**ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH**

TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA ĐIỆN – ĐIỆN TỬ

**BỘ MÔN ĐIỆN TỬ**

---------------o0o---------------

****

**ĐỒ ÁN MÔN HỌC 2**

**NHẬN DIỆN TIẾNG NÓI**

**GVHD: TRẦN VĂN HOÀNG**

**SVTH: BÙI HUY AN**

**MSSV: 41200001**

**TP. HỒ CHÍ MINH, THÁNG 6 NĂM 2016**

**MỤC LỤC**

[1. GIỚI THIỆU 5](#_Toc452714433)

[Tổng quan 5](#_Toc452714434)

[Yêu cầu: 5](#_Toc452714435)

[2. LÝ THUYẾT 6](#_Toc452714436)

[Sự biểu hiện của tiếng nói: 6](#_Toc452714437)

[Các yếu tố ảnh hưởng đến kết quả nhận dạng: 7](#_Toc452714438)

[Các hướng nghiên cứu: 7](#_Toc452714439)

[3. THIẾT KẾ VÀ THỰC HIỆN: 9](#_Toc452714440)

[Các sơ đồ khối: 9](#_Toc452714441)

[4. KẾT QUẢ THỰC HIỆN 13](#_Toc452714442)

[5. TÀI LIỆU THAM KHẢO 14](#_Toc452714443)

**DANH SÁCH HÌNH MINH HỌA**

[Hình 1. Dạng sóng ở miền tần số. 6](#_Toc452711651)

[Hình 2. Sơ đồ khối thực hiện trích đặc trưng MFCC 7](#_Toc452711652)

[Hình 3. Sơ đồ khối thực hiện huấn luyện bằng phương pháp VQ 8](#_Toc452711653)

[Hình 4. Sơ đồ khối thực hiện tìm tâm của tín hiệu. 9](#_Toc452711654)

**DANH SÁC CÁC CÔNG THỨC:**

[Equation 3‑1 10](#_Toc452714444)

[Equation 3‑2 11](#_Toc452714445)

[Equation 3‑3 11](#_Toc452714446)

[Equation 3‑4 11](#_Toc452714447)

[Equation 3‑5 12](#_Toc452714448)

# GIỚI THIỆU

## Tổng quan

Nhận dạng tiếng nói là một quá trình nhận dạng mẫu, với mục đích là phân lớp (classify) 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. 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. 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 cũng khó khăn trong việc tổng quát hoá từ các mẫu tiếng nói những biến thiên quan trọng cần thiết trong nhận dạng tiếng nói.

Các nghiên cứu về nhận dạng tiếng nói dựa trên ba nguyên tắc cơ bản:

* Tín hiệu tiếng nói được biểu diễn chính xác bởi các giá trị phổ trong một khung thời gian ngắn (short-term amplitude spectrum). Nhờ vậy ta có thể trích ra các đặc điểm tiếng nói từ những khoảng thời gian ngắn và dùng các đặc điểm này làm dữ liệu để nhận dạng tiếng nói.
* Nội dung của tiếng nói được biểu diễn dưới dạng chữ viết, là một dãy các ký hiệu ngữ âm. Do đó ý nghĩa của một phát âm được bảo toàn khi chúng ta phiên âm phát âm thành dãy các ký hiệu ngữ âm.
* Nhận dạng tiếng nói là một quá trình nhận thức. Thông tin về ngữ nghĩa (semantics) và suy đoán (pragmatics) có giá trị trong quá trình nhận dạng tiếng nói, nhất là khi thông tin về [âm học](https://vi.wikipedia.org/wiki/%C3%82m_h%E1%BB%8Dc) là không rõ ràng.

Cách tiếp cận nhận dạng tiếng nói bằng thống kê bao gồm: sử dụng mô hình Markov ẩn, mạng nơ-rơn, sử dụng cơ sở tri thức, v.v..

## Yêu cầu:

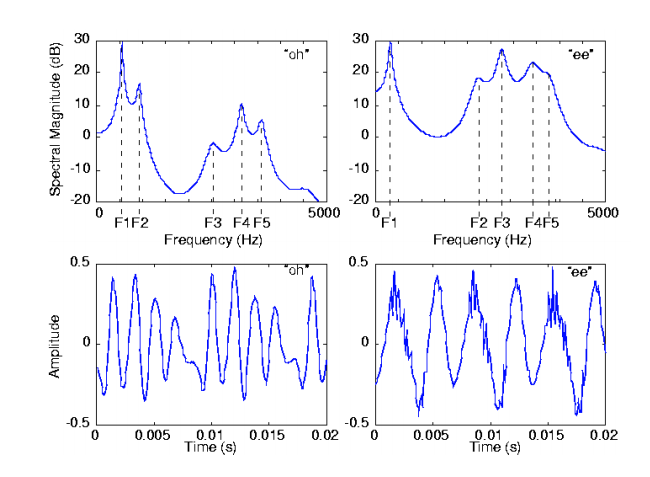
* Độ chính xác cao.
* Thời gian nhận dạng thấp.
* Chống nhiễu từ môi trường tốt.

