NHẬN DIỆN NGƯỜI NÓI SỬ DỤNG MFCC VÀ GMM

NHÓM

Mục lục

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.

Bài 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, bài 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.

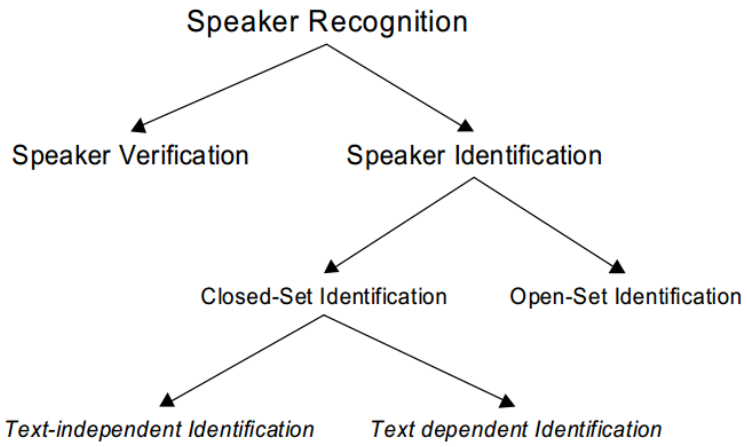
I. Giới thiệu chung

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 diện 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.

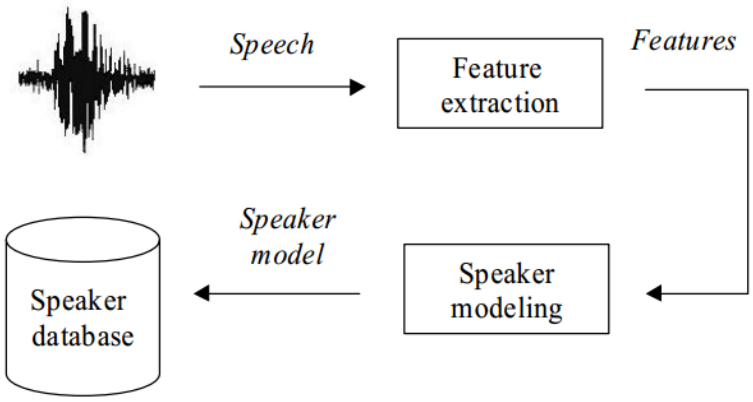
- Hệ thống nhận dạng người nói còn có thể dùng để sử dụng trong điều tra hình sự.

4. Phương pháp được 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).

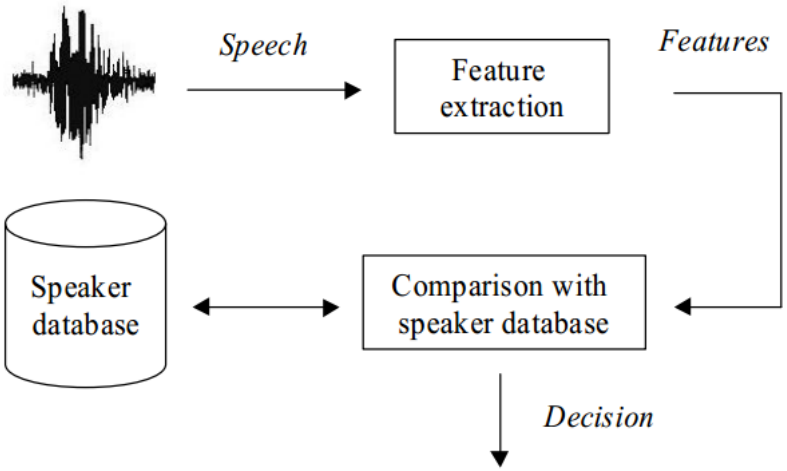
II. Hệ thống nhận diện người nói

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



Hình 2: Sơ đồ pha đăng ký người nói

- 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.



Hình 3: Sơ đồ pha nhận diện 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. Phần tiếp theo của bài báo cáo sẽ trình bày các ý chính trong thuật toán rút trích đặc trưng và mô hình hóa.

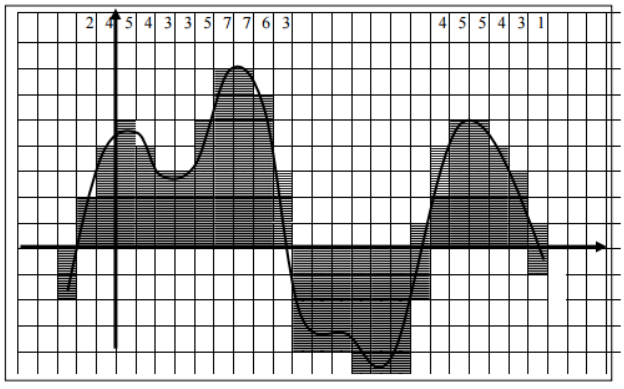
III. Rút trích đặc trưng tiếng nói sử dụng đặc trưng MFCC

1. Tổng quan về mã hóa âm thanh, tiếng nói và đặc trưng của tiếng nói

a. Mã hóa âm thanh

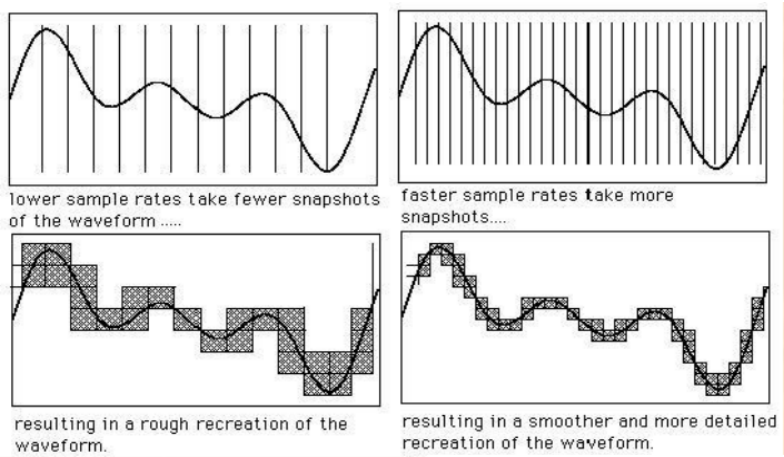
Có nhiều phương pháp mã hoá âm thanh.

