Gắn nhãn các trạng thái là $\{1, 2, \dots, N\}$ và kí hiệu trạng thái ở thời điểm t là q_t .
 - M : Số các kí hiệu quan sát khác nhau cho mỗi trạng thái. Những kí hiệu quan sát tương ứng với kết xuất vật lý của hệ thống được mô hình hóa. (Với thí nghiệm gieo đồng xu, các kí hiệu quan sát là mặt "sấp" và "ngửa". Kí hiệu tập các kí hiệu quan sát $V = \{v_1, v_2, \dots, v_M\}$)
 - Ma trận phân bố xác suất chuyển trạng thái $A = \{a_{ij}\}$

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], 1 \leq i, j \leq N$$

Các thành phần của một mô hình Markov ẩn

- 4. Ma trận phân bố xác suất các ký hiệu quan sát
 $B = \{b_j(k)\}$
 $b_j(k)$ là xác suất nhận được ký hiệu quan sát v_k ở trạng thái j :
 $b_j(k) = P[o_t = v_k | q_t = S_j], \quad 1 \leq k \leq M$
- 5. Ma trận phân bố xác suất trạng thái ban đầu $\pi = \{\pi_i\}$
 $\pi_i = P[q_1 = S_i], 1 \leq i \leq N$

Với các giá trị thích hợp của N, M, A, B và π , HMM có thể được sử dụng để đưa ra một chuỗi quan sát $O = O_1 O_2 \dots O_T$ như sau: (Mỗi quan sát O_t là một trong các ký hiệu từ tập V . T là số lượng các quan sát của chuỗi)

- 1) Chọn trạng thái ban đầu $q_t = S_i$ theo phân bố trạng thái ban đầu π .
- 2) Đặt $t = 1$
- 3) Chọn $O_t = v_k$ theo phân bố xác suất ký hiệu ở trạng thái S_i , tức là $b_i(k)$.
- 4) Chuyển đến trạng thái mới $q_{t+1} = S_j$, theo phân bố xác suất chuyển trạng thái cho trạng thái S_i , tức là a_{ij} .
- 5) Đặt $t = t + 1$; trở lại bước 3) nếu $t < T$; nếu không chấm dứt thủ tục

Ba bài toán cơ bản của mô hình Markov ẩn

- Các ràng buộc về mặt thống kê như sau:

$$\sum_{j=1}^N a_{ij} = 1 \quad \text{với mọi } i = 1, 2, \dots, N$$

$$\sum_{k=1}^M b_j(k) = 1 \quad \text{với mọi } j = 1, 2, \dots, M$$

$$\sum_{i=1}^N \pi_i = 1$$

- HMM có thể được biểu diễn gọn dưới dạng $\lambda = (A, B, \pi)$

Ba bài toán cơ bản của mô hình Markov ẩn

1. Đánh giá xác suất (Evaluation Problem)

Cho dãy quan sát $O = (O_1 O_2 \dots O_T)$ và mô hình Markov ẩn $\lambda = (A, B, \pi)$. Hãy tính $P(O|\lambda)$ là xác suất của dãy quan sát theo mô hình.

2. Tìm dãy trạng thái tối ưu (Decoding Problem)

Cho dãy quan sát $O = (O_1 O_2 \dots O_T)$ và mô hình λ . Làm thế nào lựa chọn được 1 dãy trạng thái $q = (q_1 q_2 \dots q_T)$ là tối ưu theo một nghĩa nào đó.

3. Ước lượng tham số của mô hình (Learning / Training Problem)

Cần phải điều chỉnh các tham số của mô hình như thế nào để $P(O|\lambda)$ đạt cực đại.

Nhận dạng tiếng nói

- Hai giai đoạn: huấn luyện (học) – nhận dạng
- Phân loại theo
 - Số lượng từ vựng
 - Từ rời rạc – liên tục
 - Một người nói – nhiều người nói
 - Nhận dạng từ – câu

Phân loại theo độ phức tạp

- Nhận dạng từ riêng lẻ, từ vựng ít (<100), một người nói
- Từ vựng nhiều hơn (vài nghìn từ), một người nói
- Như trên nhưng cho hệ thống nhiều người nói
- Nhận dạng các từ đi với nhau, từ vựng ít (hàng chục từ)
- Nhận dạng câu ngắn, từ vựng hạn chế, một người nói
- Như trên nhưng cho hệ thống nhiều người nói
- Nhận dạng lời nói liên tục, một hoặc nhiều người nói

