XÁC THỰC NGƯỜI NÓI KHÔNG PHỤ THUỘC VÀO TỪ KHÓA

(TEXT-INDEPENDENT SPEAKER VERIFICATION)

Trần Cao Trưởng

TÍNH CẤP THIẾT ĐỀ TÀI

- Lĩnh vực nhận dạng người nói có nhiều ứng dụng
  - Điều khiển truy nhập máy tính, thiết bị..
  - Nhận dạng tội phạm
  - Nhận thông tin phản hồi...
- Trên thế giới có nhiều công trình nghiên cứu vấn đề này nhưng ở Việt Nam mới có rất ít công trình nghiên cứu về lĩnh vực này
- Là hướng có thể mở rộng nghiên cứu

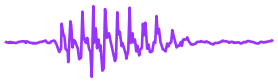
NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

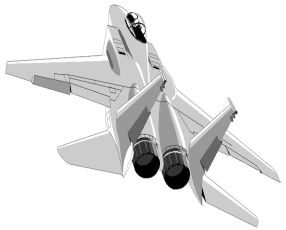
NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

TRÍCH CHỌN THÔNG TIN TỪ TIẾNG NÓI



Tín hiệu tiếng nói



Speech Recognition

Language Recognition

Speaker Recognition

Nội dung nói

Bạn có khỏe không?

Tên ngôn ngữ

Tiếng Việt

Tên người nói

Anh Trưởng

NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

NỘI DUNG

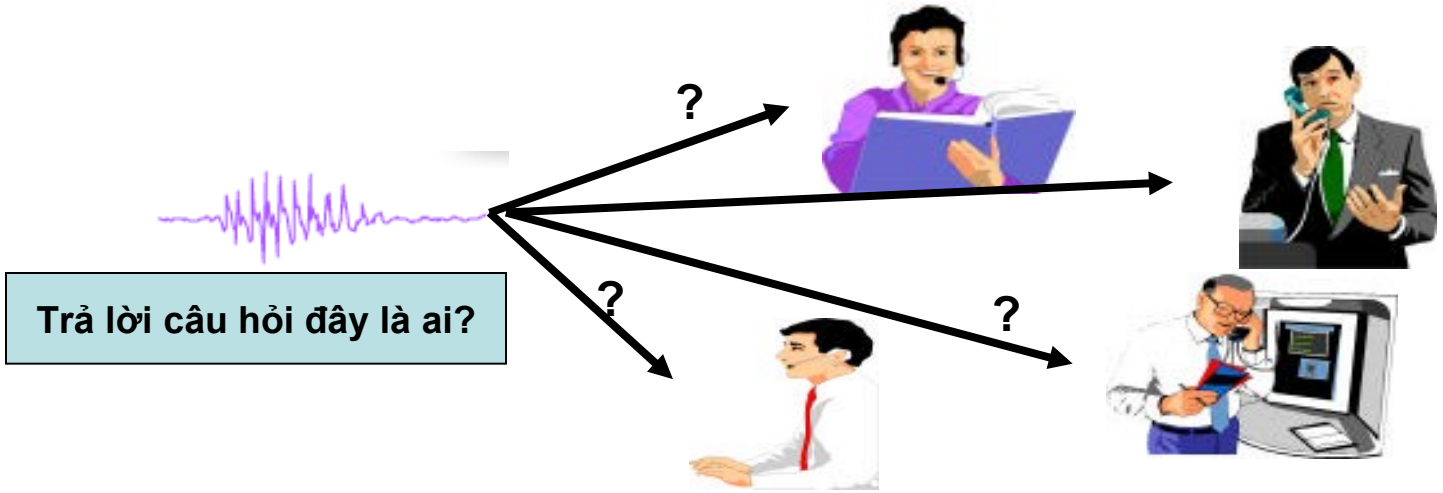
- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

CƠ SỞ LÝ THUYẾT

- Tiếng nói ngoài thông tin ngữ nghĩa còn chứa các thông tin như trạng thái tình cảm khi nói hay những thông tin riêng của giọng nói.
- Các thông tin này không bất biến từ lúc người biết nói đến lúc già, nhưng có tính khá ổn định trong giai đoạn dài của cuộc đời.
- Khi con người đã trưởng thành, những thói tật khi nói, những đặc trưng khu biệt trong cấu âm sẽ hình thành và mang tính ổn định cao.

PHÂN LOẠI THEO CHỨC NĂNG

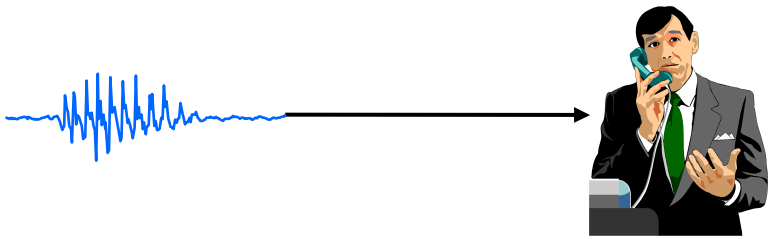
- Speaker Identification: Kiểm tra xem giọng nói cần kiểm tra là của ai trong hệ thống



PHÂN LOẠI THEO CHỨC NĂNG(tiếp)

- Speaker Verification: Xác thực xem ID người cần kiểm tra có chính xác là người đó hay là không?

