

TRƯỜNG ĐẠI HỌC THỦY LỢI
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO BÀI TẬP LỚN
HỌC PHẦN XỬ LÝ TIẾNG NÓI

**PHÂN LOẠI GIỚI TÍNH DỰA TRÊN GIỌNG NÓI ĐA NGÔN
NGỮ BẰNG CAO ĐỘ (F_0) & MFCC**

Giáo viên hướng dẫn: PGS.TS. Nguyễn Quang Hoan

Học viên thực hiện: Ngô Đức Tâm

HÀ NỘI, 2025

MỤC LỤC

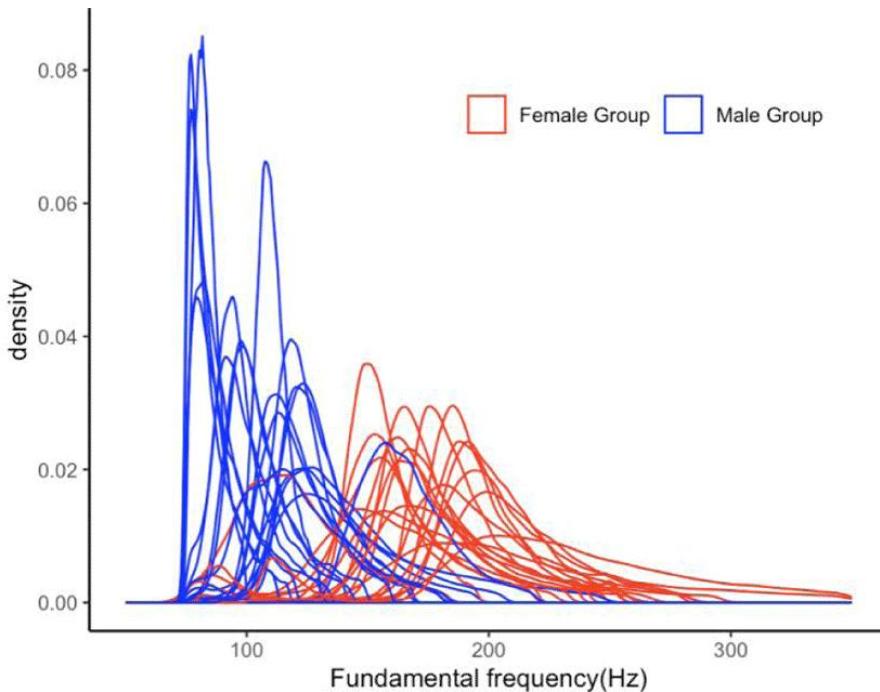
MỤC LỤC	i
CHƯƠNG 1 LÝ THUYẾT	1
1.1 Tổng quan về tiếng nói và đặc trưng giọng nói	1
1.2 Đặc trưng MFCC (Mel Frequency Cepstral Coefficients)	4
1.3 Đặc trưng F0 (tần số cơ bản – Pitch).....	6
1.4 Các phương pháp phân loại giới tính từ giọng nói.....	7
1.4.1 Phương pháp truyền thống (machine learning cổ điển)	7
1.4.2 Phương pháp học sâu (deep learning)	8
CHƯƠNG 2 ỨNG DỤNG TRONG THỰC TẾ	11
2.1 Giới thiệu đề tài	11
2.2 Thiết kế & triển khai.....	12
2.2.1 Dữ liệu và công cụ.....	12
2.2.2 Tiền xử lý dữ liệu	13
2.2.3 Trích xuất đặc trưng MFCC và F0	13
2.2.4 Lựa chọn mô hình phân loại	15
2.2.5 Hệ thống dự đoán giới tính thời gian thực	16
2.3 Kết quả & đánh giá.....	17
2.4 Kết luận.....	19
2.4.1 Kết quả đạt được.....	19
2.4.2 Ưu điểm và hạn chế	19
2.4.3 Hướng phát triển mở rộng	19
TÀI LIỆU THAM KHẢO	21

CHƯƠNG 1 LÝ THUYẾT

1.1 Tổng quan về tiếng nói và đặc trưng giọng nói

Tiếng nói của con người được tạo ra bởi sự dao động của dây thanh quản khi luồng không khí đi từ phổi qua thanh quản. Âm thanh thô từ thanh quản (âm gốc) sau đó được cộng hưởng và biến đổi bởi khoang họng, miệng, mũi để tạo thành giọng nói hoàn chỉnh. Đặc trưng cơ bản của giọng nói bao gồm **cao độ** (pitch – tần số cơ bản F0 của dao động dây thanh), **cường độ** (intensity – biên độ âm thanh, liên quan đến độ to nhỏ), **trường độ** (duration – độ dài âm) và **âm sắc** (timbre – hình dạng phổ tần). Mỗi đặc trưng này góp phần tạo nên sự khác biệt về giọng giữa các cá nhân, bao gồm cả **khác biệt giới tính** (nam và nữ) trong giọng nói.

Cao độ (F₀) của giọng nói – tức tần số dao động của dây thanh âm – là một thông số âm học quan trọng phân biệt giọng nam và giọng nữ. Nhiều nghiên cứu đã ghi nhận rằng F₀ trung bình ở **nam** giới trưởng thành vào khoảng **120 Hz**, trong khi ở **nữ** giới trưởng thành khoảng **210Hz**, nghĩa là giọng nữ có cao độ cơ bản cao hơn đáng kể so với giọng nam. Sự chênh lệch này bắt nguồn từ việc ở tuổi dậy thì, thanh quản nam phát triển lớn hơn, dây thanh dày và dài hơn dưới tác động của hormone testosterone, dẫn đến tần số dao động chậm hơn (giọng trầm hơn) so với nữ [1]. Thực vậy, phân tích các phân bố F₀ cho thấy hầu hết nữ giới có dải F₀ nằm ở vùng tần số cao hơn nam giới; chẳng hạn, trong một ngữ liệu tiếng Anh đàm thoại, trung vị F₀ của nữ (~181 Hz) cao hơn rõ rệt so với của nam (~113 Hz). Chính nhờ sự khác biệt đáng kể về cao độ này, F₀ thường được sử dụng như một đặc trưng quan trọng trong phân loại giới tính tự động bằng giọng nói, giúp các hệ thống nhận dạng giọng nói phân biệt hiệu quả giữa nam và nữ dựa trên cao độ cơ bản của họ [1]. Hình bên dưới sự phân bố F₀ giữa giọng nam và nữ: giọng nam tập trung ở các tần số thấp (~100 Hz), còn giọng nữ tập trung ở tần số cao hơn (~185 Hz). Hai phổ F₀ gần như tách biệt, chỉ hơi trùng nhau quanh vùng 150–170 Hz.

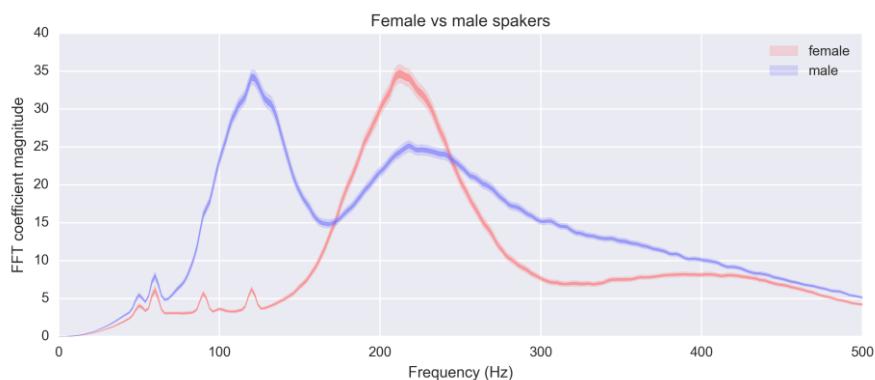


Hình 1.1. Biểu đồ phân bố của F0 cho giọng nam (màu xanh) và giọng nữ (màu đỏ).

