**SINH TRẮC HỌC GIỌNG NÓI - Chaper 8: VOICE BIOMETRICS - Handbook of Biometrics**

**1. Giới thiệu**

Trong thời gian gần đây, dữ liệu về người dùng điện thoại di động trên toàn cầu, số lượng điện thoại cố định đang trong tình trạng hoạt động, việc triển khai VoIP (Voice over IP networks - Mạng hội thoại IP), cho thấy rằng giọng nói (voice) là một đặc điểm sinh trắc học (nhân trắc học) dễ dàng tiếp cận nhất mà không cần phải có thêm thiết bị thu nhận và hệ thống truyền dẫn.

Những dữ kiện trên cho thấy giọng nói một lợi thế so với những đặc điểm sinh trắc học (nhân trắc học) khác, đặc biệt là khi người dung hoặc các hệ thống điều khiển từ xa được tính đến. Tuy nhiên, đặc điểm giọng nói không chỉ liên quan đến các đặc trưng cá thể mà còn liên quan với môi trường xung quanh và vấn đề xã hội, do vậy việc sản sinh giọng nói là một kết quả của một quá trình hết sức phức tạp.

Vì thế, việc truyền giọng nói sẽ làm mất đi nhiều đặc trưng của giọng nói tùy thuộc vào đặc điểm của giọng nói và sẽ bị ảnh hưởng bởi ngữ cảnh, gây khó khăn trong viêc xử lý. May mắn thay, các công nghệ và ứng dụng tiên tiến có thể khắc phụ những vấn đề trên, cho phép cải thiện độ hiệu quả và tin cậy cho việc xác nhận từ xa hoặc nhận diện giọng nói chỉ dựa vào tín hiệu giọng nói truyền qua sóng điện thoại [39], [16]

1.1 **Ứng dụng**

Do tính phổ biến của tín hiệu giọng nói, phạm vi ứng dụng có thể có của sinh trắc học (nhân trắc học) giọng nói rộng hơn so với các đặc điểm sinh trắc học thông thường khác. Chúng ta có thể phân biệt ba loại ứng dụng chính tận dụng thông tin sinh trắc học có trong tín hiệu giọng nói như sau:

- **Voice authentication** (Xác nhận giọng nói) (Access control - Điều khiển truy cập, thường là điều khiển từ xa bằng điện thoại) và back-ground recognition (Nhận dạng lý lịch) (natural voice checking - kiểm tra bằng giọng nói tự nhiên) [11]

- **Speaker detection** (Nhận diện người nói) (ví dụ như: phát hiện danh sách đen trong các tổng đài điện thoại hoặc trong nghe lén và giám sát, …), hay còn được gọi là speaker spotting

- **Forensic speaker recognition** (Nhận dạng người nói trong Pháp y) (sử dụng giọng nói làm bằng chứng trước tòa án hoặc làm thông tin tình báo trong các cuộc điều tra của cảnh sát hình sự) [42]

**1.2 Công nghệ**

Nguồn thông tin chính được mã hóa trong tín hiệu thoại chắc chắn là nội dung ngôn ngữ. Vì lý do đó, không có gì ngạc nhiên khi tùy thuộc vào cách sử dụng hoặc kiểm soát nội dung ngôn ngữ, chúng ta có thể phân biệt hai loại công nghệ nhận dạng người nói rất khác nhau với các ứng dụng tiềm năng khác nhau.

**1.2.1 Text-dependent**

Đầu tiên, công nghệ text-dependent, trong đó user được yêu cầu phải nói ra một từ khóa (Ví dụ: “Open, Sesame”, “Vừng ơi! Mở ra!”) hoặc chuỗi (Ví dụ: “12-34-56”), đã trở thành chủ đề chính của nhân trắc học điều khiển truy cập và ứng dụng xác minh giọng nói [38], [16]. Mức độ bảo mật của mật khẩu hệ thống có thể được cường hóa bằng việc yêu cầu kiến thức về mật khẩu và cũng yêu cầu chủ sở hữu thực sự của mật khẩu phải nói ra. Để mà tránh việc ghi trộm mật khẩu thực có thể xảy ra, các hệ thống Text-dependent có thể được cải tiến để yêu cầu các lời nhắc ngẫu nhiên, không mong đợi đối với người gọi, điều này không thể dễ dàng bị kẻ mạo danh tạo ra.

**1.2.1 Text-independent**

Loại thứ hai của công nghệ nhận dạng người nói được biết như là Text-independent. Nó là nhân tố thúc đẩy sự phát triển của hai loại ứng dụng còn lại, cụ thể là **Speaker detection** (Nhận diện người nói) và **Forensic speaker recognition** (Nhận dạng người nói trong Pháp y). Vì nội dung ngôn ngữ là nguồn thông tin chính được mã hóa trong lời thoại, tính độc lập với văn bản đã là một thách thức lớn và là chủ đề nghiên cứu chính của cộng đồng Nhận dạng người nói trong suốt hai thập kỷ qua. NISTSRE (Speaker Recognition Evaluations) được tiến hành hàng năm kể từ năm 1996 [35], [39] đã thúc đẩy sự xuất sắc trong nghiên cứu trong lĩnh vực này, với tiến bộ phi thường đạt được qua từng năm dựa trên đánh giá mờ với các cơ sở dữ liệu và giao thức chung, và đặc biệt là việc chia sẻ thông tin giữa những người tham gia hội thảo tiếp theo sau mỗi lần phát triển.

**2. Thông tin nhận dạng trong tín hiệu giọng nói**

Sự tạo thành giọng nói là quá trình cực kỳ phức tạp mà kết quả của nó phụ thuộc vào nhiều tham biến ở nhiều cấp độ khác nhau, bao gồm từ các yếu tố xã hội học (Ví dụ như: trình độ học vấn, ngữ cảnh nôn ngữ, sự khác biệt vùng miền) cho đến các vấn đề sinh lý (Ví dụ như: chiều dài đường thanh âm, hình dạng và mô và cấu trúc động của các cơ quan cấu âm). Những ảnh hưởng này sẽ xuất hiện đồng thời trong mỗi hành động lời nói và một số hoặc tất cả chúng sẽ chứa đựng những đặc điểm cụ thể của giọng nói. Vì lý do đó, chúng ta cần làm rõ và phân biệt rõ ràng các cấp độ và nguồn thông tin giọng nói khác nhau mà chúng ta có thể trích xuất để mô hình hóa tính đặc biệt của giọng nói

**2.1 Sự hình thành ngôn ngữ và hình thành lời nói**

Quá trình mà con người có thể xây dựng một thông điệp được mã hóa bằng ngôn ngữ đã là một chủ đề nghiên cứu trong nhiều năm trong lĩnh vực Ngôn ngữ học tâm lý (Psycholinguistics). Nhưng một khi thông điệp đã được mã hóa trong não người, vẫn cần một quá trình sinh lý và ngữ âm khớp phức tạp để cuối cùng tạo ra một dạng sóng lời nói (giọng nói) chứa thông điệp ngôn ngữ (cũng như nhiều nguồn thông tin khác, một trong số đó là nhận dạng người nói) được mã hóa như một sự kết hợp của các đặc điểm phổ thời gian. Quá trình này là chủ đề nghiên cứu của các nhà ngữ âm học và một số lĩnh vực liên quan đến phân tích giọng nói khác

(kỹ sư, bác sĩ, v.v.). Thông tin chi tiết về quá trình tạo ngôn ngữ và tạo giọng nói có thể xem trong [50], [27], [41]. Quá trình tạo giọng nói rất phức tạp và bản thân nó sẽ xứng đáng có vài chương sách, nhưng chúng ta ở đây quan tâm đến những khía cạnh liên quan đến việc mã hóa một số loại thông tin riêng lẻ trong tín hiệu lời nói cuối cùng được truyền ra khỏi miệng người nói. Trong cả hai giai đoạn của quá trình tạo giọng nói (tạo ngôn ngữ và tạo giọng nói), các đặc điểm cụ thể của người nói đều được giới thiệu. Trong lĩnh vực sinh trắc học giọng nói - còn được gọi là nhận dạng người nói - hai thành phần này tương ứng với nó thường được gọi là đặc điểm cấp cao (ngôn ngữ) và cấp thấp (âm thanh).

**2.2 Thông tin đa cấp - Multiple information levels**

Các thí nghiệm với những thính giả cho thấy rằng, như kinh nghiệm của chúng ta cho chính chúng ta biết, con người nhận biết giọng nói bằng sự kết hợp của nhiều mức thông tin khác nhau, và điều gì là quan trọng, với những trọng số khác nhau đối những người nói khác nhau (ví dụ: một giọng nói có thể hiển thị các đường viền cao độ rất đặc trưng, và một cái khác có thể có một giọng mũi mạnh làm cho chúng "âm thanh" khác nhau). Các hệ thống tự động sẽ tận dụng nguồn tin khác nhau sẵn có, kết hợp chúng lại theo một cách tốt nhất có thể cho giọng nói

**Idiolectal characteristics - Đặc trưng Idiolectal**

Đặc trưng Idiolectal của một người nói nằm ở cấp độ cao nhất tính đến thời điểm hiện tại mà thường được đưa vào các công nghệ, mô tả người nói “sử dụng” một hệ thống ngôn ngữ cụ thể. Từ “sử dụng” được xác định bởi nhiều yếu tố, một số chúng khá ổn định như trình độ học vấn, điều kiện xã hội học và gia đình và nơi cư trú. Nhưng cũng có một số dữ kiện cấp cao phụ thuộc mạnh vào điều kiện môi trường ví dụ như một bác sĩ nam không sử dụng ngôn ngữ theo cách tương tự khi nói chuyện với đồng nghiệp của anh ấy tại bệnh viện (xã giao), với gia đình anh ấy ở nhà, hoặc với bạn bè của anh ấy chơi bài.

