**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**🙠🕮🙢**

**XỬ LÝ NGÔN NGỮ NÓI**

**BÀI TẬP 1: NHẬN DẠNG NGƯỜI NÓI**

GVHD: PGS. TS. Vũ Hải Quân

Nhóm thực hiện:

Ngành Khoa Học Máy Tính – Cao học khóa 23

1. Đỗ Đặng Minh - 13 11 015
2. Huỳnh Công Toàn - 13 11 026
3. Hồ Văn Tấn - 13 11 058

*Tp. Hồ Chí Minh, tháng 04 năm 2014*

# Thông tin nhóm

|  |  |  |  |
| --- | --- | --- | --- |
| **MSHV** | **Họ tên** | **Số điện thoại** | **E-mail** |
| 13 11 015 | Đỗ Đặng Minh | 0168-993-5242 | [masterminh219@gmail.com](mailto:masterminh219@gmail.com) |
| 13 11 026 | Huỳnh Công Toàn | 0121-516-1090 | [alex7huynh@gmail.com](mailto:alex7huynh@gmail.com) |
| 13 11 058 | Hồ Văn Tấn | 090-290-9334 | [tanhv90@gmail.com](mailto:tanhv90@gmail.com) |

Mục lục

[Thông tin nhóm 1](#_Toc390172883)

[I. Tổng quan về nhận dạng tiếng nói 2](#_Toc390172884)

[II. Cơ sở lý thuyết 3](#_Toc390172885)

[1) Trích chọn đặc trưng 3](#_Toc390172886)

[2) Hàm Gauss 3](#_Toc390172887)

[3) Gaussian mixture model 4](#_Toc390172888)

[III. Thực nghiệm 4](#_Toc390172889)

[1) Dữ liệu thực nghiệm 4](#_Toc390172890)

[2) Phương pháp thực hiện 5](#_Toc390172891)

[3) Mô tả ứng dụng 6](#_Toc390172892)

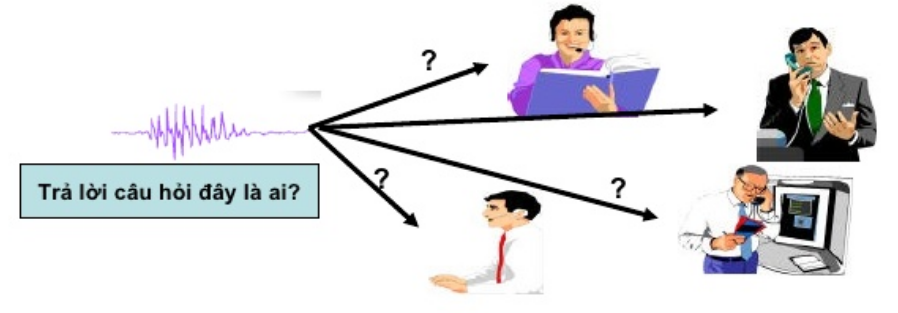
[4) Thống kê kết quả 6](#_Toc390172893)

[IV. Tài liệu tham khảo 9](#_Toc390172894)

# Tổng quan về nhận dạng tiếng nói

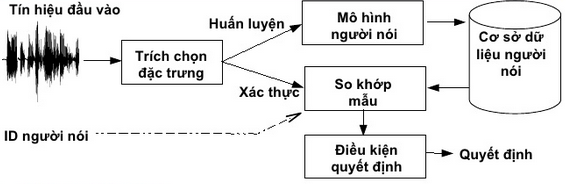
Trong khoa học máy tính và kỹ thuật điện tử, nhận dạng tiếng nói là quá trình phiên dịch ngôn ngữ nói thành chữ.

Nhận dạng giọng nói hoặc xác định người nói là quá trình xác định người đang nói hơn là tìm hiểu họ đang nói gì. Nhận dạng giọng nói có thể đơn giản hóa quá trình phiên dịch giọng nói của hệ thống, đồng thời cũng có thể dùng để xác định danh tính của một người trong quy trình an ninh nghiệp vụ.