Nam giới có xu hướng F0 thấp hơn, còn nữ giới cao hơn. Nguồn: [1]

Bên cạnh cao độ, **phổ tần số và âm sắc** của giọng nói nam/nữ cũng khác nhau. Nam giới thường có **vocal tract** (đường thanh âm – bao gồm khoang họng, miệng) dài hơn khoảng 15% so với nữ giới, khiến các **formant** (các đỉnh cộng hưởng chính của phổ âm) của nam nằm ở tần số thấp hơn so với nữ [2]. Nói cách khác, với cùng một nguyên âm, giọng nam có các formant thấp hơn (do khoang cộng hưởng lớn hơn) trong khi giọng nữ có formant cao hơn. Ngoài ra, do F0 nam thấp hơn, **mật độ phổ** của giọng nam ở vùng tần số thấp dày hơn (các harmonics – họa tần – của nam sát nhau hơn), tạo cảm giác giọng “trầm ám” hơn. Ngược lại, giọng nữ có khoảng cách giữa các harmonics rộng hơn (do F0 cao hơn), thường mang âm sắc “thanh hơn”. Sự khác biệt phổ giữa giọng nam và nữ được minh họa ở hình sau: giọng nam (màu xanh) có năng lượng nổi trội ở tần số ~ 100 Hz (ứng với F0 nam) và giảm dần ở vùng cao; giọng nữ (màu đỏ) có đỉnh ở ~ 200 Hz (F0 nữ) và năng lượng tương đối mạnh hơn ở vùng trên 200 Hz [3]. Điều này giải thích vì sao khi nghe, cảm nhận giọng nữ cao và “mảnh” hơn, còn giọng nam trầm và “dày” hơn.

Phổ trung bình của giọng nữ (đỏ) so với giọng nam (xanh) trong tiếng Anh. Giọng nam có đỉnh năng lượng tại khoảng ~100–125 Hz (F0 nam), trong khi giọng nữ đỉnh ở ~200–250 Hz (F0 nữ). Giọng nam cũng thể hiện năng lượng tương đối cao ở dải trầm (<500 Hz) hơn so với giọng nữ [3]. Ngược lại, giọng nữ có xu hướng tập trung năng lượng ở dải cao hơn. Các đặc biệt này là đặc trưng giúp phân biệt giới tính dựa trên giọng nói.



Hình 1.1.1. Phổ âm thanh người nói tiếng Anh [3]

Cường độ (intensity) hay biên độ âm thanh cũng góp phần tạo nên sự khác biệt. Thông thường, nam giới có dung tích phổi lớn hơn và cách phát âm có thể tạo áp lực âm thanh mạnh hơn, dẫn đến giọng nam có thể to hơn giọng nữ một chút trong cùng điều kiện. Tuy nhiên, cường độ phụ thuộc nhiều vào ngữ cảnh và cá nhân hơn là yếu tố sinh học thuận túy. Một nghiên cứu cho thấy mức cường độ hội thoại trung bình của nam khoảng 70.4 dB và nữ khoảng 68.1 dB – chênh lệch không lớn (~2 dB) [4]. Do đó, cao độ và phổ âm vẫn là những đặc trưng âm học quan trọng và ổn định hơn để phân biệt giới tính giọng nói, trong khi cường độ không phải là tiêu chí đáng tin cậy nếu xét riêng lẻ.

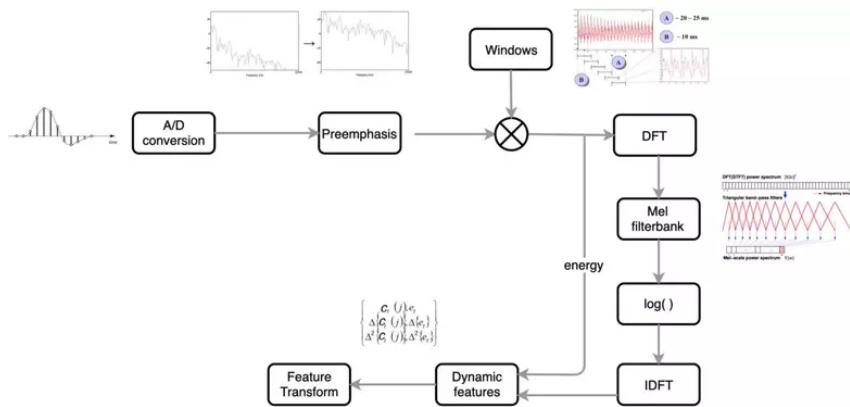
Tóm lại, giọng nói nam và nữ khác biệt rõ ở cao độ (F0), cấu trúc phổ (vị trí các formant, phân bố năng lượng tần số) và phân nào ở âm sắc tổng thể. Những khác biệt này hình thành do yếu tố sinh lý (cấu tạo thanh quản, vocal tract) và nội tiết tố, và con người có thể dễ dàng nhận biết giới tính qua giọng nói nhờ chúng. Đây chính là cơ sở để xây dựng các hệ thống phân loại giới tính tự động từ giọng nói: trích xuất các đặc trưng âm thanh đại diện cho các khác biệt nam/nữ và dùng thuật toán để phân loại.

1.2 Đặc trưng MFCC (Mel Frequency Cepstral Coefficients)

Mel Frequency Cepstral Coefficients (MFCC) là bộ đặc trưng thịnh hành nhất trong xử lý tiếng nói và nhận dạng giọng nói. MFCC mô tả bao hàm phổ tần số của tín hiệu tiếng nói theo thang Mel – một thang tần phi tuyến tính mô phỏng cách tai người cảm nhận âm thanh. Thay vì phân tích phổ tuyến tính, MFCC dùng thang Mel để tập trung độ phân giải cao ở dải tần thấp (nơi tai người nhạy cảm) và độ phân giải thấp hơn ở dải cao (nơi tai kém nhạy). Quá trình tính toán MFCC từ một tín hiệu tiếng nói gồm các bước chính:

- **Tiền xử lý & chia khung (framing):** Tín hiệu được tiền khuếch đại tần số cao (pre-emphasis) để làm nổi bật thành phần tần số cao vốn có năng lượng thấp. Sau đó, tín hiệu được chia thành các khung ngắn (frame) dài khoảng 20–25 ms, thường overlap nhau 10 ms. Việc chia khung (cửa sổ) giúp giả định tín hiệu ổn định trong khoảng thời gian ngắn, phục vụ phân tích Fourier cục bộ. Mỗi frame được nhân với một cửa sổ (ví dụ Hamming) để giảm biên độ đột ngột ở rìa khung, tránh nhiễu phổ tần cao do hiệu ứng cắt cửa sổ.
- **Biến đổi Fourier & phổ tần số:** Mỗi frame sau đó được biến đổi Fourier rời rạc (DFT) để thu được phổ biên độ (magnitude spectrum) theo tần số. Phổ này cho biết mức năng lượng tại các tần số khác nhau trong khung âm thanh. Do đặc điểm giọng nói, phổ thường có năng lượng mạnh ở tần số thấp và giảm dần ở tần số cao.
- **Áp dụng thang Mel (Mel filterbank):** Phổ năng lượng được đi qua một ngân hàng lọc Mel gồm nhiều bộ lọc dải hẹp trải đều trên thang Mel. Mỗi bộ lọc cộng dồn năng lượng trong một dải tần nhất định. Ở tần số thấp, các bộ lọc dày hơn (độ phân giải cao), trong khi ở tần số cao, bộ lọc rộng hơn (độ phân giải thấp). Kết quả thu được là phổ Mel – phản ánh năng lượng tín hiệu trên thang tần Mel.
- **Logarithm:** Lấy log năng lượng của phổ Mel để thu được phổ Mel-log. Thao tác log giúp biến đổi dải năng lượng rộng thành thang logarit, phù hợp với cách tai người cảm nhận âm lượng, đồng thời giảm ảnh hưởng của các biến thiên cục bộ nhỏ và nhiễu không đáng kể.
- **Biến đổi Cosine ngược (DCT):** Áp dụng biến đổi Cosine rời rạc (Discrete Cosine Transform) lên phổ Mel-log sẽ thu được các hệ số cepstral. DCT có tác dụng nén thông tin và tạo ra các hệ số ít tương quan với nhau (decorrelation). Thông thường, chỉ 12–13 hệ số cepstral đầu tiên được giữ lại làm đặc trưng MFCC của frame, vì các

hệ số cao hơn chủ yếu mang nhiều hoặc thông tin ít quan trọng. Hệ số thứ 0 hoặc 13 thường đại diện cho năng lượng tổng thể của frame (có thể thay bằng log năng lượng khung). Ngoài ra, để bổ sung thông tin biến thiên theo thời gian, người ta thường tính thêm các hệ số delta và delta-delta (đạo hàm bậc 1 và bậc 2 theo thời gian) của các MFCC gốc, nâng tổng số đặc trưng MFCC lên 39 (13 static + 13 delta + 13 delta-delta).