Cách đơn giản nhất là mã hoá bằng cách xấp xỉ dao động sóng âm bằng một chuỗi các byte thể hiện biên độ dao dộng tương ứng theo từng khoảng thời gian bằng nhau. Các đơn vị thời gian này cần phải đủ nhỏ để không làm "nghèo" âm thanh. Đơn vị thời gian này gọi là tần số lấy mẫu (sample rate). Giá trị tại mỗi lần lấy mẫu được biểu diễn trong một miền giá trị xác định được gọi là độ sâu số (bit depth). Khi phát, một mạch điện sẽ khôi phục lại sóng âm với một sai lệch chấp nhận được.



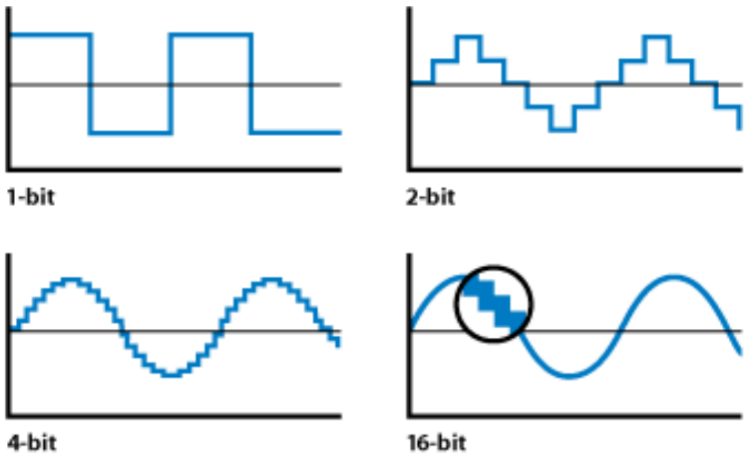
Hình 4: Số hóa tín hiệu âm thanh

Tần số lấy mẫu khác nhau:



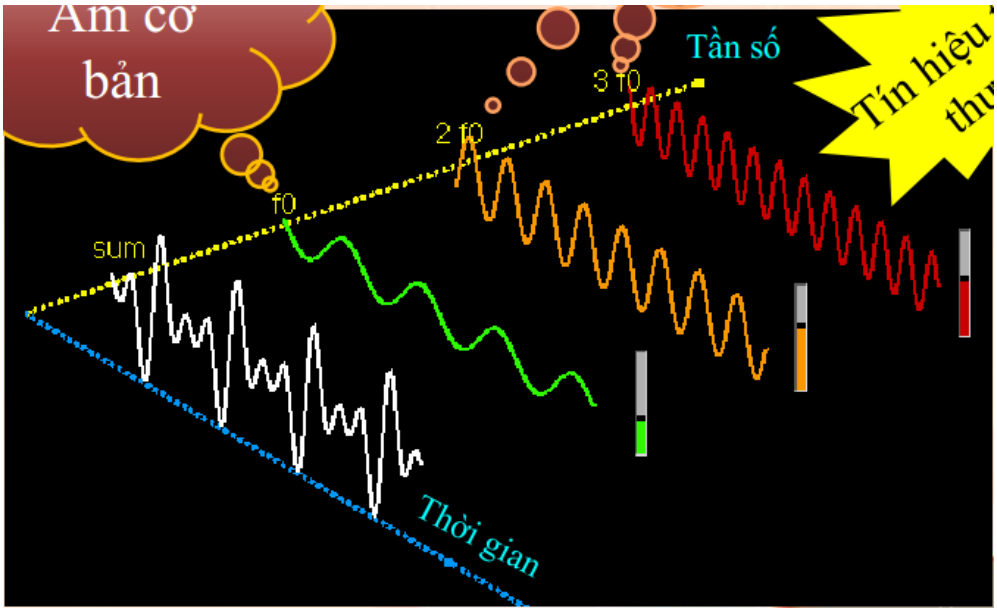
Hình 5:Các phương pháp lấy mẫu với tần số khác nhau

Độ sâu số khác nhau:



Hình 6: Lấy mẫu với độ sâu số (depth) khác nhau

Một cách khác là phân tích dao động âm thanh thành tổng các dao động điều hoà (các dao động hình sinh với tần số và biên độ khác nhau) và chỉ lưu lại các đặc trưng về tần số, và biên độ.