Đây có phải là giọng nói của anh Trưởng không?



PHÂN LOẠI THEO PHƯƠNG PHÁP

- Nhận dạng phụ thuộc vào từ khóa(text-dependent)
  - Hệ thống nhận biết nội dung nói của người nói (mật khẩu)
  - Không mất thời gian huấn luyện
  - Độ chính xác cao. Nhưng bảo mật không cao do kẻ giả mạo ghi âm từ khóa
- Nhận dạng không phụ thuộc vào từ khóa(text-independent)
  - Hệ thống không quan tâm đến nội dung nói mà chỉ quan tâm đến giọng nói
  - Dữ liệu huấn liệu càng nhiều độ chính xác càng cao.

NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

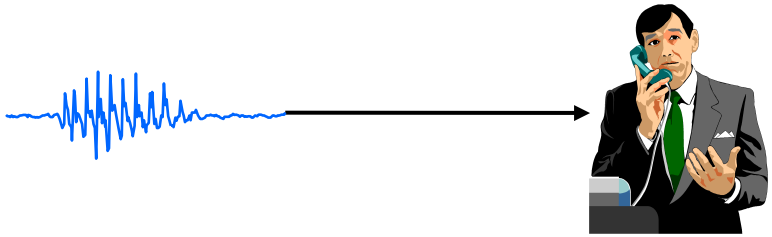
NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

MỤC ĐÍCH

- Xác minh liệu người đang nói có đúng là người mà máy tính đã được biết trước hay không (tính xác thật của giọng nói) nhưng không phụ thuộc vào nội dung người nói.

Đây có phải là giọng nói của anh Trưởng không?

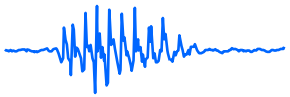


HAI GIAI ĐOẠN CỦA HỆ THỐNG

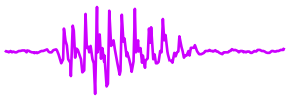
Huấn luyện

Ghi âm giọng nói của mỗi người

Đặc trưng của mỗi người nói



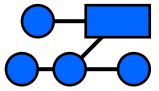
Trưởng



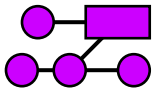
Quý

Trích chọn đặc trưng

Mô hình huấn luyện

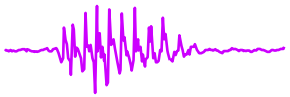


Truong



Quy

Xác thực



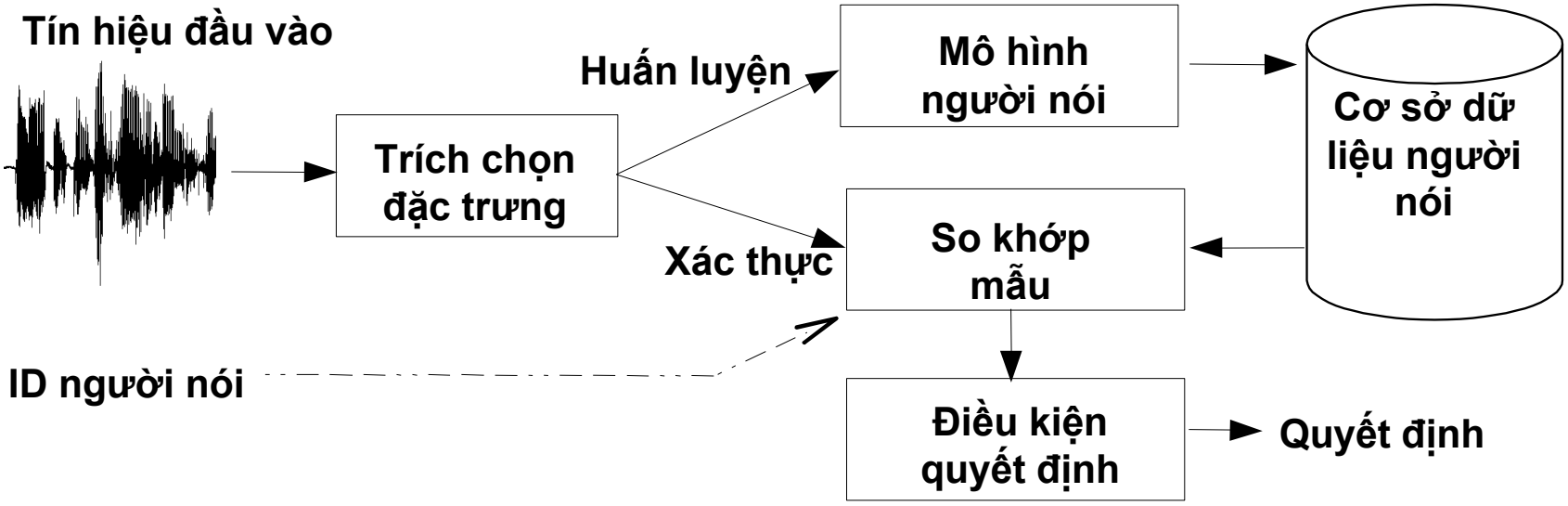
Trích chọn đặc trưng

Quyết định xác thực

Chấp nhận?