Nhận dạng tiếng nói đóng vai trò quan trọng trong giao tiếp giữa người và máy. Nó giúp máy móc hiểu và thực hiện các hiệu lệnh của con người. Hiện nay trên thế giới, lĩnh vực nhận dạng tiếng nói đã đạt được nhiều tiến bộ vượt bậc. Tuy nhiên, hầu hết các công trình vẫn còn thực hiện ở trong phòng thí nghiệm, đặc biệt là đối với các thứ tiếng ít phổ biến như tiếng Việt của chúng ta.

Nhận dạng tiếng nói là một quá trình nhận dạng mẫu, với mục đích là phân lớp thông tin đầu vào là tín hiệu tiếng nói thành một dãy tuần tự các mẫu đã được học trước đó và lưu trữ trong bộ nhớ. Các mẫu là các đơn vị nhận dạng, chúng có thể là các từ, hoặc các âm vị. Nếu các mẫu này là bất biến và không thay đổi thì công việc nhận dạng tiếng nói trở nên đơn giản bằng cách so sánh dữ liệu tiếng nói cần nhận dạng với các mẫu đã được học và lưu trữ trong bộ nhớ.



Trích chọn đặc trưng là quá trình biến đổi giọng nói thô thành đặc trưng của người nói. Mô hình người nói có được nhờ huấn luyện người dựa vào một phương pháp cụ thể. Mô số mô hình người nói phổ biến là: lượng tử hóa vectơ (vector quantization), hỗn hợp Gauss (GMM), SVM. So khớp mẫu là quá trinh so khớp tiếng nói đưa vào và cơ sở dữ liệu đã huấn luyện từ trước.

Khó khăn cơ bản của nhận dạng tiếng nói đó là tiếng nói luôn biến thiên theo thời gian và có sự khác biệt lớn giữa tiếng nói của những người nói khác nhau, tốc độ nói, ngữ cảnh và môi trường âm học khác nhau, đặc biệt là nhiễu. Xác định những thông tin biến thiên nào của tiếng nói là có ích và những thông tin nào là không có ích đối với nhận dạng tiếng nói là rất quan trọng. Đây là một nhiệm vụ rất khó khăn mà ngay cả với các kỹ thuật xác suất thống kê mạnh.

Các cách tiếp cận nhận dạng tiếng nói bằng thống kê bao gồm: sử dụng mô hình Markov ẩn, Dynamic time warping, mạng nơ-ron, ... Trong phạm vi bài tập 1 với yêu cầu nhận dạng người nói, nhóm xin trình bày tập trung vào phương pháp Gaussian mixture model (GMM).

# Cơ sở lý thuyết

1. Trích chọn đặc trưng

Các đặc trưng mang thông tin về tiếng nói là các tần số formant, tần số âm cơ bản và năng lượng.

Các đặc trưng mang thông tin về đường bao phổ:

* Hệ số dự đoán tuyến tính (LPC).
* Các hệ số cepstrum.
* Các hệ số cepstrum dự đoán tuyến tính (LPCC).
* Các hệ số cepstrum tần số mel (MFCC).

Trích chọn đặc trưng trong xử lý ngôn ngữ nói là biến đổi từ giọng nói thô thành những đặc trưng của người nói. Đặc trưng được sử dụng trong bài này là đặc trưng MFCC (Mel Frequency Cepstral Coefficients), đặc trưng MFCC thể hiện hình thái của cuống họng, … khi ta nói.

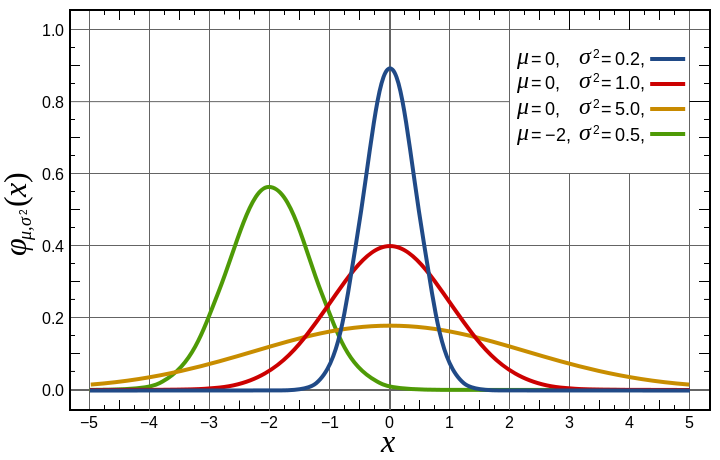
Các bước trính chọn đặc trưng MFCC:

* Phân khung (Frame blocking): chia tín hiệu đầu vào thành các đoạn nhỏ khoảng 20ms­30ms. Phân khung tín hiệu mỗi khung N mẫu, hai khung kề nhau lệch M mẫu.
* Lấy cửa sổ (Windowing): giảm sự gián đoạn của tín hiệu ở đầu và cuối mỗi khung vừa được chia.
* Biến đổi FFT (Fast Fourier Transform): chuyển đổi mỗi khung với N mẫu từ miền thời gian sang miền tần số.
* Biến đổi sang thang đo mel (Mel-frequency wrapping).
* Hệ số cepstrum (cepstral coefficients).

1. Hàm Gauss

Hàm Gauss là một hàm có dạng: với các hằng số thực *a, b, c, d*.

Biểu đồ của một hàm Gauss là một đường cong đối xứng đặc trưng "hình quả chuông". Đường cong này rớt xuống rất nhanh khi tiến tới cộng/trừ vô cùng. Tham số a là chiều cao tối đa đường cong, b là vị trí tâm của đỉnh và c quyết định chiều rộng của "chuông".



1. Gaussian mixture model

Mô hình hỗn hợp Gauss (GMM) là một dạng mô hình thống kê được xây dựng từ việc huấn luyện các tham số thông qua dữ liệu học.

Về cơ bản, mô hình GMM xấp xỉ một hàm mật độ xác suất bằng hợp các hàm mật độ Gauss.

Phân phối chuẩn, còn gọi là phân phối Gauss, là một phân phối xác suất cực kì quan trọng trong nhiều lĩnh vực. Nó là họ phân phối có dạng tổng quát giống nhau, chỉ khác tham số vị trí (giá trị trung bình μ) và tỉ lệ (phương sai σ2).

Hàm mật độ xác suất của phân phối chuẩn với trung bình μ và phương sai σ2 (hay độ lệch chuẩn σ) là một ví dụ của một hàm Gauss.

Cho trước m phân phối Gauss p1, p2,…,pm. Hàm mật độ xác suất của mô hình chính là tổng trọng của m phân phối Gauss theo công thức:

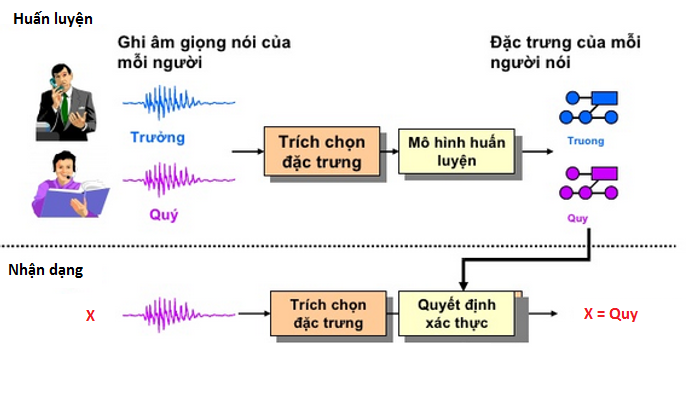
trong đó là trọng số của phân phối Gauss thứ i, với điều kiện và . Các trọng số này thể hiện mức độ ảnh hưởng của mỗi phân phối Gauss đối với mô hình GMM. Như vậy, phân phối Gauss có phương sai và trọng số lớn thì sẽ có ảnh hưởng lớn đến kết xuất của mô hình.

# Thực nghiệm

1. Dữ liệu thực nghiệm

Dữ liệu huấn luyện là 5 tập tin wav (mỗi tập tin có độ dài 45 giây) của 5 người khác nhau.

Dữ liệu để kiểm thử là 25 tập tin (mỗi tập tin có độ dài 3 giây) của 5 người khác nhau. Mục tiêu là xác định xem giọng nói trong tập tin wav thuộc về ai trong bộ dữ liệu đã huấn luyện.