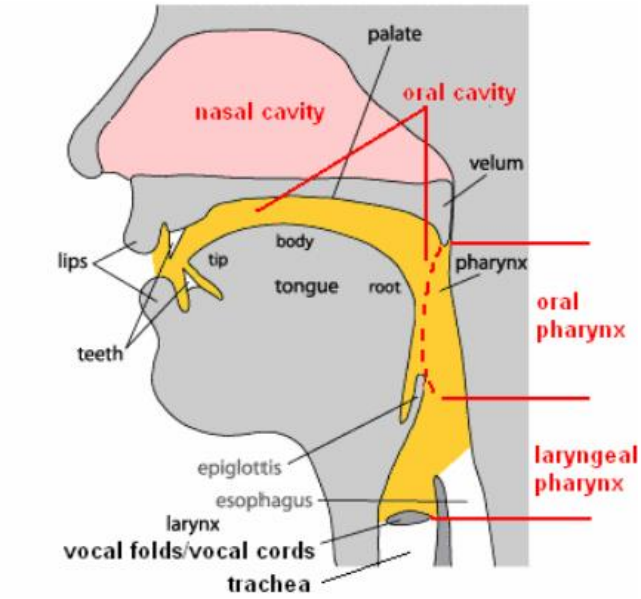
Hình 7: Chuyển tín hiệu miền thời gian thành tín hiệu tần số

b. Sự hình thành giọng nói

Giọng nói là âm thanh được thực hiện bởi một người bằng cách sử dụng những nếp gấp thanh quãng để nói chuyện, đọc sách, hát, cười, khóc, la hét vv Giọng nói là một phần của âm thanh mà con người có thể tạo ra, trong đó các nếp gấp thanh quãng (dây thanh âm) là nguồn âm thanh chính.

Các bộ phận để tạo ra tiếng nói của con người có thể được chia thành ba phần; phổi, các nếp gấp trong thanh quản, và các bộ phận cấu âm.

Phổi bơm phải đầy đủ luồng không khí và áp suất không khí làm rung thanh quản. Các nếp gấp thanh quản (dây thanh âm) là một van rung băm nhỏ các luồng không khí từ phổi vào các xung âm thanh đã hình thành các nguồn âm thanh quản. Các cơ bắp của thanh quản điều chỉnh độ dài và căng của những nếp gấp thanh âm để điều chỉnh cao độ và âm sắc. Các bộ phận cấu âm (gồm lưỡi, vòm miệng, má, môi, vv) lọc những âm thanh phát ra từ thanh quản và đến mức độ nào đó có thể tương tác với các luồng không khí thanh quản để tăng cường hoặc suy yếu nó như một nguồn âm thanh. Các nếp gấp thanh quãng kết hợp với các bộ phận cấu âm, có khả năng tạo ra các các âm rất phức tạp.



Hình 8: Các bộ phận tạo thành tiếng nói

c. Rút trích thông tin từ tiếng nói.

Tìn hiệu tiếng nói bao gồm rất nhiều loại thông tin khác nhau về người nói. Thông tin này bao gồm các thông tin "cấp cao" như hệ ngôn ngữ, ngữ cảnh, phong cách nói, tình trạng cảm xúc v.v... Việc sử dụng các thông tin cấp cao này vào việc sử dụng để nhận diện người nói đã được nghiên cứu khá nhiều nhưng rất khó để thực hiện và không thể ứng dụng trong thực tế. Thay vào đó, các thông tin cấp thấp như cao độ (pitch), cường độ, tần số, băng tần, phổ âm thanh v.v.. được sử dụng và áp dụng thành công hơn.

Thông tin được lưu trữ trong tiếng nói rất nhiều, tuy nhiên ta chỉ cần rút trích lượng thông tin vừa đủ để phân biệt giữa những người nói với nhau. Quá trình rút trích lượng thông tin này được gọi là quá trình rút trích đặc trưng người nói.

Dựa trên những phân tích trên, đặc trưng của người nói cần có những đặc điểm sau:

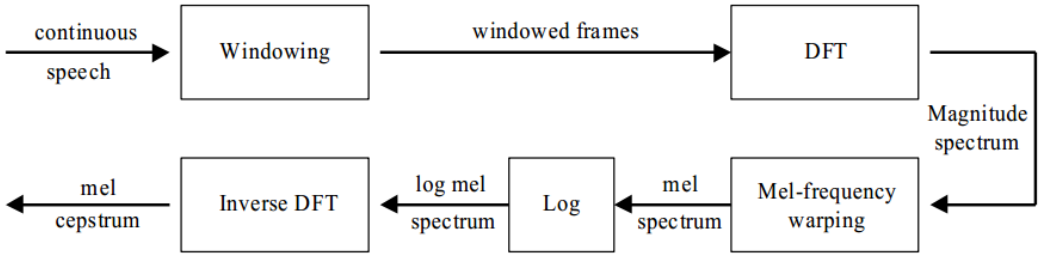
Có khả năng phân biệt giữa các người nói khác nhau nhưng đồng thời cũng không quá nhạy với những thay đổi nhỏ trong cách nói chuyện của cùng một người nói.

- Có thể đo đạc được độ chính xác.
- Ổn định qua thời gian.
- Biểu hiện một cách tự nhiên và thường xuyên trong tiếng nói.
- Thay đổi ít khi thay đổi môi trường thu âm.
- Không dễ bị đánh lừa bởi những người bắt chước.

Tuy nhiên, rất khó để có thể rút trích các đặc trưng thỏa mãn tất cả các tính chất trên. Thông thường, các đặc trưng này được rút trích dựa trên phổ âm thanh. Phần tiếp theo sẽ trình bày phương pháp MFCC.

