ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

NGUYỄN HỮU LỢI - 16C11021

TRẦN QUỐC TRƯỜNG – 16C11032

XỬ LÝ TIẾNG NÓI

BÀI TẬP 1 :

Nhận Dạng Người Nói Sử Dụng MFCC VÀ GMM

GIẢNG VIÊN HƯỚNG DẪN

PGS.TS Vũ Hải Quân

Khóa K26

Nhận dạng Người Nói Sử dụng MFCC và GMM

Tóm tắt

Sinh trắc học – hay công nghệ sử dụng các đặc điểm sinh học của con người để nhận diện là một lĩnh vực rất đa dạng và có nhiều ứng dụng quan trọng trong thực tiễn. Trong các lĩnh vực của sinh trắc học, tiếng nói nhận được rất nhiều sự quan tâm do tính tự nhiên của giọng nói, sự dễ dàng trong thu thập và sử dụng giọng nói trong quá trình nhận diện người nói. Nhiều phương pháp đã được nghiên cứu và đạt được những hiệu quả nhất định trong quá trình nhận diện người nói.

Báo cáo sẽ lần lượt trình bày giới thiệu chung về giọng nói, các bài toán trong nhận diện người nói và các phương pháp nhận diện người nói. Sau đó, báo cáo sẽ đi sâu vào phương pháp rút trích đặc trưng MFCC và mô hình hóa người nói sử dụng GMM. Cuối cùng, bài báo cáo sẽ trình bày một số kết quả thực nghiệm nhận diện người nói dựa trên phương pháp vừa được trình bày.

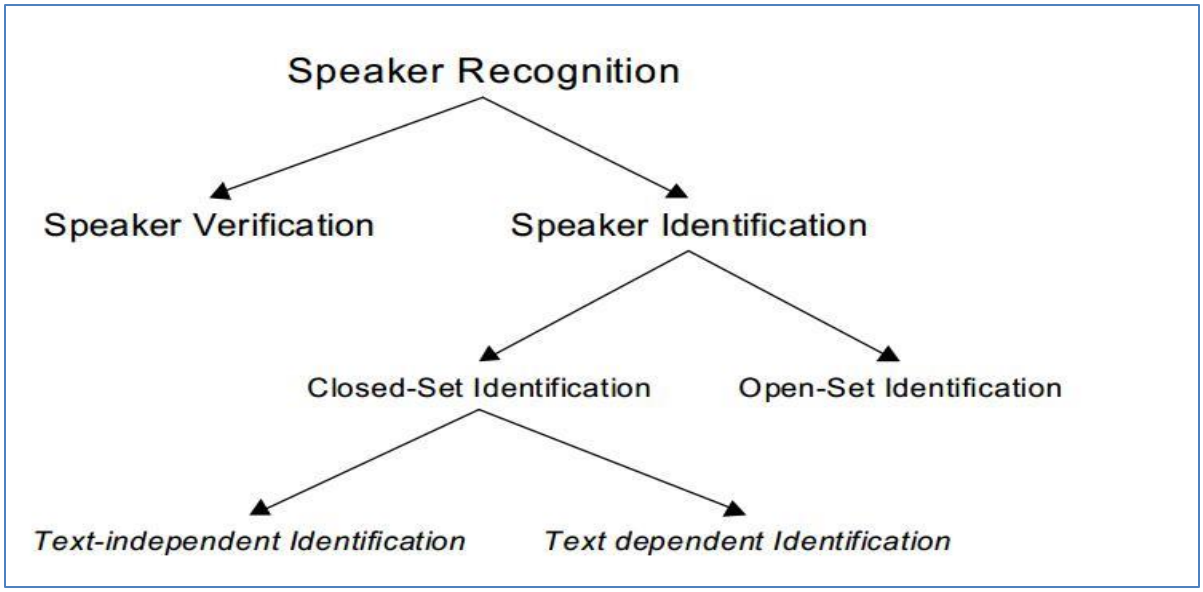
I. Giới thiệu chung

1. Tiếng nói và nhận diện người nói

Tiếng nói là hình thức giao tiếp cơ bản nhất của con người. Tiếng nói của con người bao gồm rất nhiều loại thông tin: Nội dung của lời nói (từ và ngôn ngữ), cảm xúc của người nói, giới tính và định danh người nói… Mục tiêu của quá trình nhận dạng người nói là rút trích, mô tả và nhận diện người dựa vào các đặc trưng của tiếng nói.