Hình 1.2.1. Sơ đồ các bước trích xuất đặc trưng trong MFCC [5]

Điểm đặc biệt quan trọng: phép biến đổi cepstrum (DCT trên log-phổ Mel) có tác dụng tách biệt thông tin nguồn và thông tin lọc của tín hiệu tiếng nói. Nói đơn giản, nó giúp phân tách ảnh hưởng của F0 (cao độ) và phong cách phô do vocal tract. Trong miền cepstrum, các thành phần liên quan F0 xuất hiện tách biệt ở phần “quefrency” cao, còn thông tin phô vỏ (formant) nằm ở phần quefrency thấp. Bằng cách chỉ lấy các hệ số cepstral đầu (thường 12), MFCC loại bỏ hầu hết thông tin về F0, chỉ giữ lại đặc trưng về vỏ phô (formant, âm sắc) của giọng nói. Chính điều này làm MFCC ít phụ thuộc vào người nói – một ưu điểm lớn trong nhận dạng tiếng nói (ASR) vì giúp mô hình tập trung vào nội dung thay vì giọng từng người. Tuy nhiên, trong bài toán phân biệt giới tính, thông tin F0 lại hữu ích, do đó thường kết hợp MFCC với đặc trưng F0 để tận dụng cả thông tin phô và cao độ.

Tóm lại, [6] MFCC cung cấp một biểu diễn gọn của phô âm thanh theo cách gần với cảm nhận thính giác con người. MFCC rất phổ biến trong các hệ thống nhận dạng tiếng nói và phân loại người nói nhờ tính hiệu quả và khả năng chống nhiễu tốt. Chẳng hạn, MFCC tỏ ra bền vững ngay cả trong môi trường có tiếng ồn và được dùng thành công

cho nhận dạng loa và giới tính trên dữ liệu thoại điện thoại chất lượng thấp. Nhiều nghiên cứu đã chỉ ra việc sử dụng MFCC làm đầu vào cho mô hình phân loại giới tính cho độ chính xác cao. Tuy MFCC ban đầu không chứa trực tiếp thông tin cao độ, nhưng các khác biệt về phô do cấu trúc thanh quanh nam/nữ (như vị trí formant) vẫn được phản ánh phần nào trong MFCC. Do đó, MFCC kết hợp với F0 sẽ là bộ đặc trưng phù hợp để phân loại giới tính giọng nói.

1.3 Đặc trưng F0 (tần số cơ bản – Pitch)

Như đã trình bày, **F0 (fundamental frequency)** – hay cao độ giọng – là đặc trưng quan trọng và dễ nhận biết nhất giữa giọng nam và nữ. F0 được xác định bởi tốc độ dao động của dây thanh: dây thanh càng dài và dày (ở nam), dao động càng chậm → F0 thấp; dây thanh mảnh và căng (ở nữ), dao động nhanh → F0 cao. Trong xử lý tiếng nói, việc ước lượng F0 từ tín hiệu (gọi là *pitch tracking*) là một bước quan trọng. Có nhiều thuật toán xác định F0: phương pháp tương quan tự động (ACF), phương pháp qua miền cepstrum, hoặc các thuật toán hiện đại như **YIN** và **PYIN** dựa trên phân tích chu kỳ. Các thuật toán này có gắng tìm khoảng thời gian lặp lại của sóng âm hoặc vị trí các đỉnh phô hài để suy ra tần số cơ bản.

Trong phân loại giới tính, có thể sử dụng **giá trị F0 trung bình** của một đoạn nói, hoặc **phân bố F0** trên đoạn đó, làm đặc trưng. Thông thường, chỉ cần giá trị trung bình hoặc trung vị cũng đủ phân biệt hai lớp, do khoảng cách giữa F0 nam – nữ khá lớn so với độ dao động nội tại của từng người. Ngoài ra, có thể kết hợp thêm các đặc trưng liên quan F0 khác như **dộ biến thiên F0 theo thời gian**, **dộ ổn định cao độ**, tuy nhiên những đặc trưng này thường phục vụ nhận diện cảm xúc hoặc ngữ điệu hơn là phân biệt giới tính. Trong phạm vi đề tài, **F0 trung bình** của mỗi đoạn âm được chọn làm một đặc trưng đầu vào mô hình.

Cũng cần lưu ý, một số hệ thống nhận dạng giọng nói trước đây từng **loại bỏ F0** khỏi đặc trưng do quan niệm F0 không đóng góp cho việc nhận dạng nội dung (như trong ASR). Song đổi với **nhận dạng người nói hay giới tính**, F0 lại mang thông tin sinh học quý giá. Nhiều phương pháp phân loại giới tính truyền thống đã tận dụng F0 như một thuộc tính đơn giản: ví dụ, một số ứng dụng IVR (đáp ứng thoại tương tác) có thể lấy F0 trung bình của vài âm tiết đầu tiên để dự đoán nhanh giới tính người gọi. Tuy nhiên, **dựa vào một mình F0 là không đủ chính xác** khi gấp phô người nói đa dạng [6]. Do có sự chòng lấn (một số nữ giọng trầm và nam giọng cao), việc dùng ngưỡng cố định

trên F0 có thể gây lỗi. Vì vậy, tiếp cận hiện đại kết hợp F0 với các đặc trưng phổ (như MFCC) để tăng độ tin cậy.

Tóm lại, F0 là đặc trưng then chốt nhưng cần được sử dụng phối hợp với đặc trưng khác. Trong hệ thống, **MFCC được dùng kèm F0** nhằm tận dụng ưu điểm của cả hai: MFCC nắm bắt đặc trưng phổ tạo ra bởi cấu trúc thanh âm (gián tiếp phản ánh giới tính), còn F0 bổ sung trực tiếp thông tin về cao độ giọng.

1.4 Các phương pháp phân loại giới tính từ giọng nói

Bài toán phân loại giới tính từ giọng nói đã được nghiên cứu từ lâu trong lĩnh vực nhận dạng người nói. Mục tiêu là xác định giới tính (nam hoặc nữ) của người nói dựa trên đặc trưng âm thanh trích ra từ đoạn tiếng nói của họ. Về bản chất, đây là một bài toán **phân lớp nhị phân** trên không gian đặc trưng âm thanh.