2. Rút trích đặc trưng Mel-Frequency Cepstrum Coefficients

Mel-Frequency Cepstrum Coefficients là đặc trưng thường được dùng để diễn tả âm thanh tiếng nói. Nó dựa trên quan sát đó là thông tin được mang bởi các thành phần có tần số thấp thường quan trọng hơn các âm thanh có tần số cao – do tiếng nói con người biến đổi chậm. Các bước để rút trích đặc trưng này như sau:



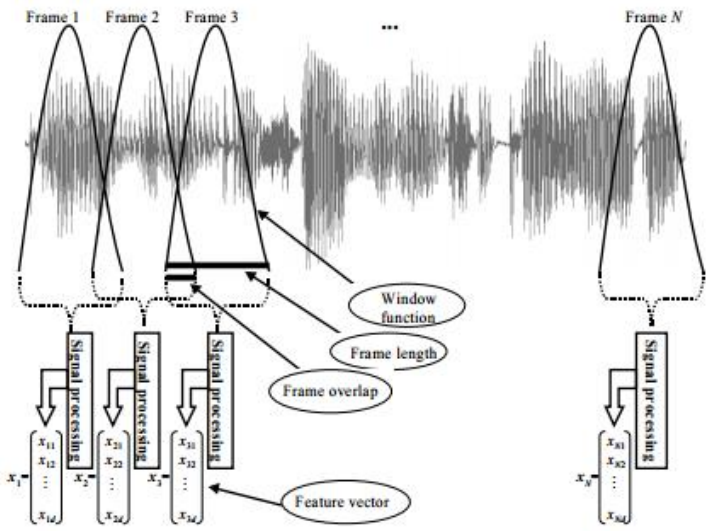
Hình 9: Mô hình các bước rút trích đặc trưng MFCC

- Bước thứ nhất đó là windowing, chia tín hiệu âm thanh ban đầu thành các frame liên tiếp nhau. Mỗi frame này sẽ được đưa vào và rút trích đặc trưng MFCC tương ứng.
- Bước thứ hai đó là biến đổi fourier rời rạc. Bước này nhằm chuyển đổi tín hiệu ban đầu thành tổ hợp của các sinusoid tương ứng với từng tần số khác nhau.
- Bước thứ ba là bước chuyển tín hiệu ở dạng tần số thu được ở bước hai sang một vùng tần số theo cảm nhận của tai người.
- Bước thứ tư là lấy log để tách tín hiệu tần số thấp và tần số cao thành 2 vùng khác nhau.
- Bước thứ năm thực hiện phép biến đổi fourier đảo, ta thu được đặc trưng MFCC.

Chi tiết của các bước như sau

a. Windowing

Tiếng nói trên thực tế thường biến đổi chậm, do đó nếu thực hiện phân tích trên một khoảng thời gian đủ ngắn (20 – 30 ms) thì những đặc trưng âm thanh của tiếng nói tương đối ổn định. Việc rút trích đặc trưng trên những khoảng thời gian này nhiều khả năng sẽ diễn tả được đặc trưng của người nói. Quá trình này được gọi là short-term analysis.

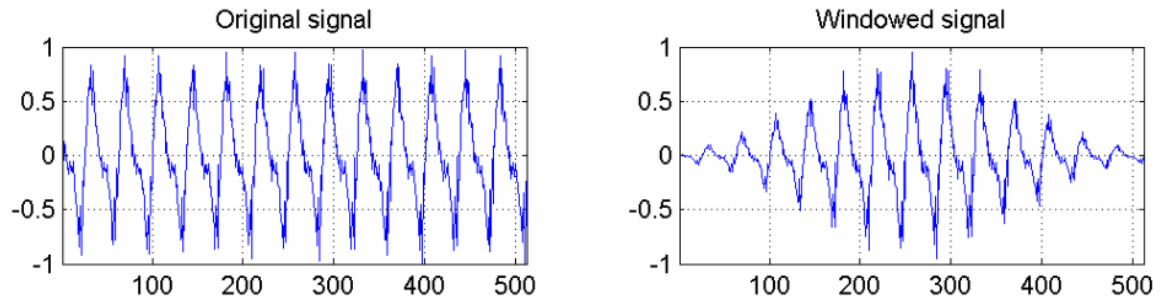


Hình 10: Quá trình framing

Tín hiệu nói ban đầu được chuyển thành các frame có kích thước cố định (20-30 ms), mỗi window sẽ có một phần chồng lên nhau (30 – 50%) với các frame cạnh nó nhằm tránh mất mát thông tin. Để tránh biến đổi đột ngội ở cuối frame, mỗi frame thường được nhân với một hàm window (window function), mà phổ biến nhất là hamming window function:

$$w(t) = 0.54 - 0.64cos\left(\frac{2t\pi}{N-1}\right)$$

Với N là kích thước của frame. Kết quả thu được sẽ được lần lượt đưa và quá trình rút trích đặc trưng.