**Phonotactics characteristics - Đặc trưng ngữ âm**

Như một đặc trưng chính thứ hai, theo chiều giảm cấp độ thông tin trong tín hiệu giọng nói, chúng ta nhận thấy “phonotactics” (ngữ âm, âm vị), trong đó mô tả cách sử dụng của người nói của các đơn vị điện thoại và khả năng nhận ra khả dụng. Ngữ âm là rất cần thiết cho việc sử dụng đúng đắn về một ngôn ngữ, và một chìa khóa trong việc học ngoại ngữ, nhưng khi chúng ta nhìn vào nét đặc trưng phonotactic người nói, cchúng ta có thể tìm thấy một số kiểu sử dụng khác biệt với những người dùng khác

**Prosody characteristics - Đặc trưng ngữ điệu**

Trong nhóm thứ ba, chúng ta thấy prosody (ngữ điệu), là sự kết hợp của năng lượng tức thời, âm điệu, tốc độ nói và thời lượng đơn vị để cung cấp cho lời nói sự tự nhiên, đầy đủ ý nghĩa và giọng điệu cảm xúc. Prosody (ngữ điệu) xác định các đối tượng âm điệu ở mức độ cụm và đoạn đàm thoại, và định nghĩa nhanh chóng hành động để phù hợp với các đối tượng trên. Nó giúp làm rõ thông điệp (“nine hundred twenty seven” có thể phân biệt “927” hoặc “900 27” bởi ý nghĩa của ngữ điệu), loại thông điệp (trần thuật, nghi vấn, mệnh lệnh), hoặc trạng thái suy nghĩ của người nói. Những theo cách mà mỗi người nói sử dụng các thành phần ngữ điệu khác nhau, nhiều đặc trưng của người bị trùng lặp, ví dụ như các đường đặc trưng bị trùng nhau ở điểm đầu và điểm của của một cụm hoặc nhóm giọng nói

**Spectral characteristics - Đặc trưng phổ**

Ở cấp độ thấp hơn, chúng ta có các đặc trưng phổ rời rạc của tín hiệu giọng nói, liên quan trực tiếp với hành động phát âm đơn lẻ, quan hệ với tạo ra ngữ âm và cũng liên quan đến sinh lý cá nhân của quá trình phát sinh giọng nói. Thông tin về phổ là trọng tâm của sự phân biệt giọng nói, thường được dùng trong các ứng dụng, và là nghiên cứu trọng tâm trong hàng 20 năm qua [43], [54] [8]. Thông tin phổ chú trọng vào rút trích các đặc điểm trong giọng nói và động lực phát âm tương ứng của người nói. Hai loại của thông tin cấp thấp thường được dùng, thông tin tĩnh có quan hệ với phân tích từng frame và thông tin động có quan hệ với cách mà thông tin phát triển trong các frame liền kề, có tính đến hiện tượng nối âm, phụ thuộc nhiều vào người nói, quá trình mà một cá nhân tự động di chuyển từ vị trí phát âm này sang vị trí phát âm tiếp theo.

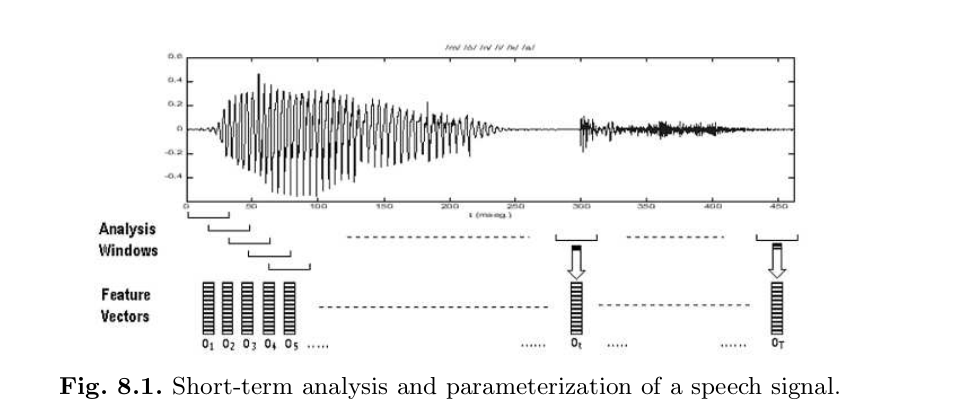
**3. Rút trích đặc trưng và Tokenization**

Bước đầu trong việc xây dưng một hệ thống nhận dạng người nói tự động là rút trích các đặc trưng đáng tin cậy và những tokens có chứa thông tin nhận dạng.

Sử dụng rút trích cả hai vectors đặc trưng short-term (spectral information, energy, pitch) và midterm, long-term tokens như phones (ngữ âm), syllables (âm tiết) và words (từ)

**3.1 Phân tích phân đoạn (Short-term Analysis)**

Để thực hiện phân tích phổ đáng tin cậy, các tín hiệu phải thể hiện các đặc tính tĩnh, điều này không dễ quan sát trong các tín hiệu lời nói thay đổi liên tục. Tuy nhiên, nếu chúng tôi giới hạn cửa sổ phân tích của mình ở độ dài ngắn từ 20 đến 40 mili giây, hệ thống ngữ âm của chúng ta không thể thay đổi đáng kể trong một khung thời gian ngắn như vậy, thu được những gì thường được gọi là tín hiệu giả tĩnh trên mỗi khung. Quá trình này được mô tả trong hình 8.1. Các tín hiệu cửa sổ đó có thể được giả định, do tính ngẫu nhiên giả, đến từ một hệ thống LTI (bất biến thời gian tuyến tính) cụ thể cho khung đó, và sau đó chúng ta có thể thực hiện, thường là sau khi sử dụng một số loại cửa sổ giống như cosine làm hamming hoặc hanning , phân tích quang phổ qua cửa sổ ngắn này, thu được các bao quang phổ thay đổi từng khung [41], [27].



**3.2 Tham số hóa (Parameterization)**

Tín hiệu cửa sổ hamming/ hanning trong thời gian ngắn này có tất cả thông tin thời gian/ phổ mong muốn, mặc dù ở tốc độ bit cao (ví dụ: số hóa giọng nói điện thoại với tần số lấy mẫu 8 kHz trong một cửa sổ 32 ms. Có nghĩa là 256 mẫu x 16 bit / mẫu = 4096 bit = 512 byte mỗi khung hình). Linear Predictive Coding (LPC) của giọng nói đã được chứng minh là một cách hợp lệ để nén đường bao quang phổ trong một mô hình toàn cực (hợp lệ cho tất cả các âm không phải âm mũi và vẫn là một giá trị gần đúng tốt cho âm mũi) chỉ với 10 đến 16 hệ số , có nghĩa là thông tin quang phổ trong khung có thể được biểu diễn trong khoảng 50 byte, bằng 10% tốc độ bit ban đầu. Thay vì các hệ số LPC, có tương quan cao giữa chúng (ma trận hiệp phương sai xa đường chéo), hệ số cepstral giả trực giao thường được sử dụng, hoặc được lấy trực tiếp như trong LPCC (vector Cepstral do LPC) từ hệ số LPC, hoặc trực tiếp thu được từ một- dựa trên phân tích quang phổ bộ lọc mel như trong MFCC (Hệ số quang phổ dựa trên tần số Mel). Một số hình thức liên quan khác được mô tả trong tài liệu, như PLP (Dự đoán tuyến tính dựa trên tri giác) [25], LSF (Tần số phổ dòng) [28] và nhiều dạng khác, không được nêu chi tiết ở đây để đơn giản hóa. Cho đến nay, một trong những yếu tố chính của sự biến đổi giọng nói đến từ việc sử dụng các kênh truyền khác nhau (ví dụ: thử nghiệm giọng nói qua điện thoại với các mẫu loa được ghi âm bằng micrô). Biểu diễn cepstral cũng có lợi thế là các kênh bất biến cộng thêm một độ lệch cepstral không đổi có thể dễ dàng bị trừ (CMS.- Cepstral Mean Subtraction), và các thành phần cepstral không phải giọng nói cũng có thể bị loại bỏ như được thực hiện trong lọc RASTA đối với các vectơ tức thời của cepstral [26] . Để tính đến quá trình coarticulation, hệ số delta (vận tốc) và delta-delta (gia tốc) được lấy từ thông tin dựa trên cửa sổ tĩnh, tính toán ước tính về cách mỗi hệ số khung thay đổi qua các cửa sổ liền kề (thường là từ ± 3, không hơn hơn ± 5)

Linear Predictive Coding (LPC) - Mã hóa dự đoán tuyến tính

Mel-Frequency based Cepstral Coefficients (MFCC)

**3.3 Ngữ âm (Phonetic) và phân tách từ (Word Tokenization)**

Mô hình Markov ẩn (HMM - Hidden Markov Models) [40] là công cụ thành công nhất và được sử dụng rộng rãi (ngoại trừ một số kiến ​​trúc ANN [37]) để mã hóa ngữ âm, âm tiết và từ, nghĩa là dịch từ lời nói được lấy mẫu thành một căn chỉnh thời gian dãy các đơn vị ngôn ngữ.