# LÝ THUYẾT

## Sự biểu hiện của tiếng nói:

**Miền thời gian:** Tín hiệu tiếng nói là tín hiệu thay đổi chậm theo thời gian. Khi quan sát tín hiệu tiếng nói trên một khoảng thời gian đủ ngắn (5 đến 20ms), khi đó các đặc trưng của nó xem như không đổi. Tuy nhiên trên khoảng thời gian dài (0.5 giây hoặc nhiều hơn) các đặc trưng của tín hiệu thay đổi phản ánh những âm thanh khác nhau được nói.

Một cách thức đơn giản để phân loại tiếng nói đó là đánh nhãn nó với 3 trạng thái: trạng thái thứ nhất là yên lặng, đây chính là khoảng thời gian mà không có tiếng nói; trạng thái thứ hai là âm vô thanh (unvoiced), đây là khoảng thời gian mà dạng sóng tiếng nói là không tuần hoàn hay nó mang tính ngẫu nhiên; trạng thái thứ ba là âm hữu thanh (voiced), trong khoảng thời gian này dạng sóng của tiếng nói mang tính tuần hoàn. Cách thức phân loại này không chính xác lắm do ảnh hưởng tại những vị trí ranh giới giữa các trạng thái nhưng những lỗi nhỏ này thường không gây hậu quả gì lớn cho hầu hết các ứng dụng.  
 Âm vô thanh có năng lượng tập trung ở tần số cao, các tần số phân bố khá đồng đều trong từ tần số thấp đến tần số cao. Các nguyên âm là âm hữu thanh (voiced), nó liên tục và tương đối dài trong một khoảng thời gian.

**Miền tần số:** Chúng ta cũng có thể biểu diễn tín hiệu tiếng nói trong miền tần số, trục tung biểu diễn tần số, trục hoành biẻu diễn biên độ. 

Hình 1. Dạng sóng ở miền tần số.

Các yếu tố ảnh hưởng đến kết quả nhận dạng:  
• Hệ thống được thiết kế cho một hay nhều người.  
• Kích thước bộ từ vựng.  
• Tiếng nói được đưa vào bằng các từ rời rạc với khoảng dừng đủ lớn hay phát âm liên tục.  
• Sự nhầm lẫn và không rõ rang giữa các âm trong bộ từ vựng.  
• Hệ thống nhận dạng trong môi trường yên tĩnh hay nhiễu.  
• Kiến thức áp dụng cho bộ nhận dạng.

Các hướng nghiên cứu:  
 Có 3 phương pháp tiếp cận chính để nhận dạng tiếng nói: ngữ âm học, nhận dạng mẫu và ứng dụng trí tuệ nhân tạo.  
 • Ngữ âm học: dựa vào lý thuyết về ngữ âm âm thanh, lý thuyết này cho rằng tiếng nói tồn tại hữu hạn, các đơn vị âm thanh có thể phân biệt được và được đặc trưng bởi một tập các tính chất biểu hiện cho tín hiệu tiếng nói trong miền thời gian hay các tính chất phổ của nó. Bước đầu tiên trong phương pháp ngữ âm học là phân đoạn và gán  
nhãn, bước thứ hai là xác định từ hợp lệ từ dãy ngữ âm được gán nhãn trong bước đầu tiên.  
 • Nhận dạng mẫu: gồm hai bước chính là huấn luyện mẫu và so sánh mẫu. Đặc tính chủ yếu của phương pháp này là sử dụng các cơ cấu toán học rõ ràng và thiết lập sự biễu diễn thích hợp các mẫu tiếng nói cho việc so sánh chính xác các mẫu huấn luyện . Tiếng nói có thể được biễu diễn ở dạng mẫu tiếng nói hoặc một mô hình thống kê, có thể được áp dụng cho một âm, một từ hoặc một nhóm từ. Trong giai đoạn so sánh mẫu, có một sự so sánh trực tiếp giữa tiếng nói chưa biết với mỗi mẫu được học trong giai đoạn huấn luyện để xác định đặc tính của mẫu chưa biết. Hướng kết hợp mẫu trở thành phương pháp nhận dạng rất phổ biến trong những năm 90.  
 • Phương pháp trí tuệ nhân tạo: nghiên cứu cách học nói và học nghe của con người, tìm hiểu các quy luật ngữ âm, ngữ pháp, ngữ nghĩa, ngữ cảnh… và tích hợp chúng bổ  
sung cho các phương pháp khác để nâng cao kết quả nhận dạng.