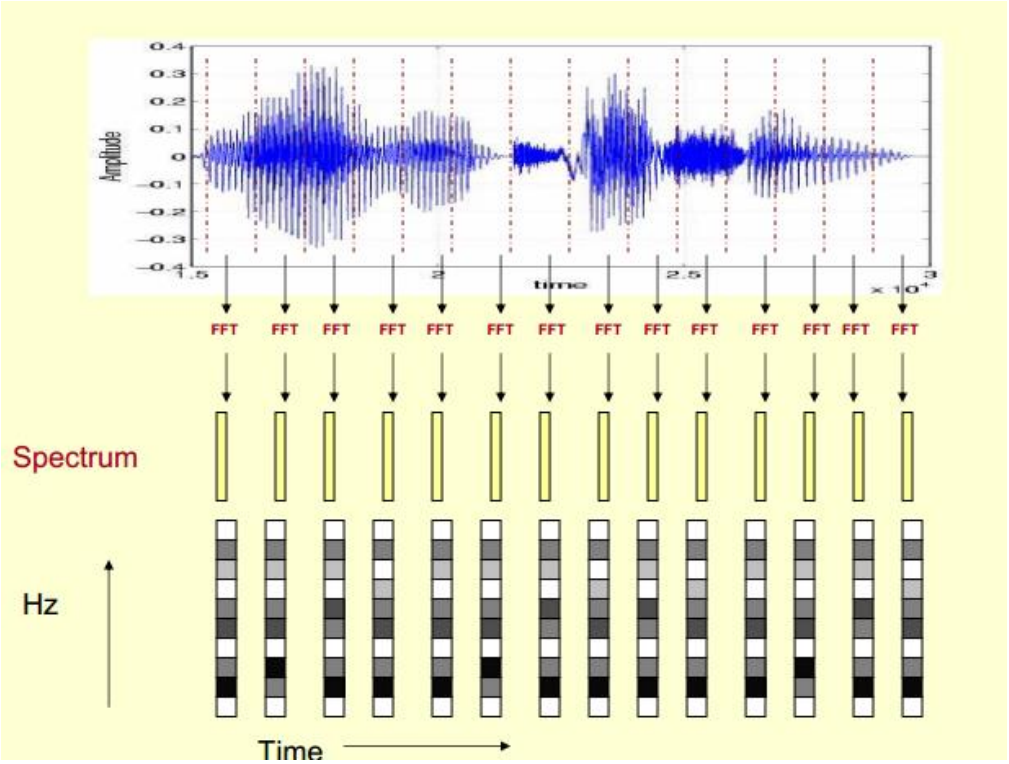


Hình 11: Tín hiệu trước và sau khi áp dụng hàm window

b. Discrete fourier transform

Mỗi frame thu được sau quá trình xử lý sẽ được đưa vào phép biến đổi fourier rời rạc để chuyển từ miền thời gian sang miền tần số. Sau khi thực hiện biến đổi này, ta thu được một biểu diễn khác

của âm thanh được gọi là cepstrum. Biểu diễn này thể hiện tốt hơn sự biến đổi và đặc trưng tiếng nói nằm trong âm thanh.



Hình 12: Cepstrum

Biến đổi Fourier hay chuyển hóa Fourier, được đặt tên theo nhà toán học người Pháp Joseph Fourier, là một biến đổi tích phân dùng để khai triển một hàm số theo các hàm số sin cơ sở, có nghĩa là dưới dạng tổng hay một tích phân của các hàm số sin được nhân với các hằng số khác nhau (hay còn gọi là biên độ). Biến đổi Fourier có rất nhiều dạng khác nhau, chúng phụ thuộc vào dạng của hàm được khai triển.

Trong toán học, phép biến đổi Fourier rời rạc (DFT), đôi khi còn được gọi là biến đổi Fourier hữu hạn, là một biến đổi trong giải tích Fourier cho các tín hiệu thời gian rời rạc. Đầu vào của biến đổi này là một chuỗi hữu hạn các số thực hoặc số phức, làm biến đổi này là một công cụ lý tưởng để xử lý thông tin trên các máy tính. Đặc biệt, biến đổi này được sử dụng rộng rãi trong xử lý tín hiệu và các ngành liên quan đến phân tích tần số chứa trong trong một tín hiệu, để giải phương trình đạo hàm riêng, và để làm các phép như tích chập. Biến đổi này có thể được tính nhanh bởi thuật toán biến đổi Fourier nhanh (FFT).

Một biến đổi Fourier nhanh (FFT) là một thuật toán hiệu quả để tính biến đổi Fourier rời rạc (DFT) và biến đổi ngược. Khi cài đặt thực tế, ta sử dụng phép FFT này lên các frame, kết quả sẽ được chuyển qua bước tiếp theo, đó là lọc Mel-frequency.

c. Lọc mel-frequency

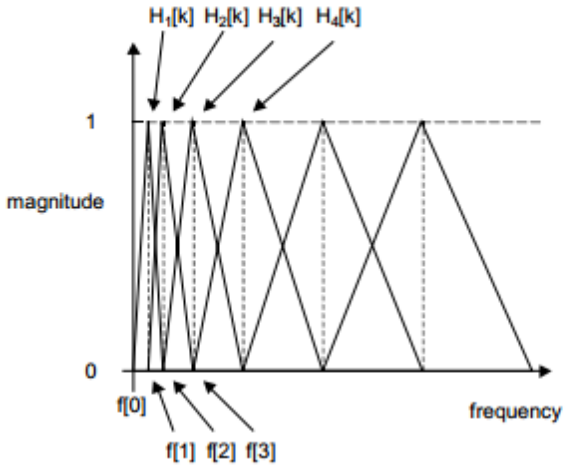
Sau bước biến đổi DFT, ta thu được thông tin về tín hiệu dưới dạng tần số và cường độ. Tuy nhiên, tai người cảm nhận âm thanh không giống với những biến đổi vật lý của âm thanh. Do đó, người ta sử dụng một thang đo tần số khác được gọi là tần số mel, được đo theo cảm nhận của tai con người. Bảng chuyển đổi tần số vật lý sang thang đo mel như sau, người ta lấy 1000 Hz làm mốc chuyển đổi giữa hai thang đo:

Hz	40	161	200	404	693	867	1000	2022	3000	3393	4109	5526	6500	7743	12000
mel	43	257	300	514	771	928	1000	1542	2000	2142	2314	2600	2771	2914	3228

Người ta xây dựng nhiều công thức để chuyển từ Hz sang mel, trong đó phổ biến nhất là công thức của Lindsay và Loman:

$$m = 2410 \log_{10}(1.6 \times 10^{-3} f + 1)$$

Thông qua một bộ lọc, người ta có thể tính toán lại tần số và biên độ ở thang đo Hz sang thang đo mel, khi đó ta thu được một vector tần số và biên độ mới.