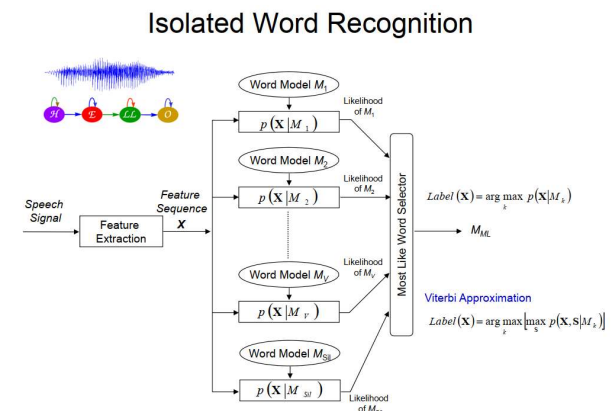
Nhận dạng người nói (Speaker Recognition)

- Kiểm tra (xác thực) (verification) giọng nói
- Định danh (identification) giọng nói

Một số vấn đề đối với hệ thống nhận dạng tiếng nói

- Phát hiện khoảng lặng, phát hiện tiếng nói
- Cải thiện chất lượng tín hiệu tiếng nói (giảm nhiễu)
- Tiếng nói được phát âm với thời hạn và nhịp điệu khác nhau
- Mô hình nhận dạng
 - Mô hình Markov ẩn (Hidden Markov Model: HMM)
 - Mạng nơ-ron

Sơ đồ khối nhận dạng từ rời rạc với HMM



Measures of ASR Performance

- Evaluating the performance of automatic speech recognition (ASR) systems is critical, and the Word Recognition Error Rate (WER) is one of the most important measures
 - There are typically three types of word recognition errors
 - Substitution**
 - Deletion**
 - Insertion**
 - An incorrect word was substituted for the correct word
 - A correct word was omitted in the recognized sentence
 - An extra word was added in the recognized sentence
- How to determine the minimum error rate?

Measures of ASR Performance

- Calculate the WER by aligning the correct word string against the recognized word string
 - A **maximum substring matching** problem
 - Can be handled by **dynamic programming**

- Example:

Correct : "the effect is clear"
Recognized: "effect is not clear"

deleted
matched inserted matched

 - Error analysis: **one deletion and one insertion**
 - Measures: **word error rate (WER)**, **word correction rate (WCR)**, **word accuracy rate (WAR)**

Might be higher than 100%

Word Error Rate = 100% $\frac{\text{Sub. + Del. + Ins. words}}{\text{No. of words in the correct sentence}} = \frac{2}{4} = 50\%$

Word Correction Rate = 100% $\frac{\text{Matched words}}{\text{No. of words in the correct sentence}} = \frac{3}{4} = 75\%$

Word Accuracy Rate = 100% $\frac{\text{Matched - Ins. words}}{\text{No. of words in the correct sentence}} = \frac{3-1}{4} = 50\%$

Might be negative

WER+WAR=100%

TỔNG HỢP TIẾNG NÓI

Khái niệm và phân loại

- Dùng phần cứng và phần mềm để tạo tiếng nói xuất phát từ biểu diễn ngữ âm, ngữ nghĩa của lời nói.
- Hệ thống chuyển văn bản thành tiếng nói (Text-To-Speech (TTS) chuyển đổi văn bản thông thường thành tiếng nói; các hệ thống khác cho ra tiếng nói từ các biểu diễn ngôn ngữ tượng trưng như phiên âm.
- Kỹ thuật tổng hợp tiếng nói:
 - Tổng hợp trực tiếp
 - Tổng hợp dựa trên mô hình
 - Formant
 - LPC
 - Mô phỏng bộ máy phát âm
 - HMM
 - Học sâu

Phân loại

- Chất lượng bộ tổng hợp: Mức độ tự nhiên
 - Mức độ rõ
 - Thanh điệu
 - Ngữ điệu
- Số lượng từ vựng:
 - Hạn chế
 - Không hạn chế