2. Các loại nhận dạng người nói

Nhận diện người nói thường được chia làm hai nhánh khác nhau là xác nhận người nói (speaker verification) và định danh người nói (speaker identification).



Hình 1: Các nhánh của bài toán nhận diện người nói

- Xác nhận người nói là quá trình xác nhận người hiện tại có phải là người mong muốn dựa vào giọng nói. Quá trình này là quá trình xác định có / không và không quan tâm cụ thể người nói là ai.

- Định danh người nói lại được chia làm hai nhánh nhỏ hơn, là định danh người nói trên tập mở và định danh người nói trên tập đóng. Định danh người nói trên tập mở cần phải xác định xem người nói là ai trong danh sách người nói đã biết, hoặc kết luận người này không thuộc danh sách người nói đã biết. Định danh người nói trên tập đóng chỉ xét dữ liệu chắc chắn là của một người trong danh sách những người đã biết.

Ngoài ra, dựa vào thuật toán, người ta cũng chia ra hai loại, đó là nhận diện người nói phụ thuộc văn bản và nhận diện người nói không phụ thuộc văn bản. Nhận diện người nói phụ thuộc văn bản yêu cầu người nói phải nói chính xác những từ đã được cho trước, trong khi đó nhận diện người nói không phụ thuộc văn bản có thể nhận diện khi người nói nói bất cứ từ gì.

3. Ứng dụng

Ứng dụng của hệ thống nhận diện người nói trên thực tế là cực kỳ đa dạng. Một số ứng dụng gần đây có thể được kể đến như sau:

- Vào tháng 5/2013, Barclays Wealth đã công bố rằng ông đã dùng hệ thống nhận dạng người nói để xác minh các khách hàng qua điện thoại trong 30 giây thông qua một cuộc trò chuyện bình thường. Hệ thống này được phát triển bởi chuyên gia phân tích giọng nói Nuance – công ty đứng sau công nghệ của siri của Apple.

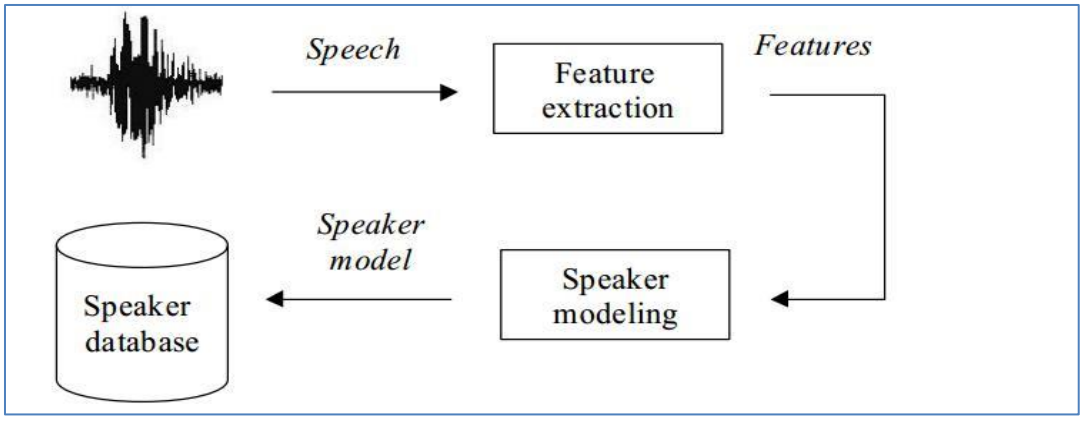
- Các ngân hàng tư nhân của Barclays là công ty dịch vụ tài chính đầu tiên triển khai sinh trắc học bằng giọng nói để xác minh khách hàng gọi đến trung tâm của họ. 93% khách hàng đánh giá hệ thống này 9/10 điểm về tốc độ, dễ sử dụng và bảo mật.
- Tháng 8/2014 tập đoàn GoVivace phát triển một hệ thống nhận dạng người nói cho phép họ tìm kiếm một người trong hàng triệu người chỉ bằng cách đơn giản là ghi âm giọng nói của họ.
- Hệ thống nhận dạng người nói còn có thể dùng để sử dụng trong điều tra hình sự.