ID: truong

CÁC THÀNH PHẦN CỦA HỆ THỐNG



- Trích chọn đặc trưng: Biến đổi từ giọng nói thô thành nhưng đặc trưng của người nói.
- Mô hình người nói: Huấn luyện người nói dựa vào một phương pháp cụ thể.
- So khớp mẫu: Tính toán làm hợp tiếng nói đưa vào và một đặc trưng người nói trong cơ sở dữ liệu đã được huấn luyện.
- Điều kiện quyết định: Việc đưa ra quyết định dựa vào việc làm hợp ở trên.

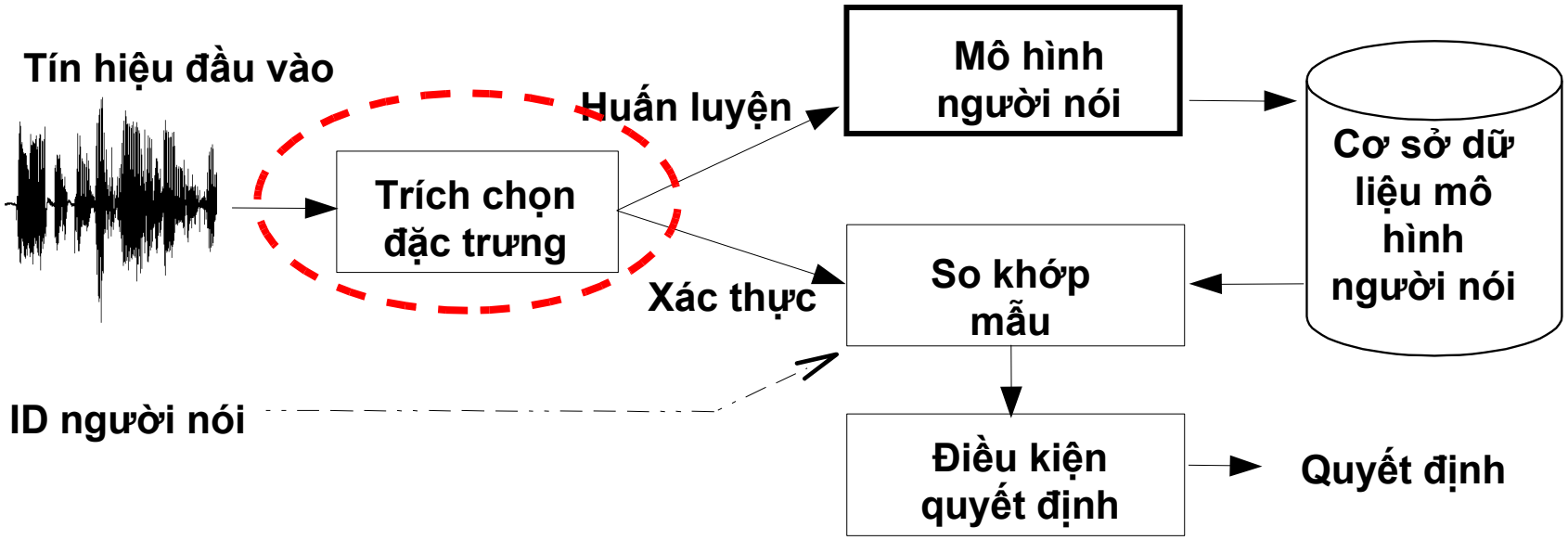
NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

TRÍCH CHỌN ĐẶC TRƯNG NGƯỜI NÓI



MONG MUỐN

- Xuất hiện một cách tự nhiên và liên tục trong khi nói.
- Ổn định đối với mỗi người nhưng phải khác nhau từ người này sang người khác.
- Ít bị thay đổi theo thời gian, sức khỏe hay trạng thái của người nói
- Ít bị ảnh hưởng bởi môi trường xung quanh (độc lập môi trường).
- Dễ dàng tính toán.

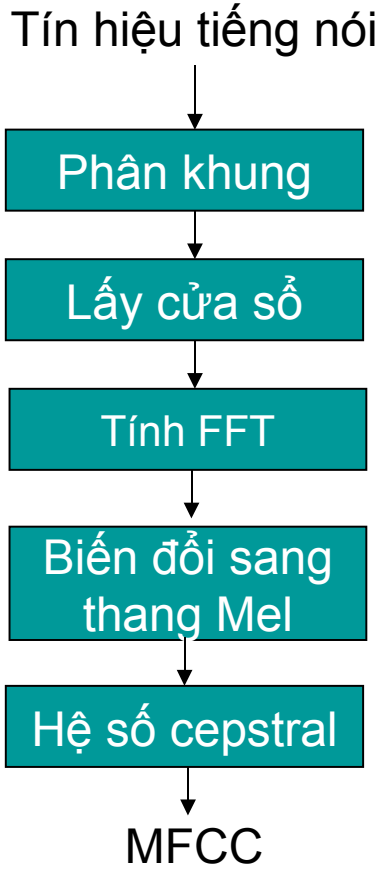
CÁC PHƯƠNG PHÁP TRÍCH CHỌN ĐẶC TRƯNG

- Các đặc trưng mang thông tin về tiếng nói:
  - Các tần số formant
  - Tần số âm cơ bản
  - 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)