Left-to-RightHMMs là máy trạng thái mà mô hình thống kê các phần giả định của giọng nói (trạng thái) và quá trình chuyển đổi (ép buộc từ trái sang phải, giữ cảm giác thời gian) giữa các trạng thái, cố gắng bắt chước bằng cách nào đó chuyển động của các cơ quan cấu âm của chúng ta, có xu hướng nghỉ ngơi (trong trường hợp tất cả các âm thanh không phát) ở các vị trí khớp (được gọi là trạng thái giả tĩnh) và liên tục di chuyển (chuyển tiếp) từ trạng thái này sang trạng thái sau.

Hiện tại, hầu hết những thông tin của mô hình HMM ở mỗi trạng thái với các hàm mật độ xác suất liên tục, điển hình là Hỗn hợp Gaussian. Loại mô hình này thườnlàmg được gọi là CDHMM (Continuous Density HMM, ngược lại với VQ-based Discrete Density).

Huấn luyện HMM thường được thực hiện thông qua ước lượng Baum-Welch, trong khi giải mã và căn chỉnh thời gian thường được thực hiện thông qua giải mã Viterbi. Hiệu suất của các phổ HMM đó được cải thiện bằng cách sử dụng các mô hình ngôn ngữ, mô hình này áp đặt một số ràng buộc về ngôn ngữ hoặc ngữ pháp đối với sự kết hợp gần như vô hạn của tất cả các đơn vị có thể. Để cho phép tăng hiệu quả, việc lược bớt tìm kiếm chùm tia cũng là một cơ chế tổng quát để đẩy nhanh đáng kể quá trình nhận dạng mà không có hoặc ít suy giảm hiệu suất.

**3.4 Phân tách ngữ điệu (Prosodic Tokenization)**

Các đặc trưng ngữ điệu cơ sở như cao độ và năng lượng cũng có được ở mức frame. Năng lượng cửa sổ thu được rất dễ dàng thông qua định lý Parseval, ở dạng thời gian hoặc dạng phổ, và cao độ tức thời có thể được xác định bằng, ví dụ, phương pháp tự tương quan hoặc dựa trên phân rã cepstral, thường được làm mịn bằng một số lọc thời gian [41]. Các đặc điểm thuận âm quan trọng khác là những đặc điểm liên quan đến thời lượng của các đơn vị ngôn ngữ, tốc độ nói và tất cả những đặc điểm liên quan đến trọng âm. Trong tất cả những trường hợp đó, cần phải phân đoạn chính xác, đánh dấu các vị trí âm tiết, đường nét năng lượng và cao độ để phát hiện các vị trí trọng âm và dấu chuyển cụm từ hoặc giọng nói. Việc phân đoạn ngữ âm và âm tiết của lời nói là một vấn đề phức tạp còn lâu mới giải quyết được [53] và mặc dù nó có thể hữu ích cho việc nhận dạng người nói [1], các hệ thống âm tiết không phải lúc nào cũng yêu cầu phân đoạn chi tiết như vậy [13].

**4. Nhận dạng giọng nói phụ thuộc văn bản**

Hệ thống nhận dạng giọng nói có thể được phân thành hai loại phụ lớn: phụ thuộc vào văn bản và không phụ thuộc vào văn bản. Đối với loại thứ nhất, người ta sử dụng nội dung từ vựng của giọng nói phát ra để nhận dạng giọng nói, trong loại thứ hai cố gắng giảm thiểu ảnh hưởng của nội dung từ vựng vốn được coi là không xác định đối với khả năng nhận dạng của giọng nói. Sự khác biệt này làm cho hai loại này của hệ thống nhận dạng giọng rất khác nhau về cả kỹ thuật được sử dụng và các ứng dụng tiềm năng. Phần này dành cho các hệ thống nhận dạng người nói phụ thuộc vào văn bản, ứng dụng chính của hệ thống này trong các hệ thống tương tác, nơi cần có sự hợp tác từ người dùng để xác thực danh tính của họ. Ví dụ điển hình của các ứng dụng này là xác thực bằng giọng nói qua điện thoại cho các hệ thống phản hồi giọng nói tương tác yêu cầu một số mức độ bảo mật như các ứng dụng ngân hàng hoặc đặt lại mật khẩu. Tương tự như các phương thức sinh trắc học khác, việc sử dụng hệ thống nhận dạng giọng nói phụ thuộc vào văn bản yêu cầu giai đoạn đăng ký trong đó người dùng cung cấp một số mẫu để xây dựng mô hình người dùng và giai đoạn nhận dạng trong đó mẫu giọng nói mới được so khớp với mô hình người dùng.

**4.1 Phân loại các hệ thống và kỹ thuật**

Chúng ta có thể phân loại hệ thống nhận dạng người nói phụ thuộc vào văn bản theo quan điểm ứng dụng thành hai loại: hệ thống văn bản cố định và hệ thống văn bản biến đổi. Trong các hệ thống văn bản cố định, nội dung từ vựng trong ghi danh và các mẫu nhận dạng luôn giống nhau. Trong các hệ thống văn bản biến đổi, nội dung từ vựng trong mẫu nhận dạng là khác nhau trong mọi thử nghiệm truy cập với nội dung từ vựng của các mẫu đăng ký. Hệ thống văn bản biến đổi linh hoạt hơn và mạnh mẽ hơn trước các cuộc tấn công sử dụng bản ghi âm từ người dùng hoặc bắt chước sau khi nghe người nói thực sự nói đúng mật khẩu. Một khả năng thú vị là việc tạo ra một lời nhắc mật khẩu được tạo ngẫu nhiên khác nhau mỗi khi người dùng được xác minh (hệ thống nhắc bằng văn bản), do đó hầu như không thể sử dụng bản ghi. Đối với các kỹ thuật được sử dụng để nhận dạng người nói phụ thuộc vào văn bản, [14] đã chứng minh rằng thông tin hiện diện ở các cấp độ khác nhau của tín hiệu giọng nói (các đặc trưng kích thích tối đa, phổ và siêu phân đoạn) có thể được sử dụng một cách hiệu quả để xác minh danh tính của người dùng. Tuy nhiên, thông tin được sử dụng rộng rãi nhất là nội dung phổ của tín hiệu tiếng nói, được xác định bởi cấu hình vật lý và động lực của đường thanh quảng. Thông tin này thường được tóm tắt dưới dạng chuỗi thời gian của các vector MFCC, mỗi vector trong số đó đại diện cho một thời lượng nói từ 20-40 mili giây. Bằng cách này, vấn đề nhận dạng người nói phụ thuộc vào văn bản được giảm xuống thành vấn đề so sánh chuỗi các vector MFCC với mô hình của người dùng. Để so sánh này có hai phương pháp đã được sử dụng rộng rãi: phương pháp dựa trên khuôn mẫu và phương pháp thống kê. Trong các phương pháp dựa trên khuôn mẫu [20], [17] mô hình của người nói bao gồm một số chuỗi vectơ tương ứng với lời nói đăng ký và việc nhận dạng được thực hiện bằng cách so sánh lời nói xác minh với lời nói đăng ký. So sánh này được thực hiện bằng cách sử dụng Dynamic Time Warping (DTW) như một cách hiệu quả để cải thiện sai lệch thời gian giữa các cách phát âm khác nhau. Trong khi các phương pháp này vẫn được sử dụng, đặc biệt đối với các hệ thống nhúng có tài nguyên rất hạn chế, các phương pháp thống kê và đặc biệt là Mô hình Markov ẩn (HMM) [40], có xu hướng được sử dụng thường xuyên hơn các mô hình dựa trên khuôn mẫu. HMMs cung cấp tính linh hoạt hơn, cho phép chọn đơn vị tiếng nói từ đơn vị âm vị phụ đến từ và cho phép thiết kế hệ thống nhắc văn bản [38], [6].