4. Phương pháp trình bày

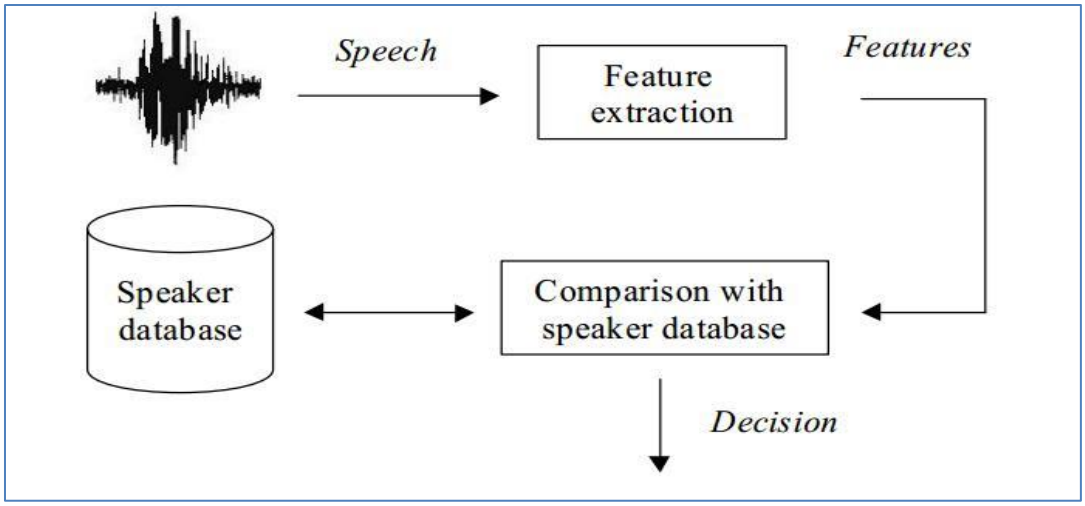
Có rất nhiều phương pháp rút trích đặc trưng như MFCC, LPCC và phương pháp phân lớp, mô hình hóa người nói như sử dụng HMM, GMM, hay không mô hình hóa và sử dụng một thuật toán phân lớp như neural networks, SVM. Bài báo cáo sẽ tập trung trình bày phương pháp nhận diện người nói không phụ thuộc văn bản trên tập đóng sử dụng đặc trưng MFCC (Mel Frequency Cepstrum Coeffecient) và GMM (Gaussian mixture model).

II. Hệ thống nhận dạng người nói

Quá trình nhận diện người nói được thực hiện qua các pha. Có hai pha trong quá trình này:



- Pha đăng ký người nói: Tiếng nói của người cần nhận diện được thu thập và sử dụng để huấn luyện mô hình. Tập các mô hình của nhiều người nói còn được gọi là cơ sở dữ liệu người nói.



- Pha định danh người nói: Dữ liệu tiếng nói của một người dùng không rõ định danh được đưa vào hệ thống và so khớp với các mô hình trong cơ sở dữ liệu người nói.

_ Chi tiết hai pha như sau:

- Cả hai pha đều có chung hai bước đầu. Bước đầu tiên là thu thập tiếng nói. Tiếng nói có thể được thu thập thông qua micro và chuyển thành tín hiệu rời rạc – tín hiệu số (digital). Tuy nhiên dữ liệu này thông thường sẽ bị nhiễu, do đó cần phải được tiền xử lý trước khi đưa vào pha bước thứ hai.
- Bước thứ hai đó là rút trích đặc trưng, nhằm mục đích giảm kích thước dữ liệu nhưng vẫn đảm bảo đủ thông tin để phân biệt người nói. Trong bài báo cáo sẽ trình bày đặc trưng MFCC.

- Ở bước thứ ba của pha đăng ký, thông tin người nói sau khi đã được rút trích đặc trưng được mô hình hóa (modeling) và lưu vào cơ sở dữ liệu. Bài báo cáo sẽ sử dụng Gaussian mixture model để mô hình hóa dữ liệu người nói và sử dụng EM (Expectation Maximization) để xây dựng GMM tương ứng với các đặc trưng MFCC được truyền vào.
- Ở bước thứ ba của pha định danh, dữ liệu rút trích được so khớp với các dữ liệu trong cơ sở dữ liệu và đưa ra quyết định xem người đó là ai.