Đánh giá chất lượng tiếng nói tổng hợp

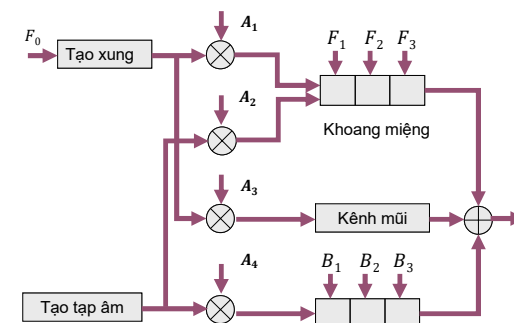
- MOS: Mean Opinion Scores

SCORES	QUALITY
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

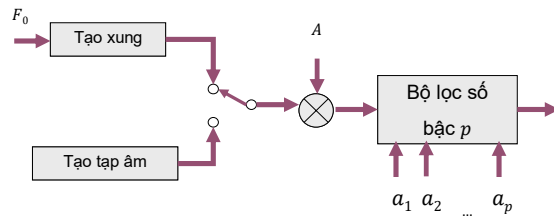
Tổng hợp trực tiếp

- Ghi âm tiếng nói tự nhiên
 - Đơn vị ghi âm
 - Ghép các đơn vị ghi âm: từ, câu.
 - Đơn vị ghi âm
 - Âm vị : hiện tượng đồng cấu âm (coarticulation)
 - Âm tiết (diphone - âm vị kép)
 - Từ
 - Tổ hợp từ
 - câu
- $\text{nam} = n + a + m$
 $\quad = n + am$
 $\quad = na + m$
 $\quad = na + am$

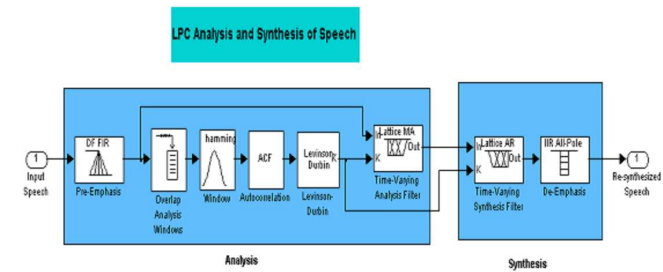
Tổng hợp theo mô hình - Tổng hợp formant



Tổng hợp theo mô hình - Tổng hợp LPC

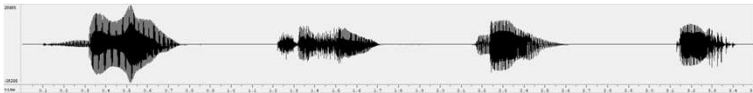


Ví dụ thực hiện bằng MATLAB LPC Synthesis-by-Analysis

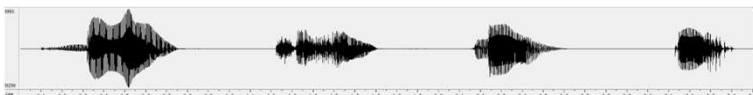


Synthesis-by-Analysis

Dạng sóng tiếng nói tự nhiên

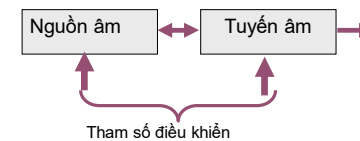


Dạng sóng tiếng nói tổng hợp



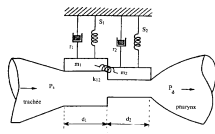
Tổng hợp theo mô hình

Mô phỏng hệ thống tổng hợp

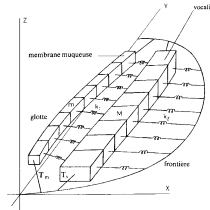


- Mô phỏng nguồn âm (nguồn tuần hoàn)
Mô phỏng dây thanh: Mô hình một khối, Mô hình hai khối,
Mô hình nhiều khối, Mô hình hai dầm...

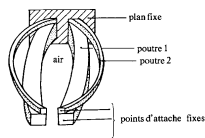
Mô hình nguồn âm



Mô hình 2 khối

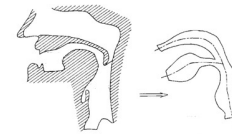


Mô hình nhiều khối

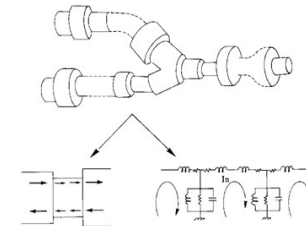


Mô hình 2 dầm

Mô phỏng tuyến âm



Ổng âm tương đương



Ổng âm được rời rạc hóa

Tổng hợp theo mô hình - Dựa trên HMM