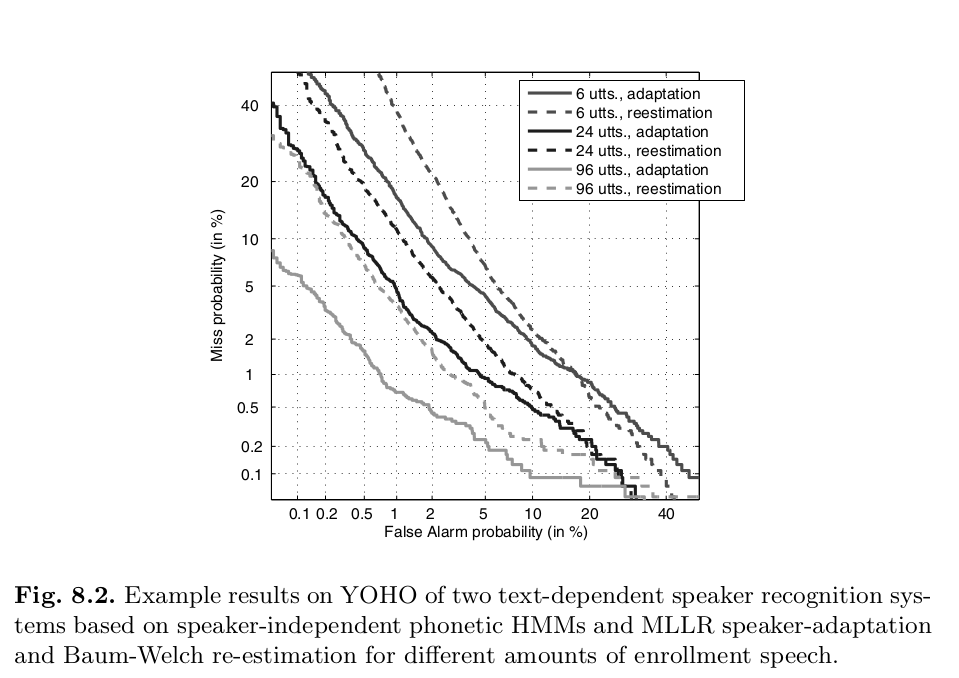
**4.2 Cơ sở dữ liệu và benchmarks**

Cơ sở dữ liệu đầu tiên được sử dụng để xác minh người nói phụ thuộc vào văn bản là cơ sở dữ liệu không được thiết kế đặc biệt cho tác vụ này như cơ sở dữ liệu TI-DIGITS [33] và TIMIT [21]. Một trong những cơ sở dữ liệu đầu tiên được thiết kế đặc biệt cho nghiên cứu nhận dạng người nói phụ thuộc vào văn bản là YOHO [5]. Nó bao gồm 96 mẫu lời nói khác nhau đã được thu thập trong 4 phiên khác nhau và 40 mẫu lời nói để kiểm tra được thu thập trong 10 lần cho mỗi người trên tổng số 138 người nói. Mỗi câu nói bao gồm các bộ ba cặp chữ số khác nhau (ví dụ: “12-34-56”). Đây có lẽ là tiêu chuẩn mở rộng và nổi tiếng nhất để so sánh và thường được sử dụng để đánh giá các hệ thống phụ thuộc vào văn bản. Tuy nhiên, cơ sở dữ liệu YOHO có một số hạn chế. Ví dụ: nó chỉ chứa lời nói được ghi lại trên một micrô duy nhất trong môi trường yên tĩnh và không được thiết kế để mô phỏng những kẻ giả mạo đã được thông báo (tức là những kẻ mạo danh nói ra mật khẩu của người dùng). Gần đây hơn, MIT Mobile Device Verification Corpus [55] đã được thiết kế để cho phép nghiên cứu xác minh loa phụ thuộc vào văn bản trong điều kiện nhiễu thực tế, trong khi BIOSEC Baseline Corpus [19] đã được thiết kế để mô phỏng giả mạo thông tin (bao gồm cả tài liệu song ngữ và một số phương thức sinh trắc học ngoài giọng nói). Một trong những khó khăn chính của việc so sánh các hệ thống xác minh người nói phụ thuộc vào văn bản khác nhau là các hệ thống này có xu hướng phụ thuộc vào ngôn ngữ, và do đó nhiều nhà nghiên cứu có xu hướng trình bày kết quả của họ trong cơ sở dữ liệu tùy chỉnh của họ, khiến cho việc so sánh trực tiếp không thể thực hiện được. Việc so sánh các hệ thống thương mại khác nhau thậm chí còn khó hơn. May mắn thay, một ấn phẩm gần đây [16] đã so sánh hiệu suất kỹ thuật của một vài hệ thống thương mại. Tuy nhiên, cũng như các phương thức sinh trắc học khác, hiệu suất kỹ thuật không phải là thứ nguyên duy nhất để đánh giá và các thước đo khác liên quan đến khả năng sử dụng của hệ thống cũng cần được đánh giá [51].

**4.3 Công trình liên quan: Text-dependent speaker recognition with HMM**

**speaker adaptation and HMM reestimation**

Như một ví dụ về hệ thống phụ thuộc vào văn bản được thử nghiệm trên cơ sở dữ liệu benchmark YOHO, chúng tôi trình bày kết quả thu được với hai hệ thống nhận dạng giọng nói phụ thuộc vào văn bản do các tác giả phát triển. Các hệ thống này mô phỏng một hệ thống được nhắc bằng văn bản dựa trên một tập hợp các HMM ngữ âm không phụ thuộc vào giọng nói và không phụ thuộc vào ngữ cảnh được đào tạo trên TIMIT. Việc ghi danh bao gồm việc sử dụng một số câu của người nói để điều chỉnh HMM cho người nói. Chúng tôi so sánh hai cách thực hiện điều chỉnh này: với một lần vượt qua ước lượng Baum-Welch và với Maximum Likelihood Linear Regression (MLLR) [32]. Cách tiếp cận trước đây là cách tiếp cận thông thường nhất nhưng yêu cầu sử dụng các HMM rất đơn giản (chỉ một hoặc một vài hàm Gauss cho mỗi trạng thái). Càng về sau thì mới lạ hơn và cho phép sử dụng các HMM phức tạp hơn. Việc xác minh giọng nói bao gồm việc tính toán điểm âm được tạo ra trong quá trình căn chỉnh bắt buộc của lời nói với phiên âm của nó bằng cách sử dụng cả HMM thích hợp với loa và HMM không phụ thuộc vào giọng nói. Điểm cuối cùng trong thử nghiệm này chỉ đơn giản là tỷ lệ giữa các điểm đó (không bao gồm chuẩn hóa điểm trong các kết quả được trình bày).



Một vấn đề quan trọng trong việc phát triển hệ thống nhận dạng giọng nói phụ thuộc vào văn bản là số lượng tài nguyên huấn luyện cần thiết là rất nhiều. YOHO chứa 4 phần với 24 câu nói mỗi phần. Đây là một số lượng rất lớn tài nguyên mà hiếm khi có thể có được trong một ứng dụng thực tế. Vì lý do này, hình 8.2 cho thấy kết quả huấn luyện hai hệ thống với 4 phần (96 giọng nói), một phần (24 giọng nói) hoặc chỉ 6 giọng nói từ một phần. Như có thể mong đợi, hiệu suất được cải thiện đáng kể với nhiều tài liệu đào tạo hơn, nhưng các hệ thống thực tế cần phải tìm ra sự thỏa hiệp giữa hiệu suất và sự dễ dàng và thuận tiện khi sử dụng. Hình 8.2 cũng so sánh hệ thống dựa trên ước lượng lại Baum-Welch và hệ thống dựa trên MLLR thích ứng, cho thấy hiệu suất tốt hơn cho các hệ thống dựa trên MLLR đối với tất cả các điều kiện tuyển sinh

**5. Nhận dạng giọng nói độc lập với văn bản**

Nhận dạng người nói không phụ thuộc vào văn bản đã bị thống trị phần lớn, kể từ những năm 1970 đến cuối thế kỷ 20, bởi các hệ thống dựa trên quang phổ ngắn. Kể từ năm 2000, các hệ thống mức cao hơn bắt đầu được phát triển với kết quả đủ tốt trong cùng các tác vụ có độ thách thức cao (theo đánh giá của NIST SR). Tuy nhiên, các hệ thống quang phổ tiếp tục hoạt động tốt hơn các hệ thống mức cao (NIST 2006 SRE là điểm chuẩn mới nhất vào thời điểm viết bài), với kết quả phát hiện tốt nhất nhờ vào kỹ thuật bù trừ tiên tiến gần đây.

**5.1 Short-term spectral systems**

Khi phân tích phổ trong khoảng thời gian ngắn được sử dụng để mô hình các đặc trưng người nói, chúng ta đang mô hình hóa các “âm thanh” khác nhau mà một người có thế tạo ra, nhờ vào đường âm và các cơ quan cấu âm của họ. Khi con người cần nhiều âm thanh (hoặc các ký hiệu âm học khác nhau) để nói bằng bất kỳ ngôn ngữ chung nào, chúng ta rõ ràng đang đối mặt với một không gian đa lớp gồm nhiều đặc trưng. Kỹ thuật Lượng tử hóa Vector (Vector Quantization techniques) hiệu quả trong các bài toán đa lớp như vậy và đã được sử dụng để xác định người người [4], điển hình là một mô hình VQ cụ thể cho một người nói, tính toán khoảng cách từ bất kỳ câu nói nào đến bất kỳ mô hình nào dưới dạng tổng trong số của khoảng cách tối tiểu trên mỗi khung đến codevector gần của codebook. Việc sử dụng các giới hạn và các điểm trung tâm thay vì dùng mật độ xác suất làm hiệu suất của VQ kém hơn so với mô hình Markov ẩn với mật độ liên tục và liên thông hay còn gọi là Ergodic HMMs [34]. Tuy nhiên, yếu tố quan trọng trong E-HMM là tích số trạng thái với hàm Gaussian mỗi trạng thái, điều này loại bỏ triệt để ảnh hưởng của quá trình chuyển đổi trong mô hình liên thông. Sau đó, một hệ thống HMM với 5 trạng thái - 4 Gaussian cho mỗi trạng thái sẽ hoạt động tương tự như 4-trạng thái 5-Gaussian, 2-trạng thái 10-Gaussian, hoặc thậm chí là 1-trạng thái 20-Gaussian, mà một cách hiểu tổng quát là GMM (Gaussian Mixture Model). Những one-state E-HMMs hoặc GMMs này có lợi thế lớn, tránh được cả ước lượng Baum-Welch cho việc huấn luyện, không cần sự liên kết giữa lời nói và trạng thái (tất cả lời nói đều được tinh chỉnh với cùng một trạng thái duy nhất) và dùng Viterbi decoding cho việc việc kiểm thử (không cần tinh chỉnh thời gian), giúp tăng tốc thời gian tính toán mà không ảnh hưởng đến hiệu suất.