➢ *Có thể thấy hai pha được thực hiện tách biệt nhau nhưng có liên quan rất gần với nhau, trong đó hai pha khó thực hiện nhất đó là rút trích đặc trưng và mô hình hóa, so khớp dữ liệu*

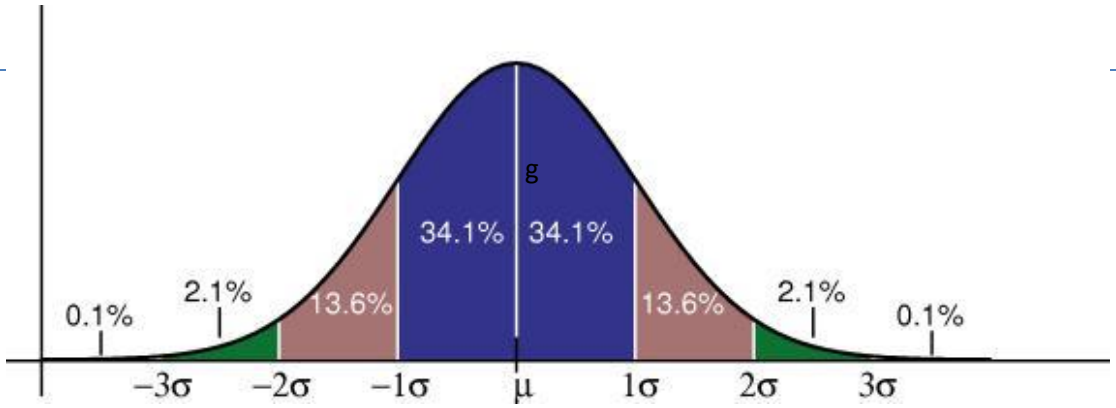
III. Mô hình hóa người nói sử dụng Gaussian Mixture Model và nhận dạng người nói

1. Phân phối Gaussian và gaussian mixture model

Phân phối chuẩn – hay còn gọi là phân phối gaussian là một phân phối quan trọng thường gặp trong đời sống và trong kỹ thuật. Phương trình mật độ xác suất của phân phối này như sau:

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

Trong đó là trung bình (hay kỳ vọng), là là độ lệch chuẩn. Phân phối xác suất có dạng như hình chuông:

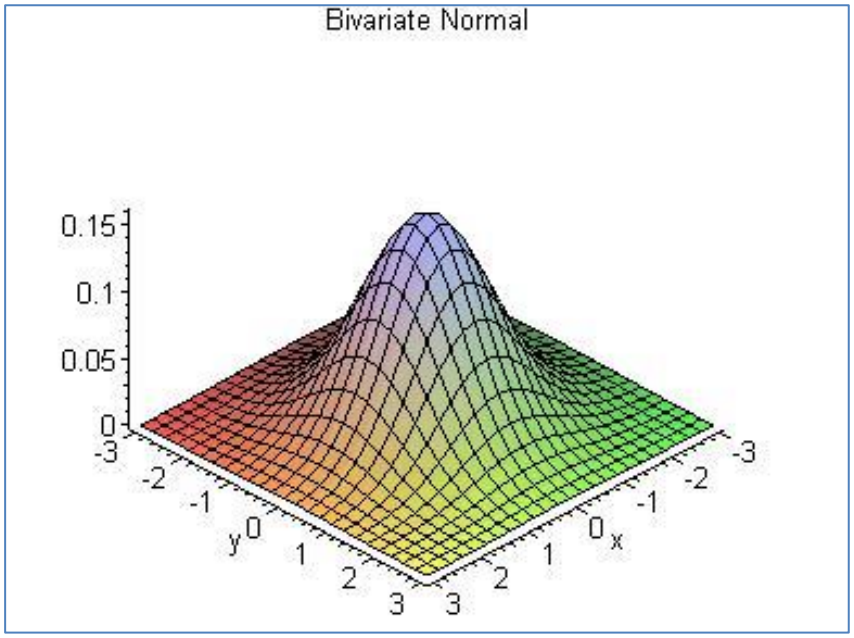


Hình 15: Phân phối mật độ xác suất của phân phối chuẩn

Với hàm nhiều biến, phương trình mật độ xác suất của gaussian như sau:

$$p(x,\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_i)'.\Sigma^{-1}(x-\mu_i)\right)$$

Với x là một vector, μ là vector kỳ vọng, Σ là ma trận hiệp phương sai, N là kích thước của vector x.