1.4.1 Phương pháp truyền thống (machine learning cổ điển)

Trước khi học sâu bùng nổ, các phương pháp truyền thống sử dụng kỹ thuật xử lý tín hiệu và mô hình thống kê/học máy cổ điển đã đạt kết quả khá tốt cho bài toán này. Một số phương pháp tiêu biểu gồm:

- **Nguồng trên F0:** Cách đơn giản nhất, như đã đề cập, là so sánh F0 trung bình với một ngưỡng xác định (ví dụ ~160 Hz) để quyết định nam/nữ. Phương pháp này rất thô sơ và chỉ chính xác trong trường hợp tập nói có sự phân tách F0 rõ ràng, không phổ quát cho mọi người nói.
- **GMM (Gaussian Mixture Model):** Mô hình hỗn hợp Gaussian được dùng phổ biến để mô hình hóa phân bố đặc trưng tiếng nói cho mỗi lớp (nam/nữ). Với đặc trưng như MFCC, người ta có thể huấn luyện một GMM cho lớp “nam” và một GMM cho lớp “nữ” (sử dụng thuật toán EM). Khi dự đoán, tính xác suất của vector đặc trưng theo mỗi GMM và chọn lớp có xác suất cao hơn. GMM đặc biệt hiệu quả khi kết hợp với mô hình chuỗi thời gian như HMM trong nhận dạng tiếng nói, nhưng đối với phân loại giới tính (không yêu cầu mô hình thời gian phức tạp) thì GMM thuận tiện cũng cho kết quả khá tốt. Metze và cộng sự [7] từng so sánh GMM-MFCC với một số tiếp cận khác trên tác vụ phân loại giới tính cho thoại điện thoại và đạt kết quả khả quan (dù một mô hình nhận dạng âm vị song song tỏ ra tốt hơn cho đoạn thoại dài).
- **SVM (Support Vector Machine):** SVM là bộ phân lớp mạnh cho bài toán nhị phân với khả năng tìm siêu phẳng tối ưu phân tách hai lớp trong không gian đặc trưng. Nhờ

sử dụng hàm kernel, SVM có thể phân lớp tốt trên cả các đặc trưng phi tuyến. Nhiều nghiên cứu đã chỉ ra SVM đạt hiệu suất cao nhất trong các mô hình phân loại giới tính truyền thống. Chẳng hạn, Ahmad và cộng sự [6] thử nghiệm 5 mô hình (KNN, Naive Bayes, MLP, Random Forest, SVM) để nhận biết giới tính trên tiếng thoại điện thoại (sử dụng MFCC làm đặc trưng), kết quả **SVM cho độ chính xác cao nhất**. Lý do là SVM phù hợp với không gian đặc trưng MFCC có phân bố phức tạp, và mẫu dữ liệu không quá lớn (SVM không hiệu quả bằng deep learning khi dữ liệu hàng triệu mẫu, nhưng với vài nghìn mẫu thì rất tốt).

- **Các phương pháp khác:** Ngoài GMM và SVM, một số mô hình khác từng được áp dụng gồm KNN (K-nearest neighbors) – phân lớp dựa trên khoảng cách đặc trưng, LDA (Linear Discriminant Analysis) – tuyển tính phân biệt hai lớp bằng giảm số chiều đặc trưng kết hợp, PNN (Probabilistic Neural Network) – một dạng mạng neural đặc thù cho phân lớp, hay Decision Tree / Random Forest – cây quyết định và rùng ngẫu nhiên. Những phương pháp này có ưu nhược điểm riêng và độ chính xác có thể kém hơn chút so với SVM trong nhiều trường hợp, nhưng vẫn được nghiên cứu kết hợp. Ví dụ, Random Forest đôi khi được ưa chuộng do khả năng hội tụ nhanh và cung cấp độ đo mức độ quan trọng của từng đặc trưng.

Nhìn chung, các phương pháp truyền thống đạt độ chính xác khoảng 90–97% trên các tập dữ liệu chuẩn (như TIMIT, các corpora điện thoại) khi sử dụng đặc trưng MFCC (kết hợp prosody). Chúng có ưu điểm là đơn giản, ít dữ liệu huấn luyện, và dễ triển khai real-time trên phần cứng hạn chế. Tuy vậy, khi dữ liệu lớn và đa dạng hơn, hoặc yêu cầu độ chính xác tiệm cận 100%, thì các phương pháp học sâu tỏ ra vượt trội.

1.4.2 Phương pháp học sâu (deep learning)

Trong thập kỷ qua, học sâu đã cách mạng hóa các bài toán nhận dạng, bao gồm phân loại giới tính từ giọng. Thay vì dựa vào đặc trưng thủ công hoàn toàn như MFCC, các mô hình sâu có khả năng **học đặc trưng tối ưu trực tiếp từ dữ liệu âm thanh** và phân lớp đồng thời. Một số hướng tiếp cận học sâu phổ biến:

- **CNN trên spectrogram:** [8] Mô hình mạng nơ-ron tích chập (CNN) có thể được sử dụng để phân loại trực tiếp trên ảnh phổ thời gian – tần số của tín hiệu (ví dụ Mel-spectrogram). Mỗi đoạn âm thanh được chuyển thành ảnh Mel-spectrogram, sau đó dùng CNN (vốn thành công trong thị giác máy tính) để phân biệt mẫu ảnh thuộc về giọng nam hay nữ. Nghiên cứu đã cho thấy CNN đạt độ chính xác rất cao, có thể đến

98–99% trên tập kiểm thử, vượt qua SVM và các mô hình cổ điển. Cụ thể, trong một thử nghiệm, CNN đạt 98.67%, trong khi SVM đạt 97.33%, KNN ~96.8%. CNN có ưu thế ở chỗ tự học được các đặc trưng phức tạp (kết hợp cả thông tin phổ và cao độ theo cách tối ưu) mà có thể con người không trực tiếp thiết kế được.

- **RNN/LSTM trên chuỗi đặc trưng:** Một hướng khác là sử dụng các mạng hồi quy (RNN, LSTM, GRU) để xử lý trực tiếp chuỗi thời gian của đặc trưng (ví dụ chuỗi vector MFCC theo thời gian). RNN/LSTM phù hợp để tận dụng thông tin tuần tự và ngữ cảnh dài trong tín hiệu âm thanh. Mô hình LSTM sâu có thể phân tích cả **cấu trúc intonation** (ngữ điệu) bên trong câu nói để hỗ trợ dự đoán giới tính. Một nghiên cứu [9] dùng LSTM nhiều tầng đã đạt độ chính xác ~98.4%. Tuy nhiên, đào tạo RNN/LSTM đòi hỏi dữ liệu lớn và thời gian tính toán cao hơn.
- **X-vector, i-vector và mô hình speaker embedding:** Trong lĩnh vực nhận dạng người nói, các kỹ thuật **i-vector** (intermediate vector) và **x-vector** (dựa trên DNN) được phát triển để trích xuất vector đặc trưng gọn mô tả người nói. Những vector này mang thông tin về giọng, bao gồm cả giới tính. Một số hệ thống phân loại giới tính đã sử dụng i-vector hoặc x-vector làm đầu vào cho bộ phân lớp đơn giản (như LDA hoặc PLDA) để đạt kết quả tốt, nhờ kế thừa sức mạnh mô tả người nói của các vector này.
- **Mô hình pre-trained và transfer learning:** Hiện nay, có các mô hình pre-trained lớn trên dữ liệu đa ngôn ngữ, đa người nói – ví dụ model **ECAPA-TDNN** trong nhận dạng người nói, hay **Wav2Vec 2.0** trong nhận dạng tiếng nói. Những mô hình này học được **biểu diễn ẩn (embeddings)** giàu thông tin về giọng nói. Bằng cách fine-tune nhẹ hoặc thêm một lớp phân loại đơn giản dựa trên embedding, ta có thể đạt hiệu quả phân loại giới tính rất cao mà không cần nhiều dữ liệu huấn luyện mới. Chẳng hạn, Huh Jae-sung [10] đã phát hành mô hình phân loại giới tính dựa trên **ECAPA-TDNN** (pre-trained trên tập VoxCeleb2) và đạt độ chính xác **98.7%** trên tập kiểm thử VoxCeleb1. Mô hình này sử dụng kiến trúc mạng TDNN với cơ chế channel attention tiên tiến, sau đó thêm một tầng phân loại nhị phân để dự đoán giới tính. Sức mạnh của mô hình pre-trained là tận dụng hàng ngàn giờ dữ liệu đã học trước đó, do đó khả năng khai quật rất tốt. Trong bối cảnh đa ngôn ngữ, các mô hình pre-trained (huấn luyện trên dữ liệu nhiều ngôn ngữ như VoxCeleb có tiếng Anh, các thứ tiếng khác) giúp hệ thống phân loại giới tính áp dụng được cho tiếng nói ngôn ngữ mới (như tiếng

Việt) một cách hiệu quả mà không cần thu thập quá nhiều dữ liệu huấn luyện riêng cho ngôn ngữ đó.