CÁC BƯỚC TRÍCH CHỌN ĐẶC TRƯNG MFCC

Bao gồm 5 bước cơ bản:

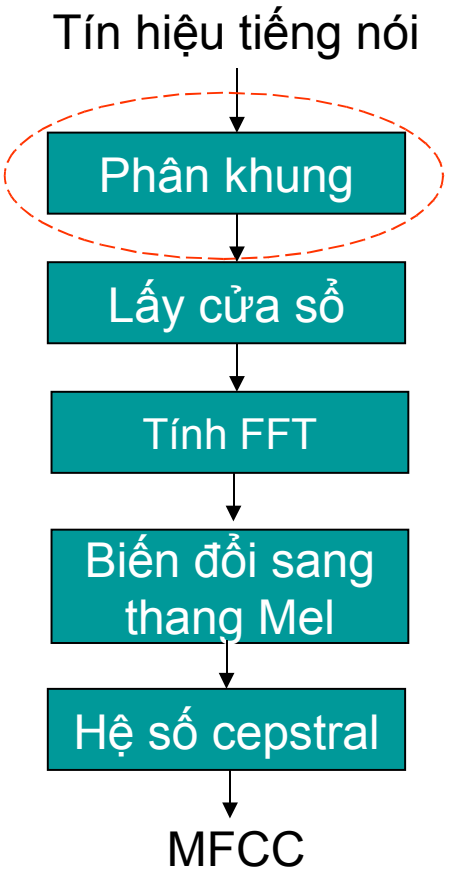
- Bước 1: Phân khung (Frame Blocking)
- Bước 2:Lấy cửa sổ (Windowing)
- Bước 3: Biến đổi FFT (Fast Fourier Transform)
- Bước 4: Biến đổi sang thang đo Mel (Mel-frequency Wrapping)
- Bước 5: Hệ số Cepstrum (Cepstral Coefficients)



CÁC BƯỚC TRÍCH CHỌN ĐẶC TRƯNG MFCC(tiếp)

Bước 1: Phân Khung

- Chia tín hiệu đầu vào thành các đoạn nhỏ khoảng 20ms-30ms.
- Phân khung tín hiệu mỗi khung N mẫu, hai khung kề nhau lệch nhau M mẫu:
- M=1/2 N

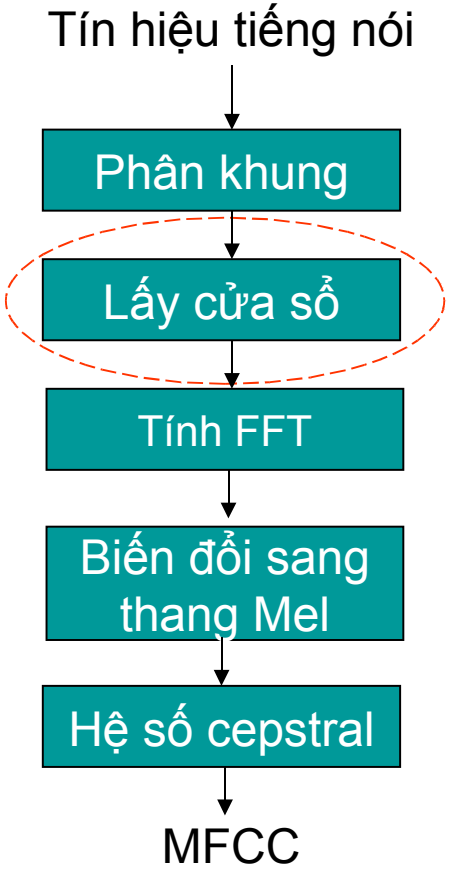


CÁC BƯỚC TRÍCH CHỌN ĐẶC TRƯNG MFCC(tiếp)

Bước 2: Lấy cửa sổ

- Lấy cửa sổ nhằm giảm sự gián đoạn của tín hiệu ở đầu và cuối mỗi khung vừa được chia.
- Dùng cửa sổ Hamming (Với α = 0.54), công thức:

$$w(n) = \begin{cases} \alpha + (1-\alpha).\cos(2\pi.n/N) & |n| \le N/2 \\ 0 & |n| > N/2 \end{cases}$$

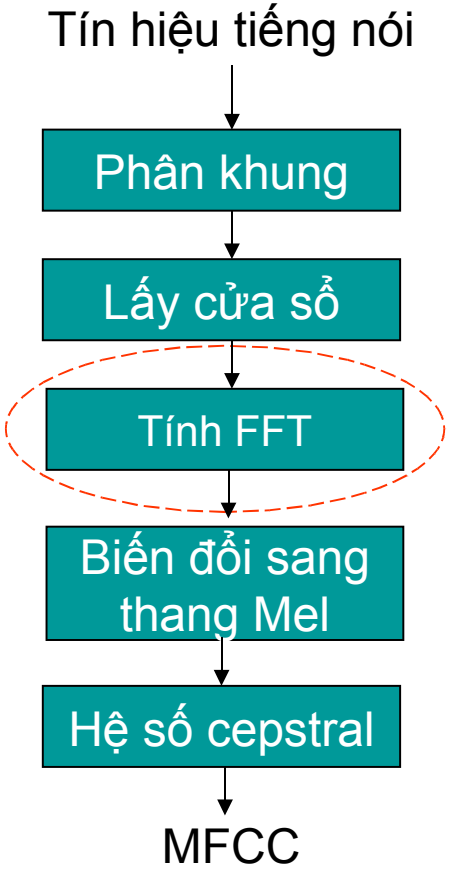


BƯỚC TRÍCH CHỌN ĐẶC TRƯNG MFCC(tiếp)