Hình 13: Bộ lọc mel

d. Lấy log và phép biến đổi fourier đảo

Tín hiệu tiếng nói của con người có thể được biểu diễn bởi hai thành phần là những thành phần biến đổi nhanh và vùng biến đổi chậm. Các đỉnh ở phổ âm thanh cùng với

Có thể biểu diễn sự tương quan của hai thông tin "nhanh" và "chậm" này như sau:

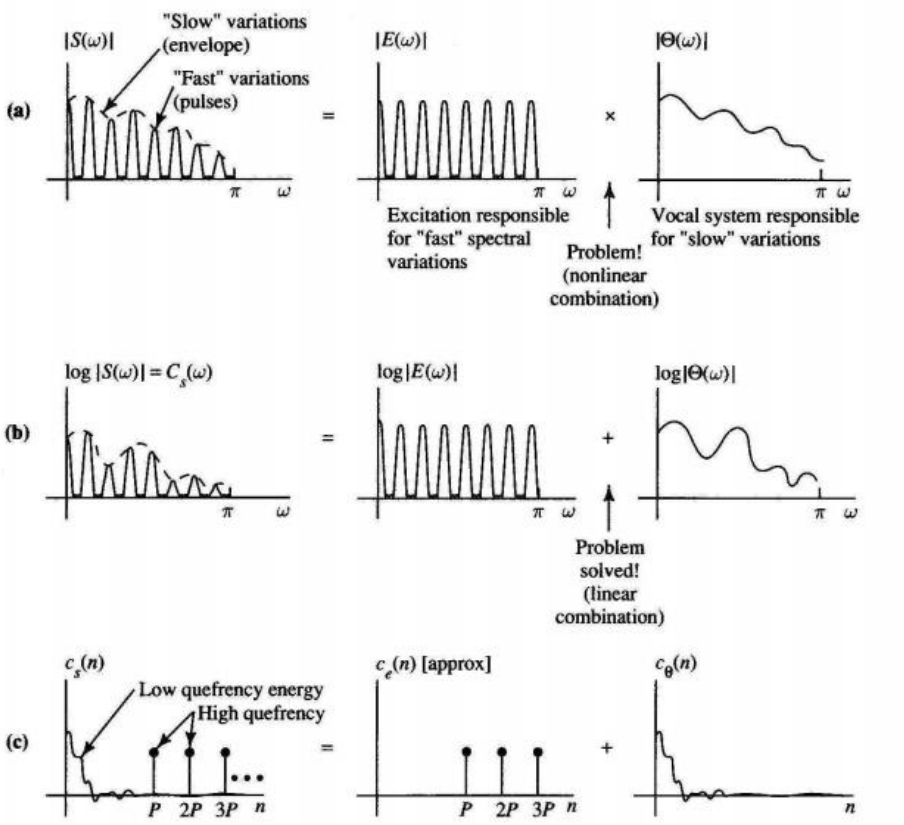
$$|S(x)| = |E(x).H(x)|$$

Trong đó E(x) là thành phần có tần số cao, H(x) là thành phần có tần số thấp, S(x) là tín hiệu gốc.

Khi thực hiện logarit trên biểu thức trên, ta có thể chuyển tổng thành tích như sau:

$$\log(|S(x)|) = \log(|E(x)|) + \log(|H(x)|)$$

Để thực hiện phân tách, người ta thực hiện một mẹo, đó là sử dụng phép biến đổi fourier trên chính log(|S(x)|) và phép biến đổi này được gọi là phép biến đổi fourier đảo. Từ kết quả của phép biến đổi này, ta có thể lọc ra hai vùng có tần số cao và thấp, vùng cần lấy là vùng có tần số thấp. Biểu diễn trực quan của cách làm này như sau:



Hình 14: IDFT và lọc kết quả để ra đặc trưng MFCC

Kết quả thu được sau toàn bộ quá trình này là đặc trưng Mel-frequency Cepstral Coefficients. Mỗi frame sẽ thu được một vector đặc trưng và các vector này sẽ được đưa vào quá trình mô hình hóa và nhận diện người nói.

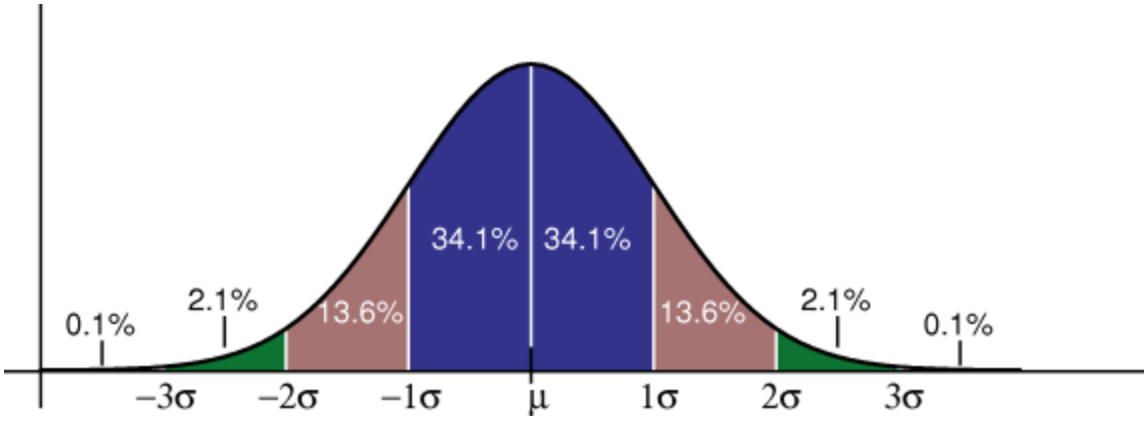
IV. Mô hình hóa người nói sử dụng gaussian mixture model và nhận diện người nói