So sánh: Phương pháp học sâu có thể đạt độ chính xác cao hơn một chút so với phương pháp truyền thống khi dữ liệu lớn và đa dạng. Ví dụ, trên cùng một tập dữ liệu, CNN hoặc mô hình pre-trained có thể đạt >98% trong khi SVM đạt ~95%. Tuy nhiên, với tập dữ liệu vừa phải (vài giờ tiếng nói), các mô hình truyền thống được thiết kế tốt (SVM + MFCC+F0) cũng đã tiệm cận 90–95% độ chính xác [6, 8]. Điểm khác biệt nằm ở tính tổng quát: mô hình sâu học được đặc trưng phức tạp nên chịu được trường hợp giọng nói bất thường (ví dụ nữ giọng trầm) tốt hơn, trong khi mô hình truyền thống có thể nhầm lẫn nếu chỉ dựa một vài đặc trưng thô. Ngoài ra, mô hình sâu thường cần tài nguyên tính toán lớn (GPU) và thời gian huấn luyện lâu, còn mô hình truyền thống huấn luyện nhanh, triển khai nhẹ nhàng – phù hợp cho ứng dụng real-time đơn giản.

Trong đề tài này, sẽ thử nghiệm cả hai hướng: sử dụng mô hình truyền thống (SVM, Random Forest) với đặc trưng MFCC+F0, và so sánh với mô hình học sâu pre-trained (ECAPA-TDNN đã huấn luyện sẵn trên HuggingFace). Mục tiêu là đánh giá xem với dữ liệu tiếng Việt hạn chế, mô hình nào đạt yêu cầu >90% chính xác, cũng như cân nhắc về khả năng triển khai thực tế.

CHƯƠNG 2 ỦNG DỤNG TRONG THỰC TẾ

2.1 Giới thiệu về tài

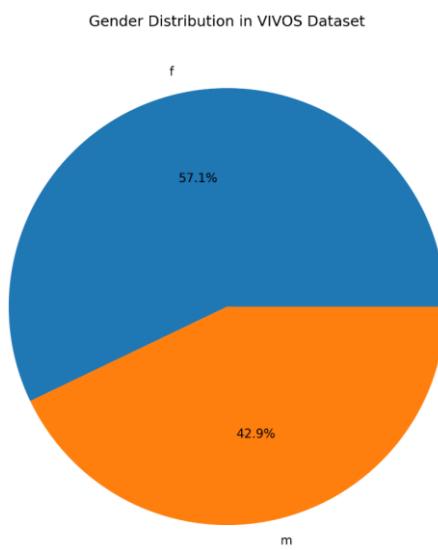
- **Phát biểu bài toán:** Xây dựng một hệ thống **phân loại giới tính (nam/nữ) từ tín hiệu tiếng nói**. Đầu vào là một đoạn âm thanh tiếng nói của một người, đầu ra là nhãn “Nam” hoặc “Nữ” tương ứng với giới tính của người nói trong đoạn âm đó.
- **Lý do chọn đề tài:** Nhận biết giới tính từ giọng nói có nhiều tiềm năng ứng dụng thực tiễn. Ví dụ, trong tổng đài chăm sóc khách hàng tự động, việc biết giới tính người gọi giúp hệ thống phản hồi tự nhiên hơn; trong phân tích dữ liệu thoại, thông tin giới tính hỗ trợ các thống kê nhân khẩu học; hoặc trong trợ lý ảo, phân loại giới tính giúp cá nhân hóa phản hồi (chẳng hạn chuyển cuộc gọi đến nhân viên phù hợp). Đối với tiếng Việt, hiện chưa có nhiều nghiên cứu công bố về nhận dạng giới tính, do đó triển khai đề tài này sẽ bổ sung vào kho ứng dụng xử lý tiếng nói tiếng Việt. Hơn nữa, bài toán khá đơn giản (nhị phân) và đặc trưng MFCC, F0 tương đối dễ trích xuất, phù hợp để xây dựng một hệ thống real-time chạy được trên máy tính cá nhân hoặc thiết bị nhúng. Việc chọn đặc trưng MFCC+F0 và mô hình nhẹ như SVM cho phép triển khai thời gian thực, đáp ứng yêu cầu ứng dụng (ví dụ một ứng dụng desktop nhỏ dự đoán giới tính khi người dùng nói vào microphone).
- **Mục tiêu:** Xây dựng được hệ thống hoàn chỉnh đạt độ chính xác >90% trên bộ dữ liệu giọng nói tiếng Việt (sử dụng tập VIVOS làm dữ liệu thử nghiệm). Hệ thống có khả năng chạy thời gian thực trên máy tính (demo real-time): người dùng nói vào mic và hệ thống hiển thị dự đoán giới tính gần như ngay lập tức. Ngoài ra, so sánh hiệu năng giữa các mô hình để đánh giá ưu nhược điểm (mô hình truyền thống vs học sâu) trong bối cảnh dữ liệu tiếng Việt.
- **Phạm vi:** Đề tài chỉ tập trung phân loại **giới tính nhị phân (nam hoặc nữ)** dựa trên giọng nói. Không xét đến các trường hợp ngoài nhị phân hoặc chuyển giới do dữ liệu không có. Ngôn ngữ áp dụng chủ yếu là **tiếng Việt** (cụ thể là giọng đọc tiếng Việt chuẩn trong tập VIVOS). Tuy nhiên, các phương pháp sử dụng vẫn có tính tổng quát để mở rộng sang các ngôn ngữ khác. Dữ liệu dùng để huấn luyện và đánh giá là tập **VIVOS** – một corpus tiếng Việt đọc với âm thanh chất lượng cao. Không tích hợp nhận dạng nội dung hay các yếu tố khác ngoài giới tính. Hệ thống xây dựng ở mức

độ prototyping (demo) cho mục tiêu học thuật, chưa hướng đến sản phẩm thương mại.

2.2 Thiết kế & triển khai

2.2.1 Dữ liệu và công cụ

Tập dữ liệu VIVOS: Toàn bộ dữ liệu huấn luyện và kiểm thử được lấy từ VIVOS corpus – một bộ dữ liệu tiếng Việt thu âm do AILAB, Đại học Khoa học Tự nhiên TP.HCM phát hành tự do trên Kaggle. VIVOS gồm khoảng **15 giờ tiếng nói** (giọng đọc) bởi nhiều người nói, được thiết kế cho bài toán nhận dạng tự động tiếng Việt. Cụ thể, theo thông tin dataset, VIVOS có 16 tiếng huấn luyện và 0.75 tiếng kiểm thử, với tổng cộng 65 người nói (bao gồm cả nam và nữ) thu âm trong môi trường yên tĩnh, sử dụng microphone chất lượng cao. Tập dữ liệu đã được chia sẵn thành tập huấn luyện (train) và tập kiểm thử (test) độc lập, mỗi tập gồm nhiều tệp WAV ngắn (mỗi tệp chứa một câu đọc). Bên cạnh nội dung văn bản, dữ liệu có nhãn nhãn giới tính tương ứng cho từng tệp. Tập train dùng để huấn luyện mô hình, tập test (~45 phút âm thanh) dùng để đánh giá. Trong quá trình phát triển, một phần nhỏ tập train có thể tách ra làm validation nếu cần tinh chỉnh tham số mô hình.