Bước 3: Tính FFT

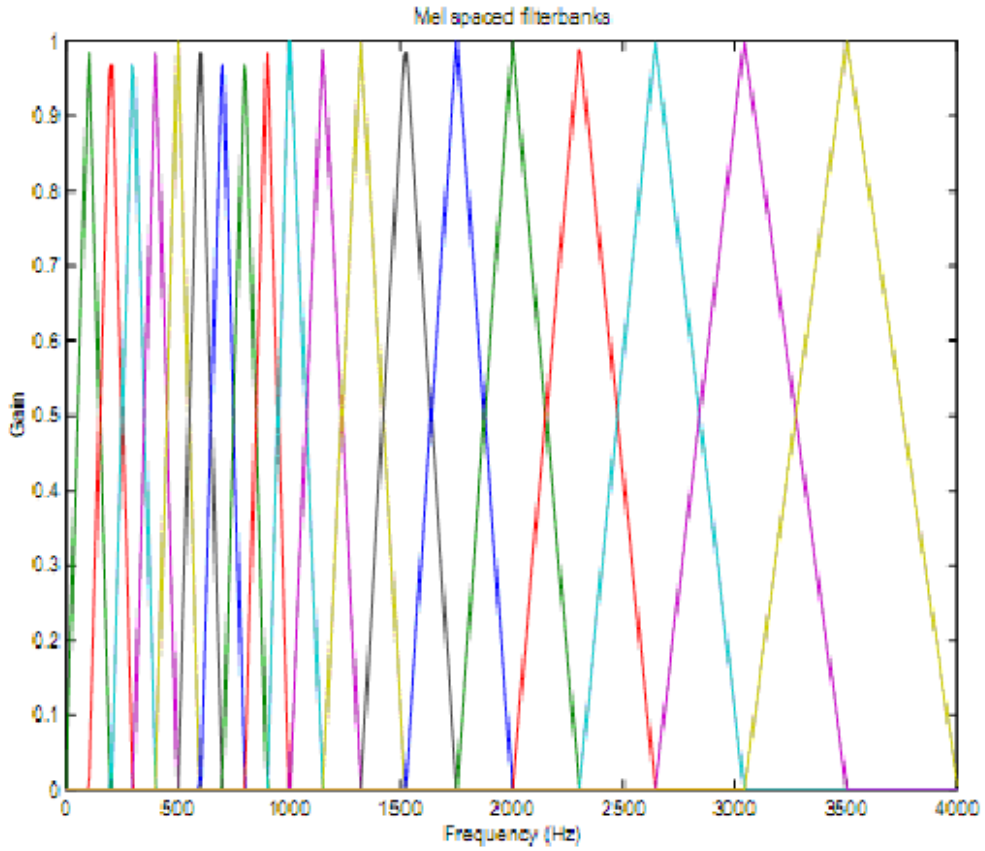
- Chuyển đổi mỗi khung với N mẫu từ miền thời gian sang miền tần số.

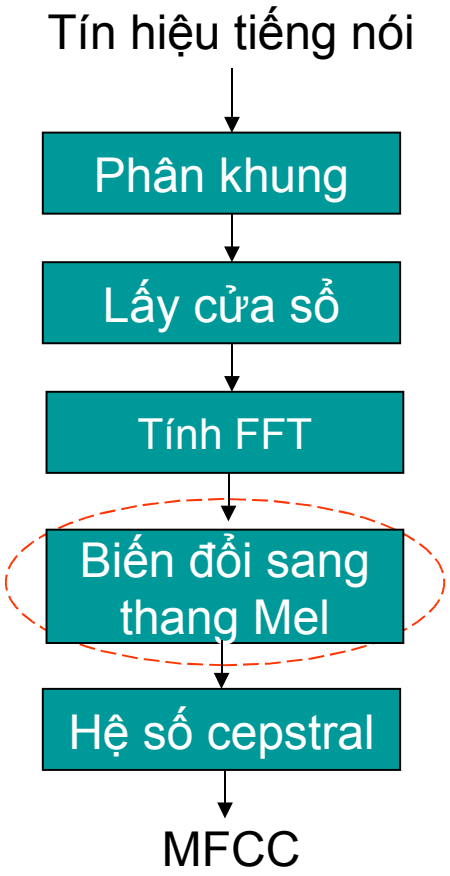
$$Y(\omega) = \int_{-\infty}^{\infty} y(t) e^{-j\omega t}$$