1. Phương pháp thực hiện

**Phương pháp Gauss**

Bước 1: rút trích đặc trưng

* Đọc tập tin wav vào mảng a bằng hàm *wavread*.
* Trích chọn đặc trưng MFCC cho mảng a bằng hàm *mfcc*, ta được ma trận *a* gồm các vectơ đặc trưng và 12 thành phần.

Bước 2: nhận dạng giọng nói

* Đưa một tập tin mẫu vào, tính xác suất của từng đặc trưng bằng hàm *mean(mvnpdf(x,m,v))*. Trong đó *mean* để lấy trung bình, *mvnpdf* là hàm Gauss, *x* là ma trận của tập tin mẫu tính như phương pháp trên, *m* là trung bình của ma trận đặc trưng *a*, *v* là phương sai của ma trận đặc trưng *a*.
* Tính lần lượt xác suất của tất cả các đặc trưng trong cơ sở dữ liệu với tập tin mẫu. Xác suất của đặc trưng nào lớn nhất sẽ quyết định giọng nói cần kiểm thử là thuộc đặc trưng nào.

**Phương pháp Gauss Mixture Model**

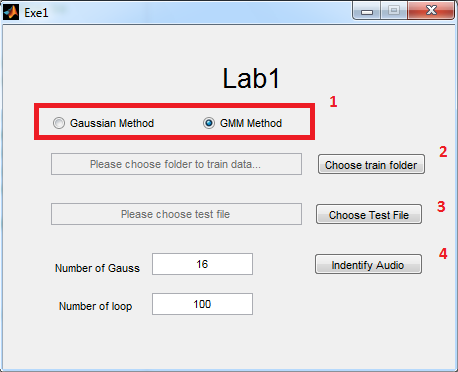
Bước 1: rút trích đặc trưng

* Đọc tập tin wav vào mảng *a* bằng hàm *wavread*.
* Trích chọn đặc trưng MFCC cho mảng a bằng hàm *mfcc*, ta được ma trận *a* gồm các vectơ đặc trưng và 12 thành phần.
* Dùng hàm *gNew(12, m, ‘diag’)* để tạo cấu trúc Gauss, 12 là số chiều vectơ đặc trưng, *m* là số phân phối Gauss (số chuông), ‘diag’ là đường chéo.
* Khởi tạo giá trị ban đầu cho cấu trúc Gauss bằng hàm *gInit(g0, a, loop)* với g0 là cấu trúc Gauss tạo ở trên, *a* là ma trận tạo bằng *mfcc* ở trên, *loop* là số vòng lặp.
* Dùng tiếp hàm *gRE(g1, a, loop)* với g1 là cấu trúc Gauss đã khởi tạo, *a* và *loop* như trên.

Bước 2: nhận dạng giọng nói

* Đưa một tập tin mẫu vào, tính xác suất của từng đặc trưng bằng hàm *mean(gPr(g2, x))*. Trong đó mean để lấy trung bình, gPr để tính xác suất bằng hàm Gauss, g2 là cấu trúc Gauss của một đặc trưng huấn luyện từ trước, *x* là ma trận của tập tin mẫu tính bằng hàm mfcc.
* Tính lần lượt xác suất của tất cả các đặc trưng trong cơ sở dữ liệu với tập tin mẫu. Xác suất của đặc trưng nào lớn nhất sẽ quyết định giọng nói cần kiểm thử là thuộc đặc trưng nào.

1. Mô tả ứng dụng



Bước 1: chọn phương pháp Gauss hoặc Gaussian Mixture Model.

Bước 2: chọn thư mục chứa các tập tin cần huấn luyện.

Bước 3: chọn tập tin wav muốn nhận dạng.

Bước 4: nhấn nút để chương trình tính toán và hiển thị hộp thoại kết quả (nhãn và xác suất).

1. Thống kê kết quả

Kết quả thống kê gồm xác suất của từng tập tin mẫu cần nhận dạng (cột) ứng với từng tập tin huấn luyện (dòng). Xác suất lớn nhất là số được in đậm.