GMM là một kỹ thuật tổng hợp trong đó một hỗn hợp các hàm Gaussians Đa chiều cố gắng mô hình hóa phân phối thống kê chưa rõ của dữ liệu người nói. GMM trở thành một kỹ thuật hiện đại vào những năm 1990, cả khi Maximum likehood (Tối đa hóa kỳ vong Expectation-Maximization, EM) hoặc huấn luyện phân loại (Maximum Mutual Information, MMI) còn được dùng. Tuy nhiên, việc sử dụng MAP tương thích với hầu hết phương tiện từ một Universal Background Model (UBM) đã mang lại cho GMMs một lợi thế lớn so với các kỹ thuật khác [43], đặc biệt khi được sử dụng với các kỹ thuật chuẩn hóa như Z-standard (chuẩn hóa điểm giả), T -norm (chuẩn hóa âm thanh), H-norm (chuẩn Z phụ thuộc vào thiết bị cầm tay), HT-norm (H + T-norm) hoặc Feature Mapping (xác định và chuẩn hóa kênh) [44].

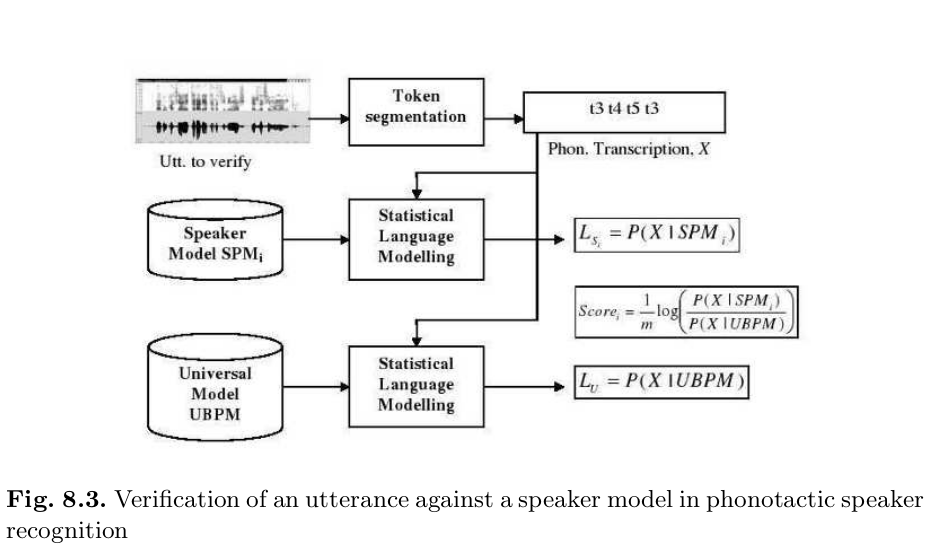
Discriminative techniques - Kỹ thuật phân lớp như Artificial Neural Networks, đã được sử dụng trong nhiều năm [18], nhưng hiệu suất của chúng chưa bao giờ đạt đến hiệu suất của GMM. Tuy nhiên, vào cuối những năm 90, SVM - Support Vector Machine được ví như một bộ phân loại hiệu năng cao được huấn luyện sẵn, mang lại cho GMM một đối thủ cạnh tranh vì hiệu suất gần như tương đương bằng việc sử dụng SVM trong không gian nhiều chiều hơn với các kernel tích hợp như GLDS (Generalized Linear Discriminant Sequence Kernel)

Gần đây, việc sử dụng SuperVectors [30], một kỹ thuật hỗn hợp GMM-SVM [9] coi là công cụ của GMM cho mọi trường hợp (cả trong huấn luyện và kiểm tra) là các điểm trong không gian nhiều chiều (số chiều bằng với số hỗn hợp của GMM nhân với số chiều của vectơ được tham số hóa) bằng cách sử dụng SVM cho mỗi người nói để phân loại các cách phát âm chưa biết từ siêu phẳng giọng nói được huấn luyện sẵn.

**5.2 Idiolectal systems (Idiolectal = idio (personal, private) + (dia)lect)**

Hầu hết hệ thống nhận dạng người nói không phụ thuộc văn bản đều dựa vào đặc trưng phổ ngắn cho đến khi công trình của Doddington [12] được công bố, nó mở ra một thế giới mới khả năng cải thiện các hệ thống nhận dạng người nói không phụ thuộc vào văn bản. Doddington đã nhận ra và chứng minh rằng lời nói của những người nói khác nhau không chỉ khác nhau về âm học, mà còn về các đặc điểm khác như cách sử dụng từ. Đặc biệt, trong công việc của mình, ông đã lập mô hình cách sử dụng từ của từng người nói cụ thể bằng cách sử dụng n-gram mô hình hóa các chuỗi từ và xác suất của chúng và chứng minh rằng việc sử dụng các mô hình đó có thể cải thiện hiệu suất của hệ thống GMM âm thanh/ phổ cơ bản. Quan trọng hơn kết quả cụ thể này là thực tế là công trình này đã thúc đẩy nghiên cứu trong việc sử dụng các cấp độ thông tin cao hơn (idiolectal, phonotactic, prosodic, v.v.) để nhận dạng giọng nói độc lập với văn bản. Sau khi công bố công trình này, một số nhà nghiên cứu đã gặp nhau tại hội thảo mùa hè SuperSID [15], nơi những ý tưởng này được phát triển thêm và thử nghiệm trên một tấm thử nghiệm chung. Các phần tiếp theo mô tả hai trong số những hệ thống thành công nhất khai thác mức độ thông tin cao hơn: hệ thống âm vị, cố gắng mô hình hóa các đặc điểm phát âm và hệ thống thuận âm, mô hình hóa các mẫu âm thanh chuyên biệt dành cho người nói.

**5.3 Phonotactic systems**



Một hệ thống nhận dạng loa phonotactic điển hình bao gồm hai khối chính:

- Bộ giải mã ngữ âm (the phonetic decoders), chuyển đổi lời nói thành một chuỗi các nhãn ngữ âm

- Giai đoạn mô hình hóa ngôn ngữ thống kê n-gram, mô hình hóa tần số của điện thoại và chuỗi điện thoại cho mỗi người nói cụ thể

Bộ giải mã ngữ âm – thường dựa trên Mô hình Markov ẩn (HMMs) - có thể được lấy từ trình nhận dạng giọng nói có sẵn hoặc được huấn luyện đặc biệt. Đối với mục đích nhận dạng người nói, việc có bộ giải mã ngữ âm không quan trọng và thậm chí không quan trọng lắm phải có bộ giải mã ngữ âm trong ngôn ngữ của người nói để được nhận dạng. Thực tế có phần đáng ngạc nhiên này đã được phân tích trong [52] cho thấy rằng các lỗi ngữ âm phụ thuộc vào giọng nói do bộ giải mã tạo ra dường như là của giọng nói cụ thể và do đó thông tin hữu ích cho việc nhận dạng giọng nói miễn là các lỗi này phù hợp với từng giọng nói cụ thể.