CÁC BƯỚC TRÍCH CHỌN ĐẶC TRƯNG MFCC(tiếp)

Bước 4: Biến đổi sang thang đo Mel



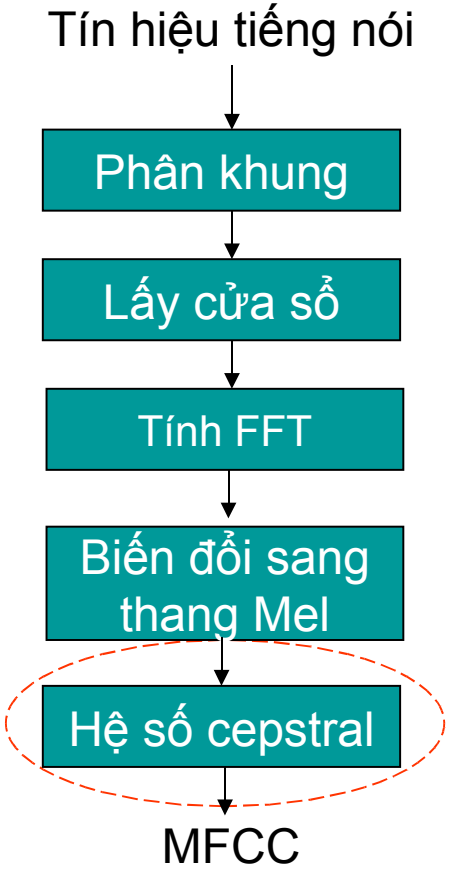


CÁC BƯỚC TRÍCH CHỌN ĐẶC TRƯNG MFCC(tiếp)

Bước 5: Hệ số Cesptral

- Kết quả của bước này là ta tính được hệ số MFCC theo công thức:

$$MFCC(i) = \frac{1}{Nfilters} \sum_{i=1}^{Nfilters} mfb(l).\cos(i(l - \frac{1}{2}) \frac{\pi}{Nfilters})$$



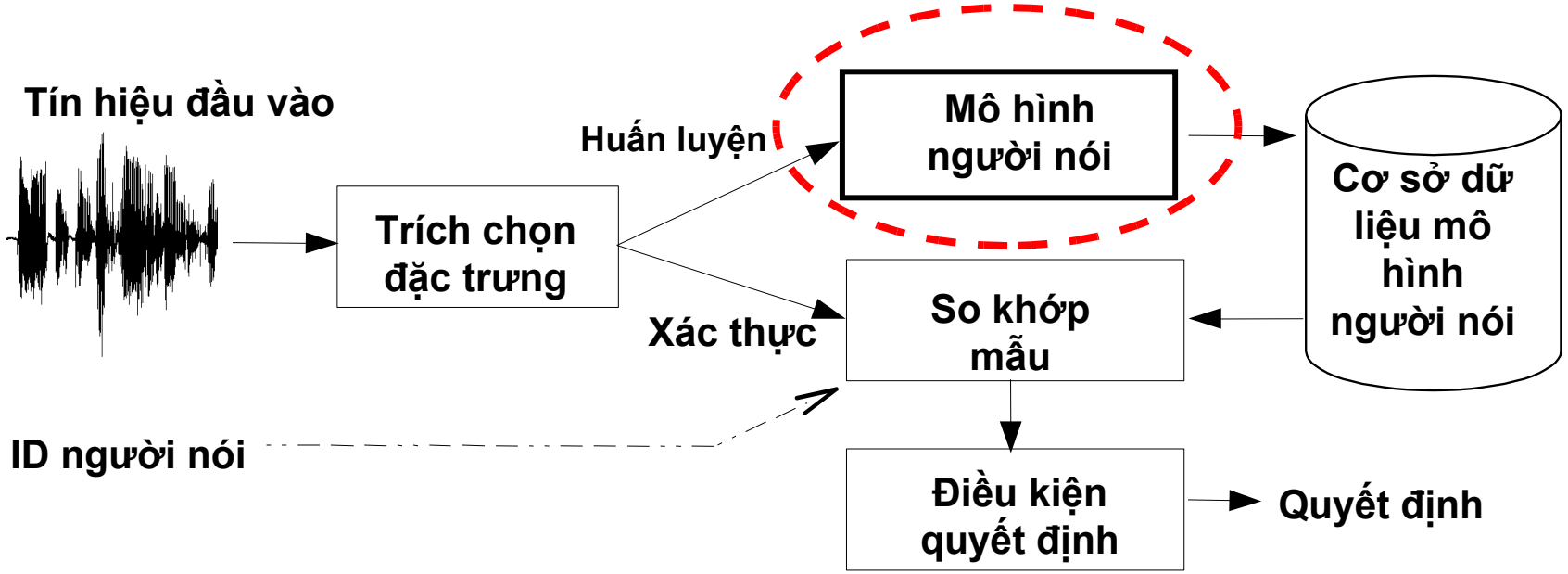
NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