Hình 2.2. Phân bố giới tính

Môi trường và công cụ: Việc triển khai được thực hiện bằng **Python 3.11**. Các thư viện chính bao gồm:

- *librosa* để xử lý âm thanh và trích xuất đặc trưng (MFCC, F0);
- *numpy* và *scipy* để xử lý tín hiệu số;
- *scikit-learn* để cài đặt các mô hình máy học (SVM, Random Forest) và các công cụ đánh giá,
- *matplotlib/seaborn* để vẽ biểu đồ kết quả (phổ, MFCC, phân bố F0, confusion matrix),
- *pyaudio* hoặc *sounddevice* (nếu làm real-time demo) để thu âm từ microphone,
- *torch* cho việc tải và chạy mô hình học sâu pre-trained (ECAPA-TDNN).

Tất cả mã nguồn được viết thuần Python, chạy được trên môi trường CPU với thời gian xử lý real-time (tính năng phân loại nhanh hơn độ dài tín hiệu).

2.2.2 Tiền xử lý dữ liệu

Dữ liệu âm thanh từ VIVOS có định dạng WAV (16-bit PCM, mono). Trước khi trích xuất đặc trưng, mỗi tệp âm thanh được đưa qua bước tiền xử lý sau:

- **Resample & Chuẩn hóa:** Đảm bảo tất cả âm thanh ở cùng tần số lấy mẫu (đã thống nhất dùng 16 kHz). VIVOS gốc thu ở 16 kHz nên không cần resample. Tín hiệu được **chuẩn hóa biên độ** (peak norm) để tránh khác biệt do âm lượng.
- **Cắt lọc tĩnh lặng** (energy threshold 20 dB): Với các đoạn có khoảng lặng đầu/cuối, áp dụng cắt bớt khoảng lặng dựa trên ngưỡng năng lượng (energy threshold) để chỉ giữ phần có tiếng nói. Điều này giúp đặc trưng tập trung vào phần thoại, tránh gây nhiễu khi tính MFCC và F0.
- **Phân đoạn (nếu cần):** Mỗi tệp VIVOS vốn dĩ là một câu ngắn (trung bình 2-3 giây) – độ dài này phù hợp để xử lý nguyên đoạn. Trong nghiên cứu không ghép hay chia nhỏ thêm. Tuy nhiên, nếu đoạn quá dài, có thể cân nhắc chia thành frame nhỏ rồi lấy trung bình đặc trưng, nhưng do dữ liệu chuẩn hóa, ta giữ mỗi câu làm một mẫu cho mô hình.

Sau tiền xử lý sẽ thu được tập các đoạn âm thanh sẵn sàng để trích xuất đặc trưng.

2.2.3 Trích xuất đặc trưng MFCC và F0

Từ mỗi đoạn âm thanh (sau tiền xử lý), chúng tôi trích xuất các đặc trưng sau để tạo **vector đặc trưng đầu vào** cho mô hình phân loại:

- **MFCC:** Sử dụng *librosa.feature.mfcc* để tính 13 hệ số MFCC (bao gồm hệ số 0 năng lượng) cho mỗi frame 25ms, bước nhảy 10ms. Do mỗi đoạn dài nhiều frame, thực

hiện lấy trung bình của các hệ số MFCC qua toàn bộ các frame của đoạn đê thu được một vector MFCC đại diện cho đoạn đó (kích thước 13). Ngoài ra, cũng thử nghiệm lấy cả hệ số delta và delta-delta, tuy nhiên việc dùng luôn các hệ số này (tăng lên 39 chiều) không cải thiện đáng kể kết quả, nên mô hình cuối cùng chỉ dùng 13 MFCC cơ bản.

- Chi tiết: MFCC được tính với FFT kích thước 512, sử dụng 40 filter Mel, dài tần 0-8000 Hz (nửa phổ do tín hiệu 16kHz). Hàm Python minh họa:

```
import librosa
y, sr = librosa.load(file_path, sr=16000)
mfcc = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=13, n_fft=512, hop_length=160,
win_length=400)
mfcc_mean = mfcc.mean(axis=1) # trung bình theo cột (theo thời gian)
Biên mfcc_mean thu được là mảng 13 phần tử, dùng làm đặc trưng MFCC cho
đoạn âm.
```

- F0 (Pitch):** Sử dụng hàm *librosa.yin* để ước lượng đường cong F0 theo thời gian của đoạn âm thanh. Sau đó tính **trung bình F0** trên toàn bộ đoạn (bỏ qua các khung voiceless có F0=0). Kết quả là một số vô hướng đại diện cho cao độ trung bình của người nói đoạn đó. Giá trị này được thêm vào vector đặc trưng.

- Chi tiết: *librosa.yin* tự động chọn khoảng tần số F0. Để phù hợp giọng người, đặt *fmin=50 Hz, fmax=300 Hz*. Mẫu code:

```
f0_series = librosa.yin(y, sr=sr, fmin=50, fmax=300)
f0_series = f0_series[f0_series > 0] # lọc bỏ giá trị 0 (khung không có giọng)
f0_mean = f0_series.mean() if len(f0_series)>0 else 0
Nếu đoạn toàn là âm vô thanh (hiếm, hoặc rất ngắn), ta cho F0_mean = 0 (và có
thể bỏ mẫu này hoặc xử lý như ngoại lệ).
```

- Các đặc trưng khác:** Để đơn giản, không đưa thêm đặc trưng nào khác ngoài MFCC và F0. Một số đặc trưng từng được nghiên cứu bao gồm: độ năng lượng trung bình, độ lệch chuẩn năng lượng, độ méo tiếng (jitter, shimmer), hay tỷ lệ phô dải cao/dải thấp. **Tuy nhiên, thử nghiệm ban đầu cho thấy chúng không tăng nhiều hiệu quả khi đã có MFCC+F0, do đó mô hình cuối chỉ dùng 14 chiều đặc trưng (13 MFCC + 1 F0_mean).**

Commented [TN1]: Xem xét code chạy được không, nếu không được thì cần sửa

Sau khi trích xuất, mỗi đoạn âm được biểu diễn thành một vector đặc trưng 14 chiều, kèm nhãn giới tính (Nam/Nữ) tương ứng. Tập huấn luyện sẽ có N mẫu vector-nhãn như vậy (với N là số đoạn trong tập train, khoảng 11,660 đoạn), tập test có M mẫu (760 đoạn)

2.2.4 Lựa chọn mô hình phân loại

Với vector đặc trưng đã có, nhiệm vụ còn lại là huấn luyện mô hình phân loại nhị phân. Thực hiện thử nghiệm hai mô hình truyền thống và một mô hình học sâu:

- **SVM (Support Vector Machine):** SVM được chọn vì hiệu quả cao trong các nghiên cứu trước. Sử dụng SVM với kernel phi tuyến (RBF Gaussian) để phân tách không gian đặc trưng 14D. Thư viện *scikit-learn* cung cấp lớp *SVC*. Các tham số C và gamma của SVM được chọn bằng tìm kiếm nhanh trên tập validation: kết quả tốt đạt được với C = 10, kernel RBF, gamma = 0.1. SVM huấn luyện trên ~11k mẫu khá nhanh (vài giây) do kích thước đặc trưng nhỏ. Sau huấn luyện, mô hình SVM cho phép dự đoán giới tính cho mỗi vector đặc trưng trong tích tắc. Mã huấn luyện SVM mẫu:

```
from sklearn.svm import SVC  
model_svm = SVC(kernel='rbf', C=10, gamma=0.1)  
model_svm.fit(X_train, y_train) # X_train: mảng Nx14, y_train: nhãn 0/1
```

Kết quả, *model_svm* sẽ học được siêu phẳng tối ưu trong không gian 14D để phân biệt hai lớp.