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

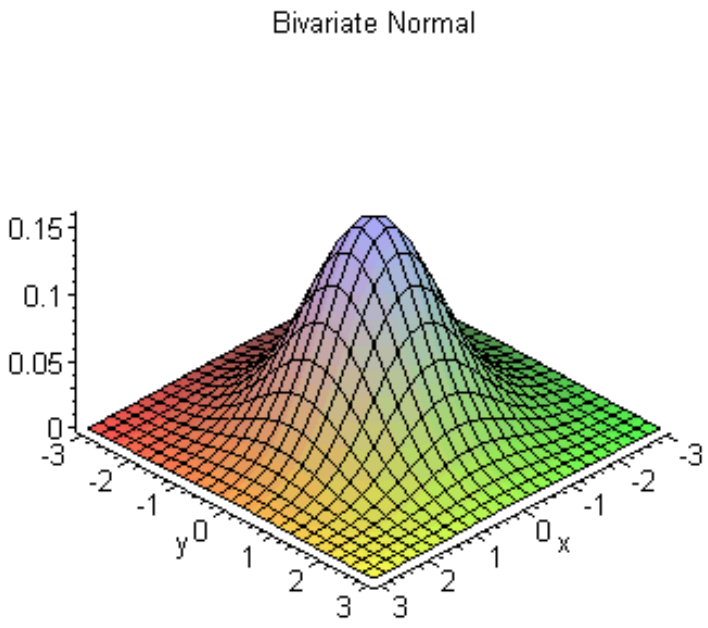


Hình 15: Phân phối mật độ xác suất của phân phối chuẩn

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.

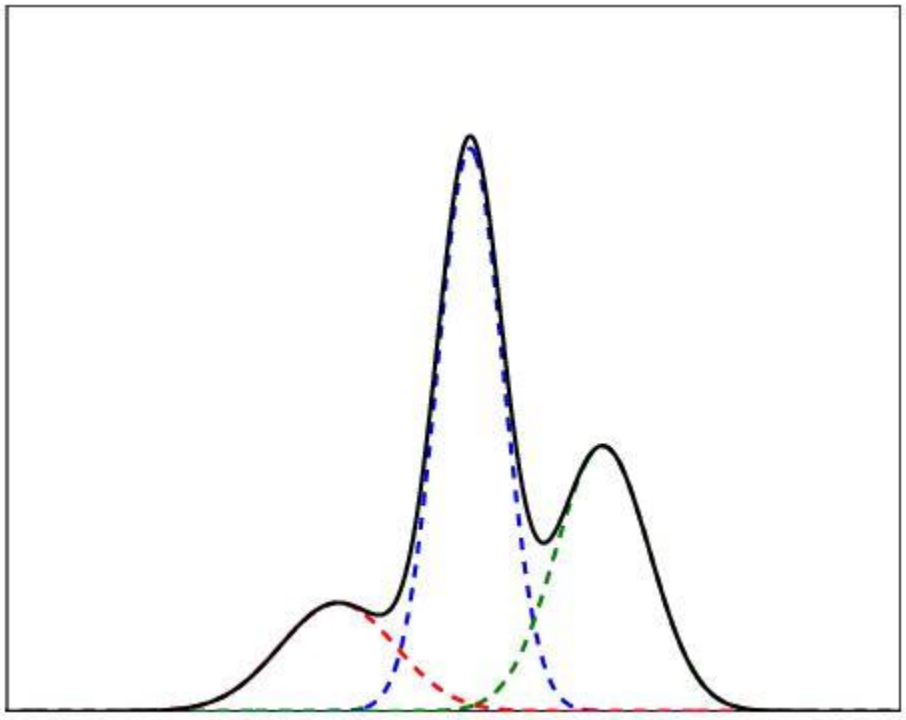


Hình 16: Phân phối chuẩn 2 biến

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 p_i là trọng số của thành phần thứ i, $b_i(x)$ là mật độ xác suất của thành phần thứ i với x, M là tổng số thành phần. Tổng của p_i 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 = {x_1, x_2…x_T) 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:

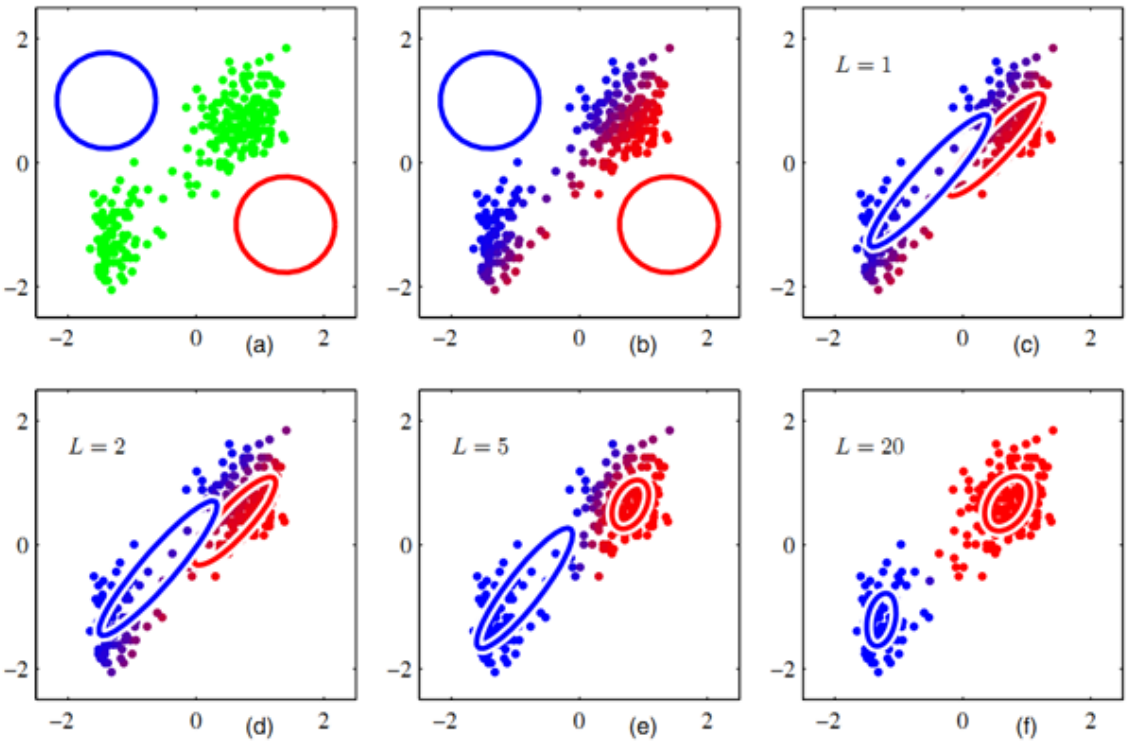
$$\mu_i = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda) x_t}{\sum_{t=1}^{T} p(i|x_t, \lambda)}$$