MÔ HÌNH HÓA NGƯỜI NÓI

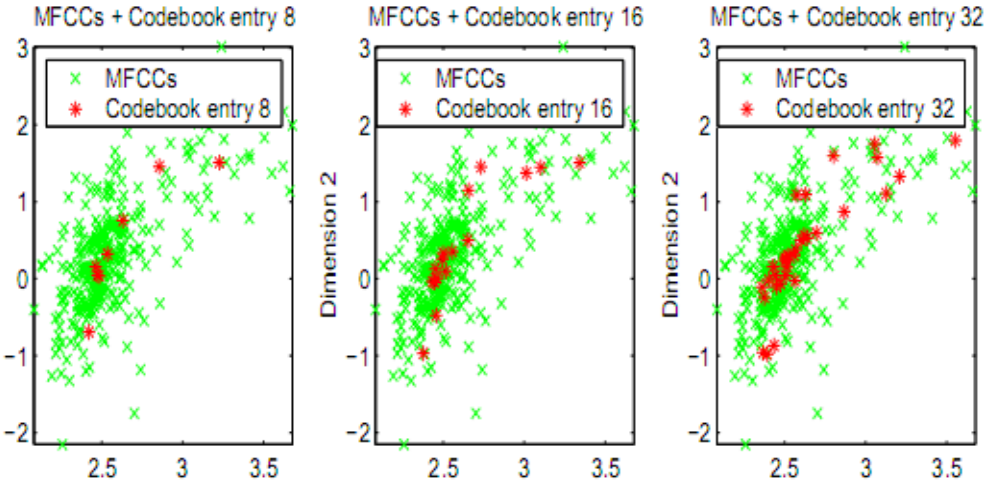


MỘT SỐ MÔ HÌNH NGƯỜI NÓI

- Mô hình lượng tử hoá vector (VQ: Vector Quantization)
- Mô hình hỗn hợp Gauss (GMM: Gaussian Mixture Model)
- Mô hình mạng Nơron nhân tạo
- Mô hình SVM

MÔ HÌNH VQ

- Tập huấn luyện gốc X là tập các vector MFCC
- Tìm cách thay X bởi tập vectơ C nhỏ hơn đặc trưng cho X gọi là codebook



- C được tạo ra bằng cách sử dụng các thuật toán gom cụm(*K*-means, Split, PNN, SOM, GA, …)
- Kích thước gom cụm cho bài toán nhận dạng người nói: 32..512

MÔ HÌNH GM

- Cho một tập có N điểm trong không gian D chiều và một họ các hàm các hàm mật độ xác suất trên R^D
- Hãy tìm hàm mật độ xác suất thích hợp nhất để có thể sinh ra các điểm trên.
- Có thể chọn là các hàm hỗn hợp Gauss: $f(x;\lambda) = \sum_{k=1}^{K} \varpi_k g(x;\mu_k, \sum_k)$
- Sử dụng thuật toán vọng số cực đại (EM_ Expectation Maximization) để tính các tham số trong hàm Gauss. Mỗi hàm hỗn hợp Gauss sẽ được biểu thị bởi:

$$\lambda = \left\{ \omega_i, \mu_i, \sum_i \right\}_{i=1}^{M}$$

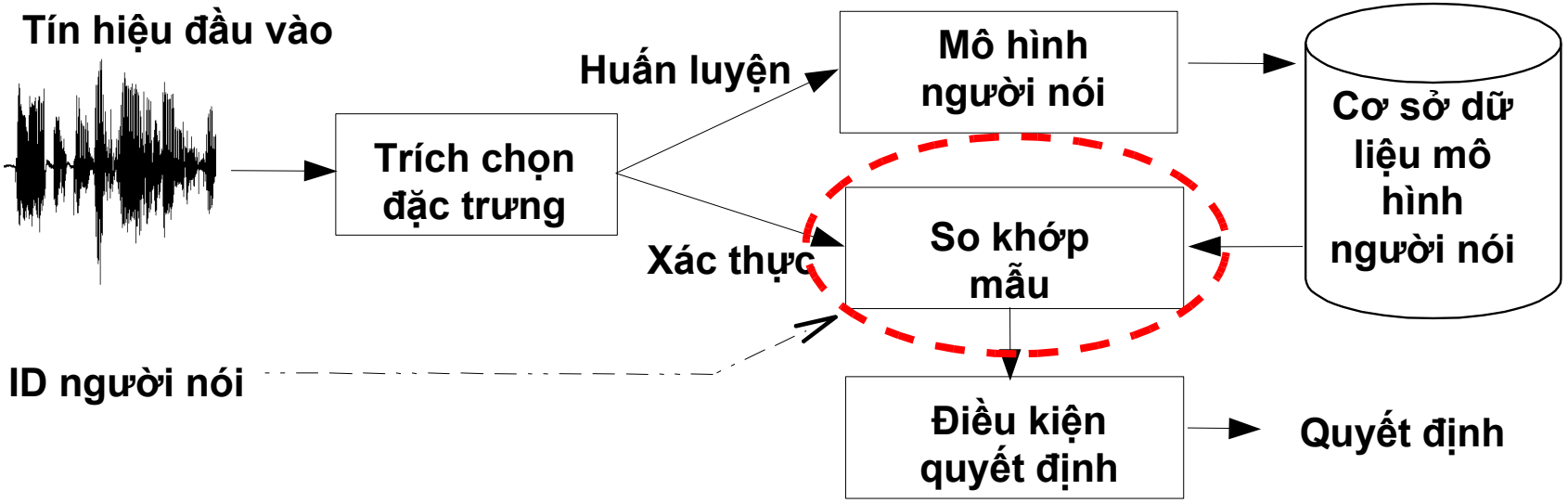
NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