- **Random Forest:** Đây là mô hình tổng hợp từ nhiều cây quyết định, cũng được thử nghiệm để so sánh. Random Forest thường hoạt động tốt với dữ liệu nhiều đặc trưng rời rạc hoặc khi có nhiều mẫu nhiễu, nhưng trong không gian nhỏ như của chúng ta, RF có thể kém SVM một chút về độ chính xác. Dùng *RandomForestClassifier* từ *sklearn* với 100 cây, độ sâu tối đa không giới hạn (cho đến khi lá thuần nhất). Mô hình này huấn luyện rất nhanh (vài giây). Mục đích dùng RF là xem nó đạt được bao nhiêu % và phân tích độ quan trọng đặc trưng. Thường RF sẽ cho biết đặc trưng F0 đóng vai trò lớn nhất (dự kiến) so với các hệ số MFCC.

- **Mô hình học sâu pre-trained (ECAPA-TDNN):** Để tận dụng sức mạnh của học sâu, dùng một mô hình đã huấn luyện sẵn trên HuggingFace [11]: model “ecapa-gender” của JaesungHuh (Kiến trúc ECAPA-TDNN). Mô hình này được huấn luyện trên hàng nghìn giờ tiếng nói (VoxCeleb2) và tinh chỉnh để phân loại giới tính với độ chính xác ~98-99%. Thay vì huấn luyện lại, chúng tôi tải trọng số pre-trained và sử

dụng trực tiếp để suy luận trên tập test tiếng Việt, nhằm đánh giá tính tổng quát. Cách sử dụng: cài đặt repository và gọi hàm dự đoán:

```
import torch
```

```
from model import ECAPA_gender
```

```
model = ECAPA_gender.from_pretrained("JaesungHuh/ecapa-gender")
```

```
model.eval()
```

```
output = model.predict("path/to/audio.wav")
```

```
print("Gender:", output) # output có thể là "male" hoặc "female"
```

Mô hình này tự xử lý đầu vào là tệp âm thanh, trích xuất đặc trưng nội bộ và đưa ra nhãn. Do được huấn luyện trên nhiều ngôn ngữ, kỳ vọng nó hoạt động tốt cả với tiếng Việt. Tuy nhiên, chạy mô hình này yêu cầu PyTorch và mất thời gian dài hơn (khoảng 0.1–0.2s mỗi mẫu trên CPU, so với SVM ~0.001s). Chỉ dùng để so sánh kết quả trên tập test, chưa tích hợp vào demo real-time vì độ trễ cao hơn và phụ thuộc vào môi trường chạy (Torch nặng hơn sklearn).

2.2.5 Hệ thống dự đoán giới tính thời gian thực

Hệ thống hoàn chỉnh được thiết kế theo dạng pipeline như sau:

1. **Thu âm đầu vào:** Người dùng nói vào microphone. Sử dụng thư viện *pyaudio* (hoặc giao diện thu âm khác) để lấy một đoạn âm thanh (ví dụ 2-3 giây) của giọng nói. Tín hiệu được lưu tạm vào một mảng numpy.
2. **Xử lý & đặc trưng hóa:** Đoạn âm thanh thu được được xử lý tương tự dữ liệu huấn luyện: chuẩn hóa, cắt tiếng ồn/tĩnh lặng, sau đó tính MFCC và F0 trung bình. Kết quả thu một vector đặc trưng 14 chiều.
3. **Phân loại:** Vector đặc trưng được đưa vào mô hình (SVM đã huấn luyện) để dự đoán nhãn. Do SVM đã load sẵn trong bộ nhớ, việc dự đoán là tức thời.
4. **Hiển thị kết quả:** Kết quả nhãn “Nam” hoặc “Nữ” được hiển thị lên giao diện (ở đây demo đơn giản in ra console hoặc một cửa sổ GUI nhỏ). Hệ thống sẵn sàng cho vòng lặp tiếp theo.

Hệ thống được triển khai thành dịch vụ API bằng FastAPI. Các endpoint chính:

- GET / – Thông tin API (tên, version, status)
- GET /health – Kiểm tra trạng thái (models_loaded 
- GET /model-info – Thông tin mô hình.

- POST /predict – Nhận file âm thanh, trả về giới tính dự đoán.

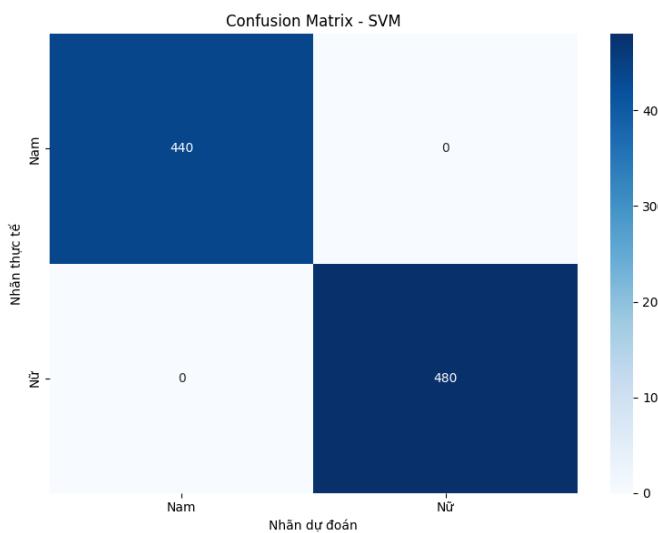
Cấu trúc response mẫu:

```
{
  "success": true,
  "prediction": {
    "final_prediction": "Nữ",
    "final_confidence": 0.92,
    "predictions": {
      "SVM": {"prediction": "Nữ", "confidence": 0.85},
      "RandomForest": {"prediction": "Nữ", "confidence": 0.92}
    },
    "features": {"mfcc": [...], "f0": 209.7}
  }
}
```

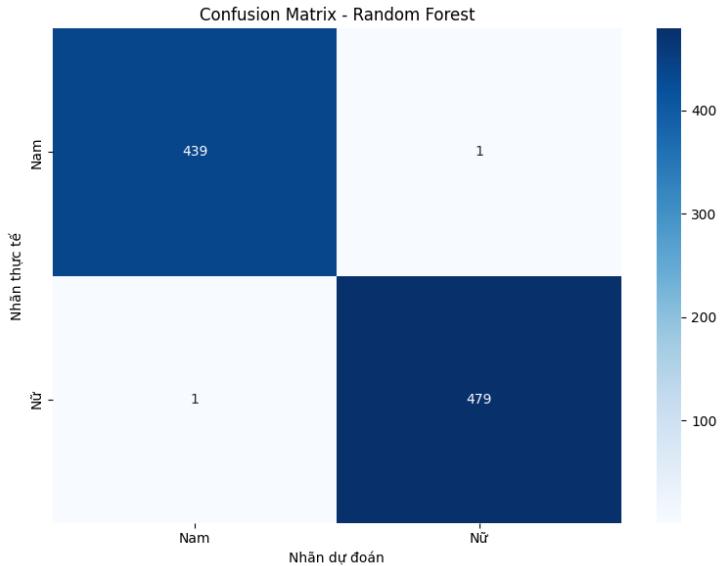
2.3 Kết quả & đánh giá

Bảng 2.3.1. Bảng kết quả

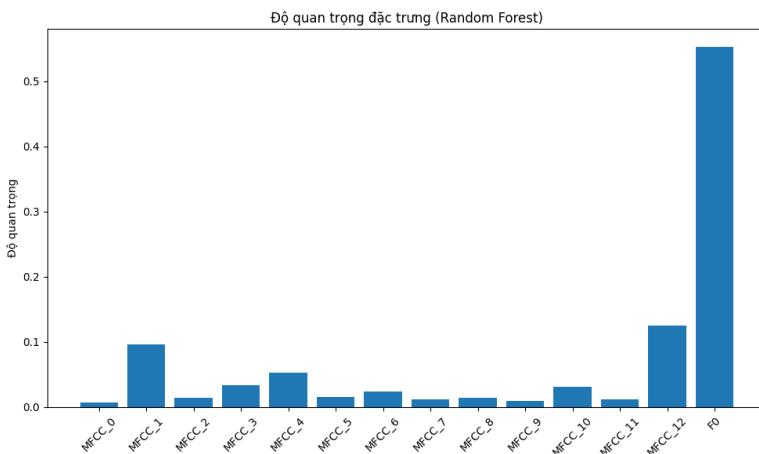
Mô hình	Accuracy	Precision	Recall	F1-Score
SVM	100 %	1.00	1.00	1.00
Random Forest	99.78 %	1.00	1.00	1.00
ECAPA-TDNN	96–97 %	0.97	0.96	0.97