Sau khi có bộ giải mã ngữ âm, bản giải mã ngữ âm của nhiều câu từ nhiều người nói khác nhau có thể được sử dụng để huấn luyện Mô hình Universal Background Phone (UBPM) đại diện cho tất cả những người nói có thể có. Các Mô hình Ngữ âm Giọng nói (SPM\_i - Speaker Phone Models) được huấn luyện bằng cách sử dụng một số bộ giải mã ngữ âm của từng người nói cụ thể. Vì giọng nói có sẵn để huấn nói một mô hình giọng nói thường bị hạn chế, các mô hình giọng nói được nội suy với UBPM để tăng tính mạnh mẽ trong ước tính tham số. Sau khi các mô hình ngôn ngữ thống kê được huấn luyện, quy trình để xác minh cách phát âm một tập test so với mô hình giọng nói SPMi được trình bày trong Hình 8.3. Bước đầu tiên là tạo ra giải mã ngữ âm của nó, X, giống như cách giải mã được sử dụng để huấn luyện SPM\_{i} và UBPM. Sau đó, giải mã ngữ âm của câu thử, X và các mô hình thống kê (SPMi, UBPM) được sử dụng để tính toán khả năng giải mã ngữ âm, X, dựa trên mô hình giọng nói SPM\_{i} và mô hình nền UBPM. Điểm nhận dạng là logarith tỷ lệ của cả hai khả năng xảy ra. Quá trình này, thường được mô tả là Nhận dạng ngữ âm, sau đó là Mô hình hóa ngôn ngữ (PRLM) có thể được lặp lại cho các bộ giải mã ngữ âm khác nhau (ví dụ: các ngôn ngữ hoặc độ phức tạp khác nhau) và các điểm nhận dạng khác nhau chỉ được thêm vào hoặc hợp nhất để có hiệu suất tốt hơn, mang lại một phương pháp được gọi là PRLM hoặc PPRLM song song. Gần đây, một số cải tiến đã được đề xuất trên các hệ thống PPRLM cơ bản. Một trong những điều quan trọng nhất về mặt cải thiện hiệu năng là việc sử dụng toàn bộ mạng lưới nhận dạng ngữ âm [24] thay vì giả thuyết rằng có một bộ giải mã tốt nhất. Mạng lưới nhận dạng là một đồ thị không chu trình có hướng chứa các giả thuyết có khả năng xảy ra nhất cùng với các xác suất của chúng. Thông tin phong phú hơn nhiều này cho phép ước tính tốt hơn n-gam trên các tài nguyên giọng nói hạn chế, và do đó cho kết quả tốt hơn nhiều. Cải tiến quan trọng khác là việc sử dụng SVM để phân lớp toàn bộ n-gam được huấn luyện với giả thuyết tốt nhất hoặc với mạng [7], [24] thay vì sử dụng chúng trong framework phân lớp thống kê.

**5.4 Prosodic systems**

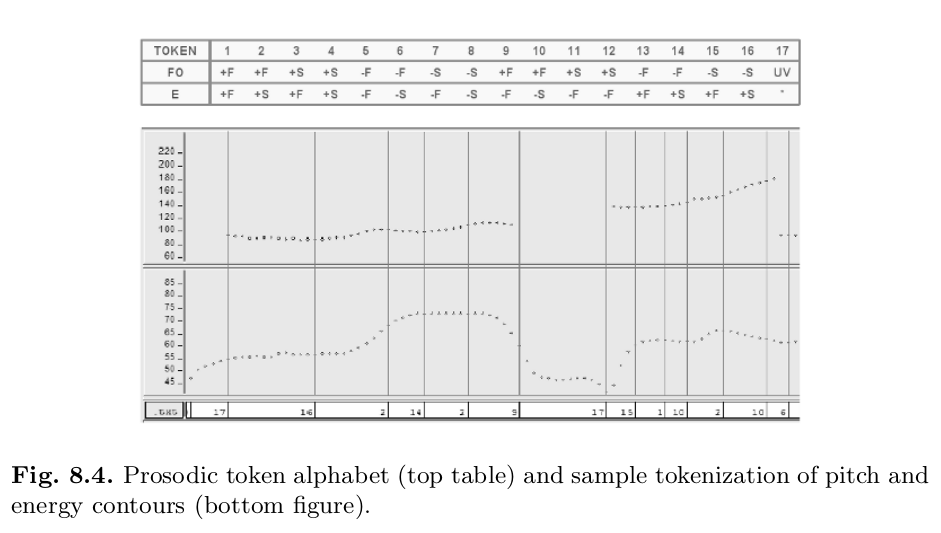
Một trong những hệ thống ngữ điệu tiên phong và thành công nhất trong việc nhận dạng giọng nói không phụ thuộc vào văn bản là công trình của Adami [13]. Hệ thống bao gồm hai khối xây dựng chính: ngữ điệu tokenizer, phân tích ưu điểm và biểu thị nó dưới dạng một chuỗi các nhãn hoặc token và giai đoạn mô hình hóa ngôn ngữ thống kê n-gram, mô hình hóa tần số của các tokens và trình tự của từng giọng nói cụ thể. Một số khả năng khác để mô hình hóa thông tin ngữ điệu cũng đã được chứng minh là khá thành công là việc sử dụng Non-uniform Extraction Region Features (NERFs) được phân định bằng khoảng dừng đủ dài [29] hoặc NERF được xác định bởi cấu trúc âm tiết của câu (SNERFs )[48].

Các tác giả đã triển khai một hệ thống ngữ điệu dựa trên công trình của Adami, trong đó khối thứ hai giống hệt nhau để nhận dạng ngữ âm và âm sắc chỉ với những điều chỉnh nhỏ để cải thiện hiệu suất. Quá trình tokenization bao gồm hai giai đoạn:

- Giai đoạn một, đối với mỗi đoạn giọng nói của đoạn thoại, quỹ đạo thời gian của các đặc điểm thuận (tần số cơ bản - hoặc cao độ- và năng lượng) được rút trích.

- Giai đoạn hai, cả hai đường bao đều được phân đoạn và dán nhãn bằng định lượng trung bình độ dốc.

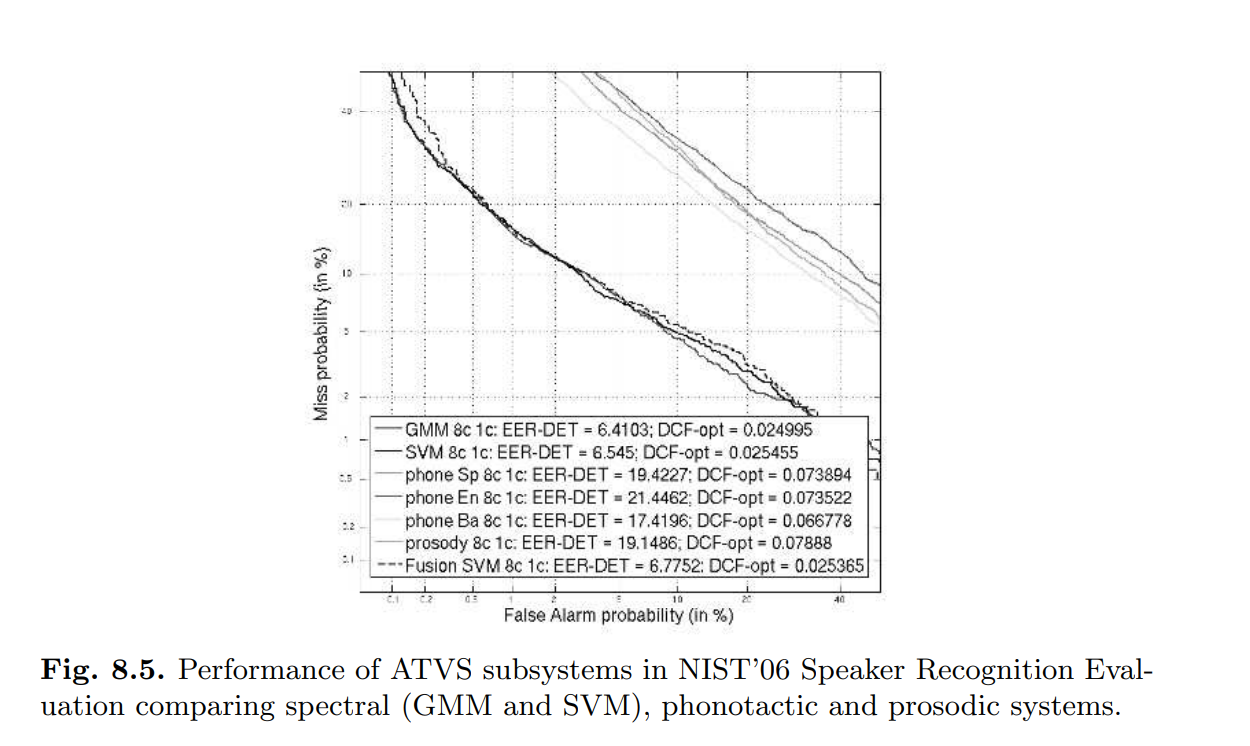
Hình 8.4 cho thấy một bảng chứa 17 token ngữ điệu. Một token đại diện cho các phân đoạn vô thanh, trong khi 16 token được sử dụng để đại diện cho các phân đoạn hữu thanh tùy thuộc vào độ dốc (phát nhanh, tăng chậm, giảm nhanh, giảm chậm) của năng lượng và cao độ. Hình 8.4 cũng cho thấy một ví dụ về câu nói được phân đoạn và gắn nhãn bằng cách sử dụng các token ngữ điệu này.