SO KHỚP MẪU


Tín hiệu đầu vào
Trích chọn đặc trưng
Huấn luyện
Xác thực
Mô hình người nói
So khớp mẫu
Cơ sở dữ liệu mô hình người nói
ID người nói
Điều kiện quyết định
Quyết định

SO KHỚP MẪU -VQ

- Giả sử ta có:
  - $X=\{x_1,x_2,\ldots, x_N\}$ là các vector đặc trưng của người chưa biết
  - $C=\{c_1,c_2,\ldots,c_K\}$ là codebook
- Công thức tính độ méo lượng tử trung bình:

$$D(X, C) = \frac{1}{N} \sum_{i=1}^{N} \min_{j} \|x_i - c_j\|^2$$

- D(X,C) càng nhỏ thi X càng gần C

SO KHỚP MẪU - GMM

$X = \{x_1, ..., x_N\}$ Là vector đặc trưng cần kiểm tra

$\lambda = \{\lambda_1, ..., \lambda_K\}$ Với $\lambda j = \{\omega_j, \mu_j, \sum_j\}$ là các biến được tính trong hàm hỗn hợp Gauss

Hàm phân bố xác suất: $$p(x_i \mid \lambda) = \sum_{j=1}^{K} \omega_j N(x_i \mid \lambda_j)$$

Trong đó *N*(*x*i |λj) được tính theo công thức:

$$N(x_i \mid \lambda_j) = (2\pi)^{-d/4} \left| \sum_j \right|^{-1/2} \exp\left\{ -1/2 (x_i - \mu_j)^T \sum_j {}^{-1} (x_i - \mu_j) \right\}$$

SO KHỚP MẪU – GMM(tiếp)

- Tính log độ hợp giữa tín hiệu đầu vào X với mẫu trong hệ thống

$$P(X \mid \lambda) = \sum_{i=1}^{N} \log p(x_i \mid \lambda)$$

- Giá trị P(X| λ) càng lớn thì độ hợp của X với λ càng cao

NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

TẠO QUYẾT ĐỊNH

- Gọi *score*(*X*, *S*i) là điểm hợp giữa tập vectơ đặc trưng *X* và mô hình của người nói *S*i
- Giả sử điểm lớn hơn nghĩa là hợp hơn
- Cho *S* = {*S*1,…,*S*N} là csdl của các người nói đã biết
- Khi đó:

$$\text{Decide} \begin{cases} \text{Accept}, & score(X, S_i) \geq \Theta_i \\ \text{Reject}, & score(X, S_i) < \Theta_i \end{cases} \qquad \Theta_i = \textit{verification threshold}$$

NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

NỘI DUNG

- Trích chọn thông tin từ tiếng nói
- Tổng quan về bài toán nhận dạng người nói
- Bài toán xác thực người nói không phụ thuộc vào từ khóa
- Trích chọn đặc trưng người nói
- Mô hình hóa người nói
- So khớp mẫu
- Tạo quyết định
- Cài đặt và kết quả thử nghiệm

KẾT LUẬN

- Những nội dung chính đã được giải quyết:
  - Nghiên cứu tổng quan bài toán nhận dạng tiếng nói
  - Nghiên cứu các phương pháp trích chọn đặc trưng người nói, chi tiết phương pháp trích chọn đặc trưng MFCC
  - Nghiên cứu mô hình VQ và mô hình GMM ứng dụng trong xác thực người nói không phụ phụ vào từ khóa
  - Cài đặt các mô hình, thuật toán và tiến hành thử nghiệm trên cơ sở dữ liệu tiếng Việt

KẾT LUẬN

- Những đóng góp khoa học và tính thực tiễn :
  - Phương pháp trích chọn đặc trưng MFCC với cơ sở dữ liệu tiếng Việt
  - Ứng dụng mô hình VQ và GMM xây dựng hệ thống xác thực người nói không phụ thuộc vào từ khóa với cơ sở dữ liệu tiếng Việt
  - Đề tài có tính thực tiễn cao trong việc xây dựng các hệ thống thanh toán thẻ tín dụng qua điện thoại; đăng nhập vào các hệ thống an ninh, máy tính bằng tiếng nói; giám định tư pháp tiếng nói…

KIẾN NGHỊ

- Thu thập số lượng lớn dữ liệu âm tiếng nói để tiến hành kiểm thử, điều chỉnh các tham số của hệ thống cho chính xác.
- Tiếp tục nghiên cứu các mô hình thuật toán như mạng nơron nhân tạo, SVM, GM.. ứng dụng cho bài toán xác thực người nói.
- Kết hợp các mô hình, thuật toán đã có với hi vọng sẽ tạo ra được mô hình tốt hơn.
- Nghiên cứu bài toán xác thực người nói phụ thuộc từ khóa, kết hợp với bài toán không phụ thuộc từ khóa nhằm làm tăng tính chính xác của các hệ thống xác thực người nói.