# THIẾT KẾ VÀ THỰC HIỆN:

## Các sơ đồ khối:



Hình 2. Sơ đồ khối thực hiện trích đặc trưng MFCC

**Khối Pre-emphasis:**

Tín hiệu sau khi được số hóa được pre-emphasis với bộ lọc có đáp ứng xung hữu hạn (bộ lọc FIR) bậc nhất vì pha của nó là tuyến tính và thực thi đơn giản. Do trong tiếng nói, các thành tố thấp hơn thường chứa đựng nhiều năng lượng hơn, vì vậy nó được xem xét hơn khi mô hình hóa so với các thành tố cao hơn. Do đó, một bộ lọc preemphasis được dùng để khuếch đại tín hiệu ở các tần số cao hơn. Hàm truyền của bộ lọc được cho bởi phương trình (3.1), trong đó tham số *a* tiêu biểu từ 0.9 đến 1. Trong miền thời gian, mối quan hệ giữa ngõ ra với ngõ vào được chỉ ra trong phương trình (3.2), với *si* là điểm thứ *i* của tín hiệu tiếng nói khi chưa qua bộ lọc và *si*′ là điểm thứ *i* của tín hiệu tiếng nói sau khi được pre-emphasis.

|  |  |
| --- | --- |
| Spi = si  - 0.97\*si-1 | Equation 3‑1 |

**Frame Blocking:**  
 Bởi vì tín hiệu tiếng nói là tín hiệu biến đổi chậm theo thời gian, trong một hệ thống nhận dạng tiếng nói thì tiếng nói được phân đoạn thành những khoảng thời gian ngắn được gọi là các frame. Để cho các thông số của frame ít thay đổi, thông thường có 50% chồng lấp giữa các frame kế cận nhau. Trong các hệ thống nhận dạng tiếng nói bằng phần mềm, tiếng nói được chia thành những frame có chiều dài 20ms với 10ms chồng lấp. Khi tiếng nói được lấy mẫu với tần số 8KHz sẽ có 160 mẫu trong mỗi frame và có 80 mẫu được chồng lấp giữa 2 frame kế cận nhau.

**Windowing:**

Một cửa sổ thường được ứng dụng để gia tăng tính liên tục giữa các frame kế cận nhau. Một trong các cửa sổ được dùng phổ biến nhất trong nhận dạng tiếng nói đó chính là cửa sổ Hamming được xác định bởi phương trình (3.3), trong đó *L* là chiều dài cửa sổ và nó bằng với chiều dài của các frame.

|  |  |
| --- | --- |
| h(l)=0.54-0.46cos l=1,2,3,….,L | Equation 3‑2 |

Tiếng nói sau khi được chia thành những frame có chiều dài 160 mẫu với 50% chồng lấp, 160 điểm cửa sổ Hamming được nhân với mỗi frame theo từng mẫu. Các frame ngõ ra cửa sổ được liên tục tại điểm đầu và điểm cuối của mỗi frame. Bước này có thể được diễn giải bởi phương trình (3.4), trong đó *f n*(*l*) là frame được pre-emphasis thứ *n*, *ham(l)* thay cho cửa sổ Hamming *wfn* (*l*) là frame thứ *n* sau khi qua cửa sổ Hamming.

|  |  |
| --- | --- |
| *Wfn*(*l*)= *f n*(*l*)⋅ *ham*(*l*) l=1,2,3,..160 | Equation 3‑3 |

**Tính năng lượng:**

Công suất của mỗi frame cũng là thành phần trong đặc trưng MFCC. Nó được tính toán như là logarithm của công suất tín hiệu, có nghĩa là đối với frame thứ ***n***, mỗi frame có 160 mẫu {*snl*, *l* =1, 2,...,160}

|  |  |
| --- | --- |
| logEn=log () | Equation 3‑4 |

**Hệ số delta:**

Chất lượng của hệ thống nhận dạng tiếng nói có thể được cải thiện nhiều hơn bằng cách thêm vào tính đạo hàm theo thời gian để có được những thông số dừng cơ bản.

|  |  |
| --- | --- |
| Dn= | Equation ‑ |



Hình 3. Sơ đồ khối thực hiện huấn luyện bằng phương pháp VQ



Hình 4. Sơ đồ khối thực hiện tìm tâm của tín hiệu.

# KẾT QUẢ THỰC HIỆN

Ưu điểm:

* Nhận diện nhanh.
* Độ chính xác cao.
* Giao diện dễ sử dụng.

Khuyết điểm:

* Tập huấn luyện còn ít nên nhận diện còn ít từ.
* Khó khăn khi phân biệt giọng của nhiều người.TÀI LIỆU THAM KHẢO
* <https://vi.wikipedia.org/wiki/Trang_Ch%C3%ADnh>
* Fundamentals of Speech Recognition
* Spoken Language Processing A Guide to Theory algorithm, and System Development