**Phương pháp Gauss**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SV0001 | SV0002 | SV0007 | SV0041 | SV0044 |
| SV0001\_23 | 0.000000520994 | **0.000002317675** | 0.000001520021 | 0.000000898821 | 0.000000685868 |
| SV0001\_25 | 0.000000537058 | **0.000002818226** | 0.000002271859 | 0.000001222677 | 0.000000707450 |
| SV0001\_28 | 0.000000624404 | **0.000003452752** | 0.000003224532 | 0.000001425355 | 0.000001013923 |
| SV0001\_31 | 0.000000423809 | **0.000003025508** | 0.000002827425 | 0.000001105947 | 0.000000809340 |
| SV0001\_34 | 0.000000610586 | **0.000002636309** | 0.000002255328 | 0.000001047086 | 0.000000807033 |
| SV0002\_35 | 0.000000597149 | **0.000003735102** | 0.000003272886 | 0.000001591670 | 0.000001086709 |
| SV0002\_38 | 0.000000637358 | **0.000004103172** | 0.000003302820 | 0.000001684048 | 0.000000892077 |
| SV0002\_43 | 0.000000653771 | **0.000003523427** | 0.000002540827 | 0.000001570174 | 0.000000733988 |
| SV0002\_47 | 0.000000807193 | **0.000006577312** | 0.000005667378 | 0.000002690644 | 0.000001338189 |
| SV0002\_50 | 0.000000448035 | **0.000003158191** | 0.000002500723 | 0.000001266472 | 0.000000722279 |
| SV0007\_20 | 0.000000751807 | **0.000006357736** | 0.000006071189 | 0.000002700616 | 0.000001329167 |
| SV0007\_22 | 0.000000613832 | 0.000004266339 | **0.000004279589** | 0.000001778953 | 0.000000946878 |
| SV0007\_25 | 0.000000815333 | **0.000006277470** | 0.000005631564 | 0.000002305169 | 0.000001243346 |
| SV0007\_30 | 0.000000524137 | 0.000004683597 | **0.000004793771** | 0.000001893029 | 0.000000900189 |
| SV0007\_40 | 0.000000449025 | **0.000003152150** | 0.000002827124 | 0.000001175933 | 0.000000766886 |
| SV0041\_48 | 0.000000272788 | **0.000002393545** | 0.000002220315 | 0.000000924629 | 0.000000512861 |
| SV0041\_53 | 0.000000586242 | **0.000004687901** | 0.000004823390 | 0.000001819343 | 0.000000942925 |
| SV0041\_61 | 0.000000346743 | **0.000002955807** | 0.000002748886 | 0.000001241565 | 0.000000647694 |
| SV0041\_67 | 0.000000325930 | **0.000002616026** | 0.000002591489 | 0.000001169783 | 0.000000745825 |
| SV0041\_73 | 0.000000273988 | **0.000002048848** | 0.000001899229 | 0.000000795375 | 0.000000475606 |
| SV0044\_30 | 0.000000231551 | **0.000000674874** | 0.000000581778 | 0.000000410623 | 0.000000575218 |
| SV0044\_31 | 0.000000151993 | **0.000000460548** | 0.000000372770 | 0.000000296511 | 0.000000416310 |
| SV0044\_32 | 0.000000145415 | **0.000000576333** | 0.000000425390 | 0.000000293347 | 0.000000350962 |
| SV0044\_33 | 0.000000193582 | **0.000000959639** | 0.000000813732 | 0.000000518460 | 0.000000555281 |
| SV0044\_34 | 0.000000203645 | **0.000001113275** | 0.000000699937 | 0.000000469130 | 0.000000398555 |