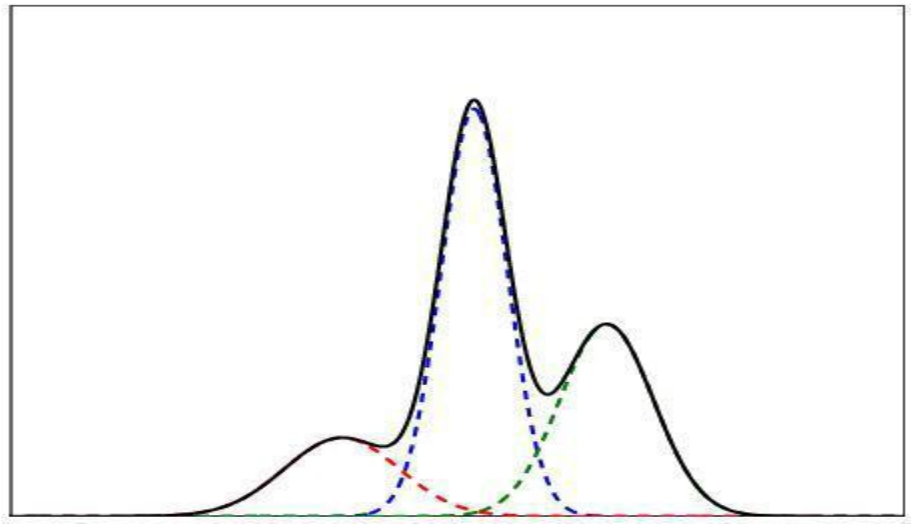


Hình 16: Phân phối chuẩn 2 biến

Mô hình trộn gaussian (gaussian mixture model) là tổng có trọng số của nhiều thành phần phân phối gaussian cơ sở, cụ thể như sau:

$$p(x) = \sum_{i=1}^{M} p_i . b_i(x)$$

Với pi là trọng số của thành phần thứ i, bi(x) là mật độ xác suất của thành phần thứ i với x, M là tổng số thành phần. Tổng của pi bằng 1.



Hình 17: Mô hình trộn gaussian

2. *Mô hình hóa người nói bằng Gaussian Mixture Model*

Có hai nguyên nhân chính khiến cho gaussian mixture model được sử dụng cho mô hình hóa người nói. Người ta thấy rằng tiếng nói cũng được tạo thành từ nhiều lớp âm thanh khác nhau, được tạo thành khi đi qua lưỡi, thanh quản, miệng tạo thành nguyên âm, phụ âm, hơi khác nhau. Mặt khác, việc sử dụng gaussian mixture model cho phép biểu diễn được số lượng rất lớn những mô hình phân phối khác nhau tương ứng với những người nói khác nhau. Do đó, GMM có thể được sử dụng để mô hình hóa các người nói khác nhau.

Việc xây dựng mô hình người nói được dựa trên các vectors MFCCs được lấy từ giai đoạn rút trích đặc trưng. Phương pháp thường đượng sử dụng đó là phương pháp maximum likelihood nhằm tìm những hệ số của mô hình gaussian sao cho xác suất của các vector huấn luyện là cao nhất. Cụ thể, likelihood có thể viết dưới dạng:

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda)$$

Với X = {x1, x2…xT) là các vector huấn luyện, λ là mô hình cần tìm.

Tuy nhiên, hàm trên là một hàm phi tuyến và không thể maximize nó một cách trực tiếp được, thay vào đó, người ta xử dụng thuật toán Expectation – Maximization (EM) lặp lại tuần tự để tìm mô hình tối ưu.

Chi tiết thuật toán: Ban đầu khởi tạo một mô hình với các hệ số ngẫu nhiên. Sau mỗi lần lặp, ước lượng lại các hệ số sau:

Trọng số

$$p_i = \frac{1}{T}\sum_{t=1}^{T} p(i|x_t, \lambda)$$

Kỳ vọng

$$\mu_i = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda)x_t}{\sum_{t=1}^{T} p(i|x_t, \lambda)}$$