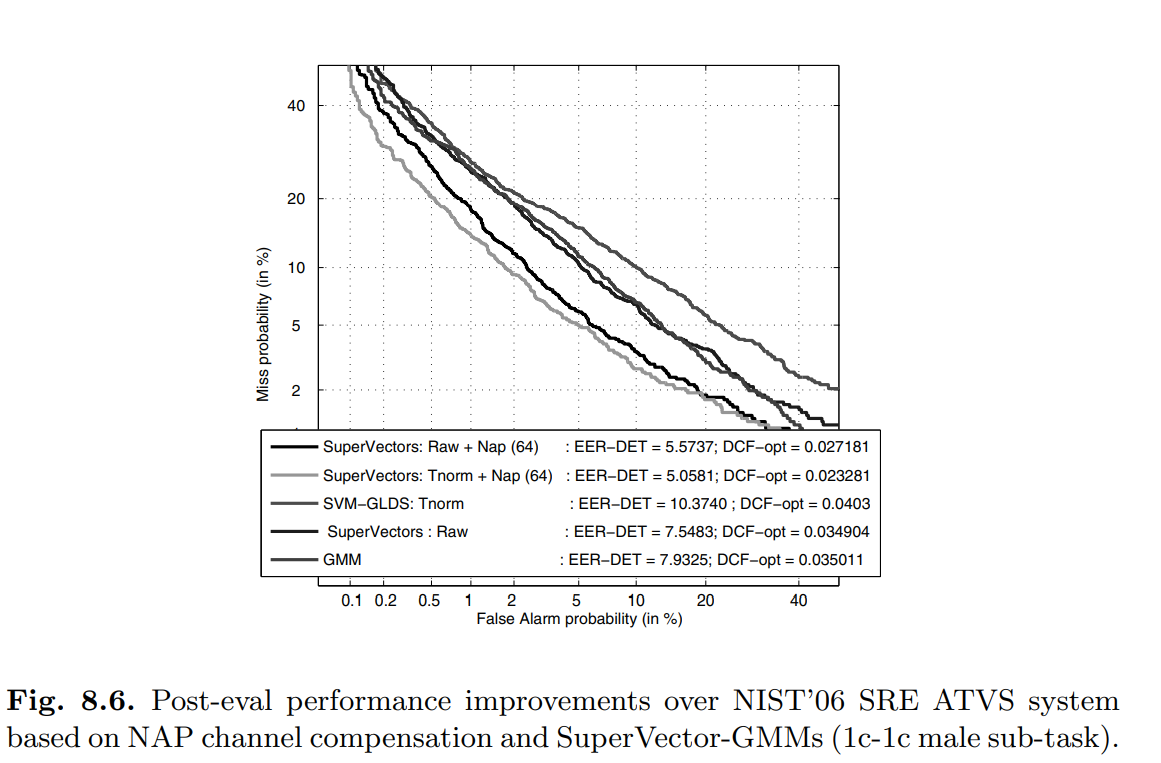


**5.5 Cơ sở dữ liệu và benchmarks**

Vào đầu những năm 1990, nhận dạng giọng nói độc lập với văn bản là một thách thức lớn, với một tương lai khó lường trước được. Vào thời điểm đó, các sáng kiến ​​nghiên cứu khiêm tốn đã được phát triển với cơ sở dữ liệu rất hạn chế, dẫn đến các công bố không đồng nhất và không có cách nào để so sánh và cải tiến các hệ thống trong các tác vụ tương tự. May mắn thay, vào năm 1996, NIST bắt đầu Đánh giá Nhận dạng Giọng nói -Speaker Recognition Evaluations hàng năm, đây chắc chắn là động lực của những tiến bộ đáng kể. Hiệu năng vượt trội của những nghiên cứu hoàn toàn nằm ngoài dự đoán của 10 năm trước. Thành công này được thúc đẩy bởi hai yếu tố. Thứ nhất, việc sử dụng cơ sở dữ liệu và giao thức chung để đánh giá các hệ thống đã cho phép so sánh công bằng giữa các hệ thống trên cùng một tác vụ. Thứ hai, các hội thảo sau đánh giá đã cho phép người tham gia chia sẻ kinh nghiệm, cải tiến, thất bại, v.v. của họ trong một môi trường hợp tác cao. Vai trò của LDC (Linguistic Data Consortium) cung cấp tài liệu nói mới đầy thách thức cũng rất đáng chú ý, vì nhu cầu liên tục tăng lên (cả về lượng lời nói và yêu cầu ghi âm). Từ các giai đoạn khác nhau của Switchboard đến cơ sở dữ liệu kiểu Fisher mới nhất, nhiều tiến bộ đã đạt được. Các bộ đánh giá trước đây (phát triển, huấn luyện và kiểm tra âm thanh và chìa khóa giải pháp) có sẵn thông qua LDC để các nhà nghiên cứu mới đánh giá hệ thống của họ mà không có áp lực cạnh tranh. Mặc dù kết quả "chính thức" đã bị hạn chế đối với người tham gia, việc theo dõi tiến trình của công nghệ là cực kỳ dễ dàng vì người tham gia thường trình bày những phát triển mới của họ trong các phiên ID người phát biểu trong các hội nghị quốc tế như ICASSP hoặc InterSpeech (trước đây là EuroSpeech), hoặc loạt Hội thảo ISCA / IEEE Odyssey.

**5.6 Công trình liên quan: the ATVS multilevel text-independent system**





6. Những ứng dụng

7. Tài liệu tham khảo

[1] A. G. Adami and H. Hermansky. Segmentation of speech for speaker and language recognition. In Proceedings of Interspeech, pages 841–844, 2003.

[2] C.G.G. Aitken and F. Taroni. Statistics and the Evaluation of Evidence for Forensic Scientists. John Wiley and Sons, 2 edition, 2004.

[3] N. Brummer and J. Preez. Application-independent evaluation of speaker detection. Computer, Speech and Language, 20:230–275, 2006.

[4] D. K. Burton. Text-dependent speaker verification using vector quantization source coding. IEEE Transactions on Acoustics, Speech and Signal Processing, 35:133–143, 1987.

[5] J. Campbell and A. Higgins. Yoho speaker verification (ldc94s16).

[http://www.ldc.upenn.edu](http://www.ldc.upenn.edu/).

[6] J. P. Campbell. Testing with the yoho cd-rom voice verification corpus. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 341–344, 1995.

[7] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek. High-level speaker verification with support vector machines. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pages 73–76, 2004.

[8] W.M. Campbell. Generalized linear discriminant sequence kernels for speaker recognition. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pages 161–164, 2002.

[9] W.M. Campbell, D.E. Sturim, and D.A. Reynolds. Support vector machines

using gmm supervectors for speaker verification. IEEE Signal Processing Letters, 13:308–311, 2006.

[10] P. Carr. English Phonetics and Phonology: An Introduction. Blackwell Publishing, Incorporated, 1999.

[11] Voice Biometrics Conference. [http://www.voicebiocon.com](http://www.voicebiocon.com/).

[12] G. Doddington. Speaker recognition based on idiolectal differences between speakers. In Proceedings of Interspeech, volume 4, pages 2517–2520, 2001.

[13] A. G. Adami et al. Modeling prosodic dynamics for speaker recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume IV, pages 788–791, 2003.

[14] B. Yegnanarayana et. al. Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system. IEEE Transactions on Speech and Audio Processing, 13:575–582, 2005.

[15] D. Reynolds et al. Supersid project: Exploiting high-level information for highaccuracy speaker recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume IV, pages 784–787, April 2003.

[16] M. Wagner et al. An evaluation of ’commercial off-the-shelf’ speaker verification systems. In Proceedings of IEEE Odyssey, 2006.

[17] V. Ramasubramanian et. al. Text-dependent speaker-recognition systems based on one-pass dynamic programming algorithm. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pages 901–904, 2006.

[18] K. R. Farrell, R. J. Mammone, and K. T. Assaleh. Speaker recognition using neural networks and conventional classifiers. IEEE Transactions on Speech and Audio Processing, 2:194–205, 1994.

[19] J. Fierrez-Aguilar, J. Ortega-Garcia, D. T. Toledano, and J. Gonzalez-

Rodriguez. Biosec baseline corpus: A multimodal biometric database. Pattern

Recognition, 40:1389–1392, 2007.

[20] S. Furui. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech and Signal Processing, 29:254–272, 1981.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, N. L.

Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus

(ldc93s1). [http://www.ldc.upenn.edu](http://www.ldc.upenn.edu/).

[22] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia. Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Computer, Speech and Language, 20:331–335, 2006.

[23] J. Gonzalez-Rodriguez, D. Ramos-Castro, D. T. Toledano, A. Montero-Asenjo, J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Fierrez-Aguilar, D. Garcia-

Romero, and J. Ortega-Garcia. Speaker recognition: the atvs-uam system at

nist sre 05. IEEE AES Magazine, 22:15–21, 2007.

[24] A. O. Hatch, B. Peskin, and A. Stolcke. Improved phonetic speaker recognition using lattice decoding. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pages 165–168, 2005.

[25] H. Hermansky, B. Hanson, and H. Wakita. Perceptually based linear predictive analysis of speech. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume 10, pages 509–512, 1985.

[26] H. Hermansky and N. Morgan. Rasta processing of speech. IEEE Transactions on Speech and Audio Processing, 2(4):578–589, October 1984.

[27] X. Huang, A. Acero, and H.-W. Hon. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall PTR, 2001.

[28] F. Itakura. Line spectrum representation of linear predictive coefficients of speech signals. Journal of the Acoustical Society of America, 57:S35, 1975.

[29] Sachin Kajarekar, Luciana Ferrer, Kemal Sonmez, Jing Zheng, Elizabeth

Shriberg, and Andreas Stolcke. Modeling NERFs for speaker recognition. In

Proceedings of IEEE Odyssey, pages 51–56, Toledo, Spain, June 2004.

[30] P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing, 13:345–354, 2005.