Tỷ lệ nhận dạng chính xác = 7/25 = 28%

**Phương pháp Gauss Mixture Model với số Gauss 25 và 100 vòng lặp**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SV0001 | SV0002 | SV0007 | SV0041 | SV0044 |
| SV0001\_23 | 0.0000418757 | **0.0001346524** | 0.0000387753 | 0.0000364559 | 0.0000318685 |
| SV0001\_25 | 0.0002430980 | **0.0010595475** | 0.0002028502 | 0.0000901465 | 0.0001082989 |
| SV0001\_28 | 0.0011792385 | **0.0411243699** | 0.0270785876 | 0.0139945126 | 0.0016999623 |
| SV0001\_31 | 0.0122939636 | 7.4643306719 | 9.0516455797 | **12.5890430062** | 0.3778269042 |
| SV0001\_34 | 0.0000441312 | 0.0001472533 | **0.0003205731** | 0.0001759411 | 0.0000862451 |
| SV0002\_35 | 0.0025192134 | **2.2153918788** | 1.7413733951 | 1.8530418738 | 0.0642489322 |
| SV0002\_38 | 0.0018023379 | **0.1214782528** | 0.0029086504 | 0.0087479165 | 0.0037847450 |
| SV0002\_43 | 0.0003187888 | **0.0578397215** | 0.0004651743 | 0.0011466140 | 0.0012297074 |
| SV0002\_47 | 0.0058243729 | 0.4388727828 | **0.4990992312** | 0.0440230181 | 0.0741702585 |
| SV0002\_50 | 0.0060134213 | **0.8751617807** | 0.1003970657 | 0.0193491156 | 0.0456328614 |
| SV0007\_20 | 0.0093779928 | 6.3230485493 | 113.9182412940 | **168.3053966757** | 0.8638193643 |
| SV0007\_22 | 0.0042994624 | 1.1824651019 | **22.8418781647** | 1.5343672430 | 0.9926376745 |
| SV0007\_25 | 0.0069562854 | 2.8417665277 | **5.1484419704** | 3.6633190188 | 0.1486926775 |
| SV0007\_30 | 0.0071494827 | 9.1236900401 | **614.9578353224** | 331.3954736565 | 1.2113728053 |
| SV0007\_40 | 0.0073746214 | 4.1634450933 | **22.6614520303** | 3.5734479939 | 0.3536010922 |
| SV0041\_48 | 0.0040364811 | 4.0426690254 | 91.2886332375 | **180.7957100595** | 1.0047812695 |
| SV0041\_53 | 0.0022474208 | 2.4380313388 | 1,840.7176988424 | **46,560.5900277658** | 2.9737188409 |
| SV0041\_61 | 0.0051447604 | 3.5676460714 | **51.3003635921** | 17.5208466158 | 0.3234304011 |
| SV0041\_67 | 0.0032027750 | 2.2558638269 | 130.8130997664 | **1,626.7889864947** | 0.2771295398 |
| SV0041\_73 | 0.0023430949 | **0.2714528569** | 0.0419526145 | 0.0080329681 | 0.0105376550 |
| SV0044\_30 | 0.0000004471 | 0.0000016257 | **0.0033026496** | 0.0004184434 | 0.0001070063 |
| SV0044\_31 | 0.0000002908 | 0.0000016878 | 0.0000065405 | 0.0000008073 | **0.0000231418** |
| SV0044\_32 | **0.0000953721** | 0.0000078793 | 0.0000027857 | 0.0000011143 | 0.0000164303 |
| SV0044\_33 | 0.0008934727 | **0.0641644782** | 0.0014728893 | 0.0024748095 | 0.0361515563 |
| SV0044\_34 | 0.0001340042 | 0.0001029008 | 0.0000740070 | 0.0000682903 | **0.0029886943** |

Tỷ lệ nhận dạng chính xác = 13/25 = 52%

**Nhận xét:**

Tỷ lệ nhận dạng của 2 phương pháp còn thấp do chưa cắt silence ở các file, độ dài tập tin huấn luyện ngắn (45 giây) và độ dài của tập tin cần nhận dạng cũng rất ngắn (3 giây), dù tập dữ liệu rõ (thu âm từ micro chất lượng tốt).

Mô hình GMM có hiệu suất tốt hơn hẳn so với mô hình Gaussian.

# Tài liệu tham khảo

Website:

1. <http://www.slideshare.net/TungLe1/nhan-dang-tieng-noi>
2. <http://www.scribd.com/doc/98608190/1111089-1111062-1111055>
3. <http://en.wikipedia.org/wiki/Gaussian_function>