Phương sai

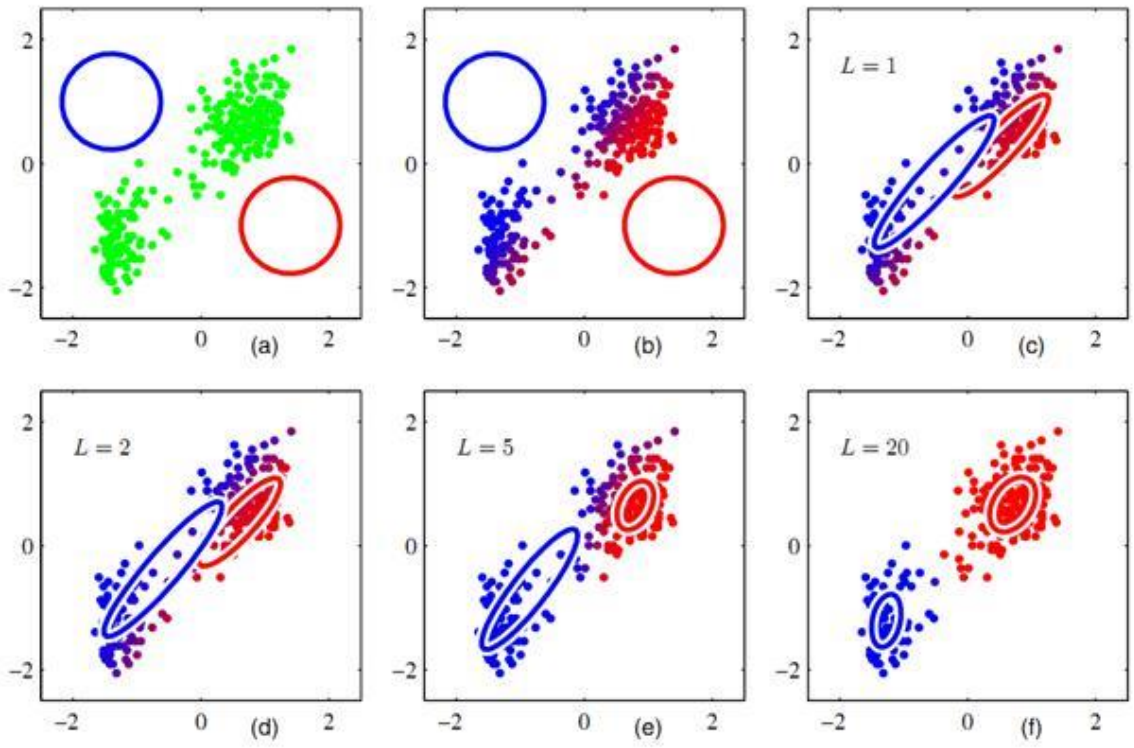
$$\sigma_i^2 = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda)x_t}{\sum_{t=1}^{T} p(i|x_t, \lambda)} - \mu_i^2$$

Với

$$p(i|x_t, \lambda) = \frac{p_i b_i(x_t)}{\sum_{k=1}^{M} p_k b_k(x_t)}$$

Trong đó, M là số mô hình gaussian cơ sở. Theo tác giả thuật toán, số M vào khoảng 20 – 32 đem lại kết quả tốt nhất.

Nhận diện người nói



Hình 18: Mô tả cách thức hoạt động của EM

3. *Nhận diện người nói*

Sau khi đã có được mô hình người nói, ta có thể nhận diện người nói với dữ liệu mới ban đầu. Dữ liệu mới sẽ được qua tiền xử lý, rút trích đặc trưng MFCC và đưa vào so khớp với các mô hình được lưu trong cơ sở dữ liệu.

Giả sử tập người nói gồm S người được biểu diễn bởi S mô hình GMM $\lambda 1$, $\lambda 2$,… λS. Mục tiêu là tìm mô hình cho xác suất tiền nhiệm cao nhất với một dữ liệu đầu vào mới thêm vào, cụ thể:

$$\hat{S} = argmax_{1 \le k \le S} \Pr(\lambda_k | X) = argmax_{1 \le k \le S} \frac{p(X|\lambda_k)\Pr(\lambda_k)}{p(X)}$$

Theo luật Bayes. Giả sử xác suất của người nói Pr (λk) đều bằng nhau, do xác suất p(X) như nhau với mọi mô hình người nói, công thức trên có thể đơn giản lại như sau:

$$\hat{S} = argmax_{1 \le k \le S} p(X|\lambda_k)$$

Trong thực tế với nhiều vector đặc trưng MFCC được rút trích từ một mẫu âm thanh ban đầu, hệ thống nhận diện người nói thực hiện tính như sau:

$$\hat{S} = argmax_{1 \le k \le S} \sum_{t=1}^{T} logp(x_t|\lambda_k)$$

IV. Thực nghiệm

1. Mô tả dữ liệu

- Dữ liệu được lấy từ bộ dữ liệu VIOS Corpus từ phòng thí nghiệm trí tuệ nhân tạo AILab của ĐH Khoa học Tự Nhiên TP.HCM. Bộ dữ liệu bao gồm 15 giờ thu gồm 46 bộ tương ứng với 46 người nói. Mỗi bộ dữ liệu sẽ chứa từ 250 tới 300 files âm thanh với chiều dài từ 2s-5s.
- Với dữ liệu huấn luyện: ta cắt bỏ phần silence và nối 80 file .wav đầu tiên của mỗi người nói để tạo thành một file âm thanh có độ dài khoảng 3-5 phút cho 1 người nói. Ta sẽ chuẩn bị dữ liệu như vậy cho 20 người nói đầu tiên từ SPK01 đến SPK20 (như vậy sẽ có 20 file .wav, mỗi file có chiều dài 3-5 phút, tương ứng với 20 người nói).
- Với dữ liệu kiểm tra: ta sẽ lấy 20 file .wav của mỗi người nói cắt bỏ silence để kiểm tra, và sẽ lấy của 20 người tương ứng ở trên, vậy tổng số file kiểm tra sẽ có là 20file * 20 người = 400 file test.

2. Mô tả thí nghiệm

Chương trình để thực hiện thí nghiệm: Matlab

_ Thực hiện huấn luyện (training):

- Đọc vào bộ dữ liệu huấn luyện.
- Rút đặc trưng của người nói sử dụng MFCC (Mel Frequency Cepstral Coefficient).
- Khởi tạo k phân phối Gauss bất kỳ.
- Dùng thuật toán gom cụm k-mean để ước lượng tạm thời các tham số của k phân phối Gauss này.

- Sử dụng thuật toán EM (Expectation Maximization) để tối đa hóa kỳ vọng trong n bước.
- Thu được mô hình nhận dạng cho người nói.

_ Thực hiện kiểm thử (testing):

- Dùng mô hình nhận dạng người nói được tạo ra ở bước huấn luyện để nhận dạng người nói trong bộ dữ liệu kiểm thử.
- Thực hiện thay đổi các tham số về kích thước bộ dữ liệu huấn luyện, số mô hình Gauss, số vòng lặp thuật toán EM và đánh giá độ chính xác cùng thời gian thực hiện của quá trình nhận dạng.

3. Source Code

```
global filesTrain;
global filesTest;
f_arr=[];
for k=1:length(filesTrain)

    path = filesTrain(k, 1)
    strPath = strjoin(path);
    a = audioread(strPath);
    b = mfcc(a);
    g1_0 = gNew(12, 16, 'diag');
    g2_0 = gInit(g1_0, b, 100);
    g3_0 = gRE(g2_0, b, 100);
    f_arr =[f_arr; g3_0];
end

strResult = '';
dem = 0;
RealValue = 1;
n = 1;

for p=1:length(filesTest)

    pathTest = filesTest(p, 1);
    strPathTest = strjoin(pathTest)
    x_w = audioread(strPathTest);
    x = mfcc(x_w);
    max = 0;
    index = 1;
    for k=1:length(f_arr)
      GMM = mean(gPr(f_arr(k), x))
      if GMM > max
          max = GMM;
          index = k;
      end
    end

    if index == RealValue
       dem = dem + 1;
    end

    max
    result = strcat(strcat('Redicted value: ',num2str(index)), strcat('  -  Real value: ', num2str(RealValue)));
    strResult = sprintf('%s\n%s',strResult, result);

    n = n + 1;
    if n > 20
       RealValue = RealValue + 1;
       n = 1;
    end
end
ratio = strcat('Ratio:  ', num2str((double(dem)*100)/length(filesTest)));
ratio = strcat(ratio, '%');
set(handles.txtRatio, 'String', ratio);
set(handles.editResult, 'String', strResult);
```