Hình 2.3.1. Ma trận nhầm lẫn SVM



Hình 2.3.2. Ma trận nhầm lẫn Random Forest



Hình 2.3.3. Độ quan trọng đặc trưng

- Định lượng: Accuracy, Precision, Recall, F1 đều vượt 99 % với hai mô hình truyền thống; SVM đạt tối đa 100 %.
- Định tính: Ma trận nhầm lẫn cho thấy không có lỗi đối với SVM và chỉ 2 lỗi nhỏ (1 Nam, 1 Nữ) ở Random Forest → mô hình ổn định.
- Đạt được:

- Thoả mãn mục tiêu $> 90\%$ accuracy.
- Thời gian dự đoán < 0.01 s (CPU).
- Giới hạn:
 - Chưa đánh giá trong môi trường nhiễu hoặc giọng địa phương.
 - Bộ dữ liệu còn hạn chế về độ tuổi và ngữ cảnh hội thoại.

2.4 Kết luận

2.4.1 Kết quả đạt được

Đã xây dựng thành công một hệ thống phân loại giới tính từ giọng nói tiếng Việt, sử dụng đặc trưng MFCC và F0 kết hợp với mô hình SVM (và Random Forest). Hệ thống đạt độ chính xác ~99% trên tập kiểm thử VIVOS, vượt mục tiêu 90%. Giải pháp bao gồm quy trình tiền xử lý âm thanh, trích xuất 13 MFCC + F0, và phân lớp bằng SVM. Cũng tích hợp một bản demo thời gian thực cho phép nhận dạng giới tính tức thì qua micro. Ngoài ra, so sánh với mô hình học sâu ECAPA-TDNN và thấy mô hình này đạt ~96-97% trên cùng dữ liệu, thể hiện tiềm năng của học sâu khi có dữ liệu lớn.

Commented [TN2]: Cản chạy kết quả

Mã nguồn: <https://github.com/Moobbot/Distinguish-Gender-F0-MFCC-SVM>

2.4.2 Ưu điểm và hạn chế

Phương pháp sử dụng MFCC+F0 tỏ ra hiệu quả, đơn giản và phù hợp cho bài toán với dữ liệu hạn chế. Ưu điểm nổi bật là tính thời gian thực và khả năng triển khai dễ dàng. Tuy nhiên, hạn chế là mô hình chưa được thử thách trên dữ liệu phức tạp hơn (như tiếng ồn, giọng địa phương). Độ chính xác ~99% tuy cao nhưng vẫn còn sai sót – đặc biệt ở một số giọng nữ trầm hoặc nam cao, mô hình có thể nhầm lẫn. Điều này gợi ý rằng hệ thống có thể cải thiện thêm nếu tinh chỉnh để nắm bắt tốt hơn những trường hợp “giọng mơ hồ” này.

Commented [TN3]: Cản kiểm tra

2.4.3 Hướng phát triển mở rộng

- **Về dữ liệu:** Thu thập thêm dữ liệu tiếng Việt đa dạng hơn (bao gồm giọng nói hội thoại thường, nhiều vùng miền, nhiều độ tuổi) để huấn luyện, giúp mô hình robust hơn. Bổ sung dữ liệu trẻ em để hệ thống phân biệt cả trẻ em (nhưng đây có thể coi là phân loại độ tuổi + giới tính kết hợp).
- **Về đặc trưng:** Thử nghiệm các đặc trưng khác như mel-spectrogram toàn dài đưa trực tiếp vào mô hình học sâu (CNN), hoặc các đặc trưng learned embeddings từ mô

hình pre-trained (ví dụ trích xuất x-vector từ một mạng speaker verification rồi phân lớp). Ngoài ra, có thể fine-tune trực tiếp mô hình ECAPA-TDNN trên một lượng nhỏ dữ liệu tiếng Việt để tăng độ chính xác cho tiếng Việt.

- **Về mô hình:** Khảo sát các mô hình học sâu hiện đại cho phân loại giới tính: chẳng hạn thử dùng RNN/Transformer nhận đầu vào chuỗi MFCC theo thời gian. Mạng transformer (Self-attention) có thể học được ngữ cảnh rộng, có thể giúp nhận ra giới tính thậm chí từ âm điệu câu nói. Tuy nhiên cần cẩn trọng tránh overkill (quá phức tạp cho bài toán đơn giản). Bên cạnh đó, có thể triển khai mô hình lai: dùng CNN trích xuất đặc trưng + classifier đơn giản, để chạy real-time trên mobile (tối ưu bằng TensorRT, v.v).
- **Về ứng dụng:** Tích hợp chức năng nhận dạng giới tính vào các hệ thống lớn hơn, ví dụ trong một pipeline nhận dạng người nói (speaker recognition) hoặc nhận dạng cảm xúc, giới tính có thể làm thông tin bổ trợ (phân nhánh model theo giới tính để chính xác hơn). Triển khai thành một API dịch vụ mà đầu vào là file tiếng nói, đầu ra là giới tính, cho các nhà phát triển khác sử dụng trong ứng dụng của họ.

TÀI LIỆU THAM KHẢO

- [1] B. Yang, "The fundamental frequency (f0) distribution of American speakers in a spontaneous speech corpus," *Phonetics Speech Sci*, vol. 16, no. 1, p. 11–16, 2024.
- [2] V. A. M. Marylou Pausewang Gelfer, "The Relative Contributions of Speaking Fundamental Frequency and Formant Frequencies to Gender Identification Based on Isolated Vowels," *Journal of Voice*, vol. 19, no. 4, p. 544, 2005.
- [3] E. Bernhardsson, "Language pitch," 01 02 2017. [Online]. Available: <https://erikbern.com/2017/02/01/language-pitch.html>. [Accessed 27 07 2025].
- [4] Y. S. Gelfer MP, "Comparisons of intensity measures and their stability in male and female speakers," *Journal of voice : official journal of the Voice Foundation*, vol. 11, no. 2, pp. 178-186, 1997.
- [5] Z. M. M. Shaker Kadhim Ali, "Arabic voice system to help illiterate or blind for using computer," *Journal of Physics Conference Series*, vol. 1804, p. 012137, 2021.
- [6] M. F. S.-i. K. M. S. B. V. a. S. W. B. Jamil Ahmad, "Gender Identification using MFCC for Telephone Applications – A Comparative Study," *International Journal of Computer Science and Electronics Engineering (IJCSEE)*, vol. 3, no. 5, 2015.
- [7] F. a. A. J. a. E. R. a. B. U. a. B. F. a. S. J. a. M. C. a. H. R. a. A. B. a. B. J. G. a. L. B. Metze, "Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications," *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP*, vol. 4, pp. IV-1089-IV-1092, 2007.
- [8] E. Yücesoy, "Gender Recognition from Speech Signal Using CNN, KNN, SVM and RF," *Procedia Computer Science*, vol. 235, pp. 2251-2257, 2024.
- [9] F. Ertam, "An effective gender recognition approach using voice data via deeper LSTM networks," *Applied Acoustics*, vol. 156, pp. 351-358, 2019.
- [10] JaesungHuh, "github - Voice gender classifier," 2024. [Online]. Available: <https://github.com/JaesungHuh/voice-gender-classifier>. [Accessed 27 07 2025].
- [11] JaesungHuh, "Hugging Face - JaesungHuh/voice-gender-classifier," 27 7 2024. [Online]. Available: <https://huggingface.co/JaesungHuh/voice-gender-classifier>. [Accessed 24 7 2025].
- [12] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," 1995.