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.



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 $\lambda_1, \lambda_2, \ldots \lambda_S$. 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 $\Pr(\lambda_k)$ 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:

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

V. Thực nghiệm

1. Chương trình

Chương trình được xây dựng dựa trên đoạn mã nguồn rút trích đặc trưng MFCC của Kamil Wojcicki (http://www.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab/content/mfcc/mfcc.m) và mô hình hóa người nói dựa trên gaussian mixture model được cài đặt sẵn trong matlab.

File mã nguồn được viết trong "test.m".

Cú pháp:

```
test(nGaussianModels)
```

Trong đó:

- nGaussianModels là số lượng mô hình (25)

2. Kết quả thực nghiệm

Trong nghiên cứu gốc của tác giả [6] sử dụng 16 người trong cùng một bộ dữ liệu, độ chính xác thu được như sau:

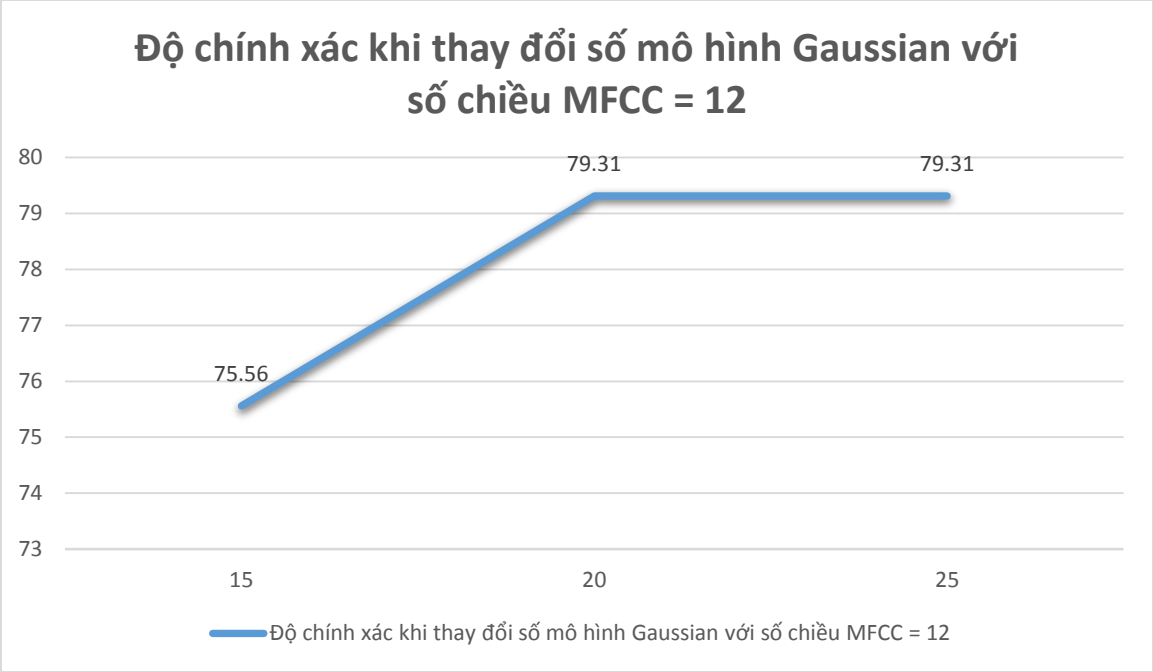
Phương pháp	Độ chính xác (%)
GMM	94.5
VQ-100	92.9
VQ-50	90.7
RBF	87.2
GC	67.1

Bảng 1: So sánh độ chính xác của GMM với một số phương pháp mô hình hóa khác

Bộ dữ liệu được nhóm sử dụng để huấn luyện và là bộ dữ liệu của Azarias Reda từ đại học Michigan. Bộ dữ liệu gồm 35 nam và 48 nữ, số lượng thực sự được sử dụng là 15 nam và 15 nữ với độ nhiễu khoảng 30%. 4 file đầu tiên dùng để huấn luyện, file cuối cùng dùng để kiểm thử.

http://www.azreda.org/audiodata/audio_data_umich.tar.gz

Khi thay đổi số lượng mô hình gaussian với số chiều MFCC là 12, độ chính xác của chương trình biến đổi như sau



Biểu đồ 1: Ảnh hưởng của số nhóm Gaussian đến độ chính xác

Có thể thấy được, số nhóm gaussian khoảng 20 là đạt được hiệu năng và độ chính xác tốt nhất. Tuy nhiên độ chính xác chỉ đạt 80% do nhóm chưa xử lý nhiễu và chất lượng tín hiệu âm thanh không cao.

Do giới hạn của hàm tính các tham số của gaussian mixture models trong matlab nên nhóm không thể tăng số chiều của đặc trưng MFCC lên được.

VI. 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