4. Mô tả kết quả

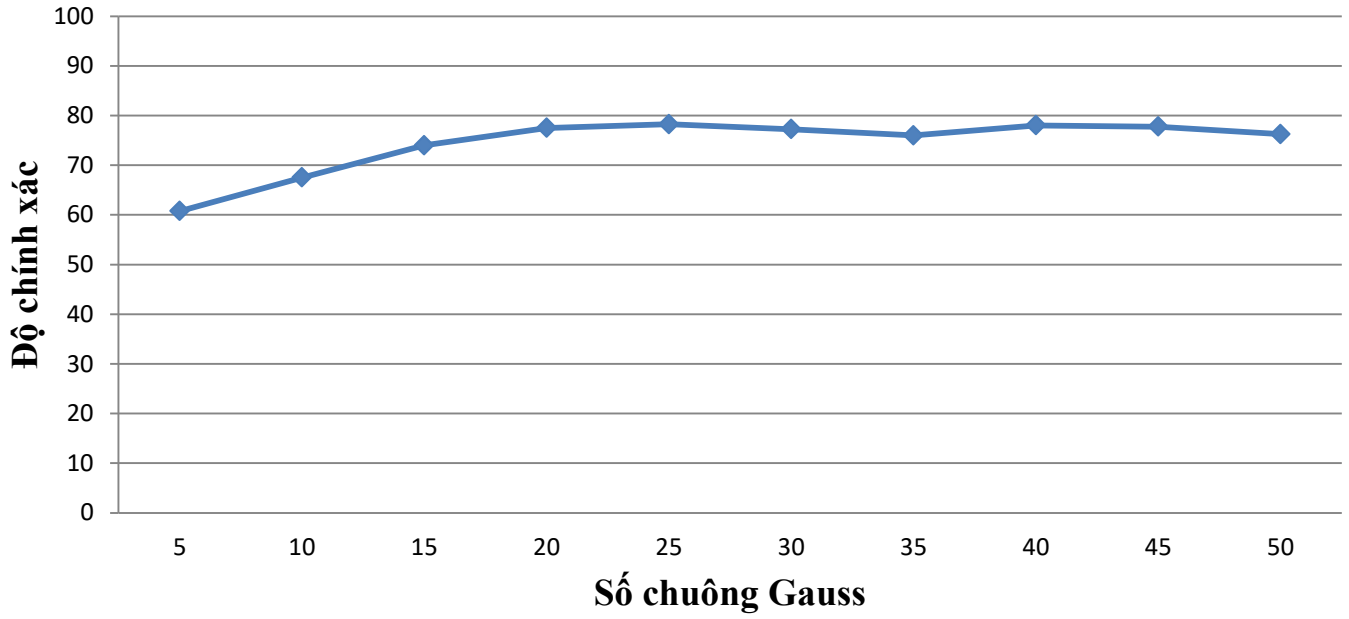
_ Khối lượng huấn luyện: 5 phút

_ Số vòng lặp K: 100

_ Số chiều MFCC: 12

_ Tham số chạy: số chuông Gauss

Số chuông Gauss	5	10	15	20	25	30	35	40	45	50
Độ chính xác	60.75%	67.5%	74%	77.5%	78.25%	77.25%	76%	78%	77.75%	76.25%



Nhận xét

Khi tăng số chuông Gauss lên thì độ chính xác của quá trình nhận dạng cũng tăng theo, nhưng đến một ngưỡng nào đó thì độ chính xác của quá trình nhận dạng sẽ bão hòa.

V. Kết luận

Nhận diện người nói có nhiều ứng dụng trong thực tế cuộc sống. Nhận diện người nói là một bài toán đã được nghiên cứu từ rất lâu và có nhiều thuật toán được sử dụng trong quá trình nhận dạng người nói.

Phương pháp nhận diện người nói sử dụng đặc trưng MFCC và mô hình hóa sử dụng GMM đem lại kết quả tương đối ổn định với độ chính xác cao, tuy nhiên độ chính xác dễ bị ảnh hưởng bởi chất lượng đầu thu và nhiễu. Do đó, quá trình tiền xử lý đóng vai trò rất quan trọng đến độ chính xác của thuật toán.

❖ Tài liệu tham khảo:

1. Anil K. Jain, Patrick Flynn, Arun A. Ross: **Handbooks of Biometric**, chapter 8: Voice Biometrics.
2. Evgeny Karpov: **Real-Time Speaker Identification**, Master's Thesis at University of Joensuu.
3. Ling Feng: **Speaker Recognition**, Master's Thesis at Technical University of Denmark.
4. Phạm Minh Nhựt: **Định danh người nói độc lập văn bản bằng mô hình thống kê**, Luận văn thạc sĩ tại Đại học Khoa học tự nhiên – Đại học Quốc Gia TPHCM.
5. Kishore Prahallad: Speech Technology Course's slides at CMU.
6. Douglas Reynolds, Richard Rose: **Robust text-independent Speaker Identification using Gaussian mixture models,** IEEE Transactions on Speech and Audio Processing, Vol 3, No. 1, 1995
7. https://drive.google.com/file/d/0BwFOqsjqEVKYU2ZzeTFzVDdWTlU/view