[31] P. Kenny and P. Dumouchel. Disentangling speaker and channel effects in

speaker verification. In Proceedings of the IEEE International Conference on

Acoustics, Speech, and Signal Processing, pages 37–40, 2004.

[32] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Computer, Speech and Language, 9:171–185, 1995.

[33] R. G. Leonard and G.Doddington Tidigits (ldc93s10). [http://www.ldc.upenn.edu](http://www.ldc.upenn.edu/).

[34] T. Matsui and S. Furui. Comparison of text-independent speaker recognition methods using vq-distortion and discrete/continuous hmms. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 157–160, 1992.

[35] Nist speaker recognition evaluation. <http://www.nist.gov/speech/tests/spk/>.

[36] J. Ortega-Garcia, J. Bigun, D. Reynolds, and J. Gonzalez-Rodriguez. Authentication gets personal with biometrics. IEEE Signal Processing Magazine, 21:50–62, 2004.

[37] Matejka Pavel, Schwarz Petr, Cernock Jan, and Chytil Pavel. Phonotactic language identification using high quality phoneme recognition. In Proceedings of InterSpeech, pages 2237–2240, 2005.

[38] CAVE Project. Cave - the european caller verification project. <http://www.ptttelecom.nl/cave/>.

[39] M. A. Przybocki, A. F. Martin, and A. N. Le. Nist speaker recognition evaluation chronicles part 2. In Proceedings of IEEE Odyssey, 2006.

[40] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77:257–286, 1989.

[41] L. R. Rabiner and R. W. Schafer. Digital Processing of Speech Signals. Prentice Hall, 1978.

[42] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia. Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework. In Proceedings of IEEE Odyssey, 2006.

[43] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted

gaussian mixture models. Digital Signal Processing, 10:19–41, 2000.

[44] D. A. Reynolds. Channel robust speaker verification via feature mapping. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume 2, pages 53–56, 2003.

[45] P. Rose. Forensic Speaker Identification. CRC, 1 edition, 2002.

[46] M. J. Saks and J. J. Koehler. The coming paradigm shift in forensic identification science. Science, 309:892–895, 2005.

[47] B. Scholkopf, S. Kah-Kay, C.J.C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. IEEE Transactions on Signal Processing, 45:2758–2765, 1997.

[48] Elizabeth Shriberg, Luciana Ferrer, Anand Venkataraman, and Sachin Ka-

jarekar. SVM modeling of “SNERF-Grams” for speaker recognition. In Proc.

Intl. Conf. Spoken Language Systems, pages 1409–1412, Jeju, Korea, October

2004.

[49] C. Soutar, D. Roberge, A. Stoianov, R. Gilroy, and B.V.K. Vijaya Kumar. Biometric encryption. (Online) [http://www.bio-scrypt.com](http://www.bio-scrypt.com/).

[50] K. N. Stevens. Acoustic Phonetics (Current Studies in Linguistics). The MIT Press, 2000.

[51] D. T. Toledano, R. Fernandez-Pozo, A. Hernandez-Trapote, and L. Hernandez-Gomez. Usability evaluation of multi-modal biometric verification systems. Interacting With Computers, 18:1101–1122, 2006.

[52] D. T. Toledano, C. Fombella, J. Gonzalez-Rodriguez, and L. Hernandez-Gomez. On the relationship between phonetic modeling precision and phonetic speaker recognition accuracy. In Proceedings of InterSpeech, pages 1993–1996, 2005.

[53] D. T. Toledano, L. Hernandez-Gomez, and L. Villarrubia-Grande. Automatic phonetic segmentation. IEEE Transactions on Speech and Audio Processing, 11:617–625, 2003.

[54] V. Wan and W. Campbell. Support vector machines for speaker verification and identification. In Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, volume 2, pages 775–784, 2000.

[55] R. Woo, A. Park, and T. J. Hazen. The mit mobile device speaker verification corpus: data collection and preliminary experiments. In Proceedings of IEEE Odyssey, 2006.

Biometrics Recognition Using Deep Learning: A Survey

Link paper:

Tác giả:

Phần Voice Biometrics (Trọng tâm, mình sẽ dịch nhiều phần này, mấy phần khác khá sơ sài)

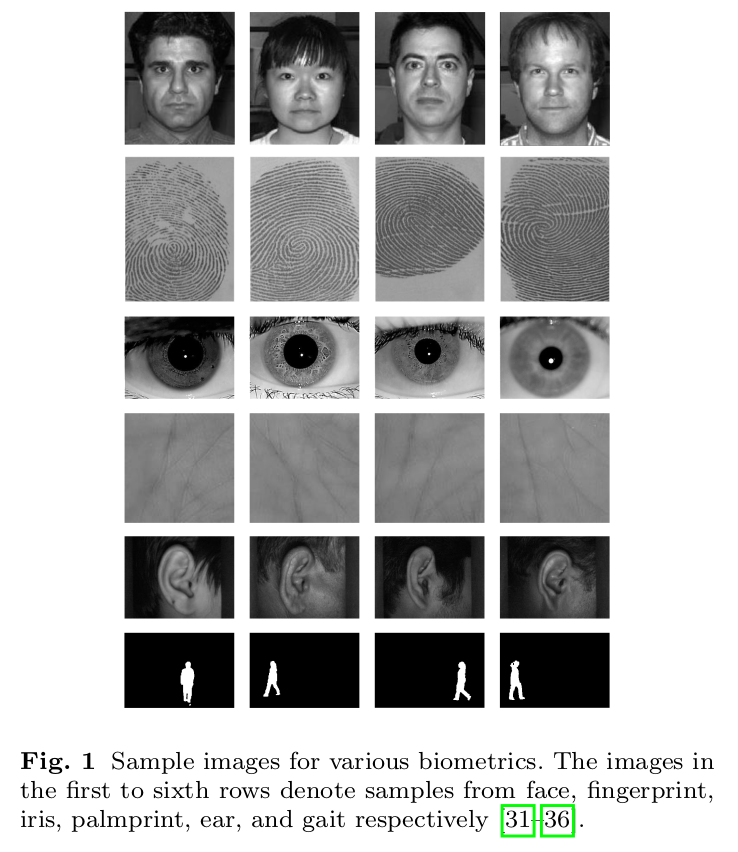
Abstract

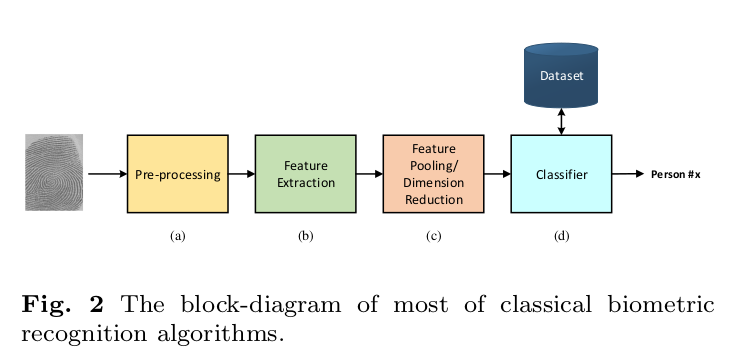
Keyword: Biometric Recognition, Deep Learning, Face Recognition, Fingerprint Recognition, IrisRecognition, Palmprint Recognition

1. Giới thiệu

Đặc trưng nhân trắc học nắm giữ vị trí độc tôn khi nói đến tác vụ nhận dạng, xác minh và các ứng dụng bảo mật. [1], [2]. Chúng không thể bị thất lạc, không giống như các tính năng dựa trên mã thông báo như khóa và thẻ ID, và chúng không thể bị quên, không giống như các tính năng dựa trên kiến thức, chẳng hạn như mật khẩu hoặc câu trả lời cho các câu hỏi bảo mật [3]. Ngoài ra, chúng gần như không thể bắt chước hoặc sao chép một cách hoàn hảo. Mặc dù gần đây đã có những nỗ lực tạo ra và giả mạo các đặc điểm sinh trắc học khác nhau [4], [5], nhưng cũng đã có những phương pháp được đề xuất để phân biệt các đặc điểm sinh trắc học giả với các đặc điểm sinh trắc học thật [6], [7], [8]. Những thay đổi theo thời gian đối với nhiều tính năng sinh trắc học cũng cực kỳ ít. Vì những lý do này, chúng đã được sử dụng trong nhiều ứng dụng, bao gồm xác thực điện thoại di động, an ninh sân bay và khoa học pháp y. Các đặc điểm sinh trắc học có thể là sinh lý học, là những đặc điểm mà bất kỳ người nào cũng sở hữu, chẳng hạn như dấu vân tay (fingerprint) [9], vân lòng bàn tay (palmprint) [10], [11], nét mặt (facial features) [12], tai (ear) [13], mắt (irises) [14], [15] và võng mạc (retinas) [16], hoặc hành vi (behavioral), biểu hiện rõ ràng trong tương tác của một người với môi trường, chẳng hạn như chữ ký (signatures) [17], dáng đi (gaits) [18] và cách gõ phím (keystroke) [19]. Giọng nói (Voice/Speech) chứa cả các đặc điểm hành vi, chẳng hạn như trọng âm và các đặc điểm sinh lý, chẳng hạn như cao độ giọng nói [20].

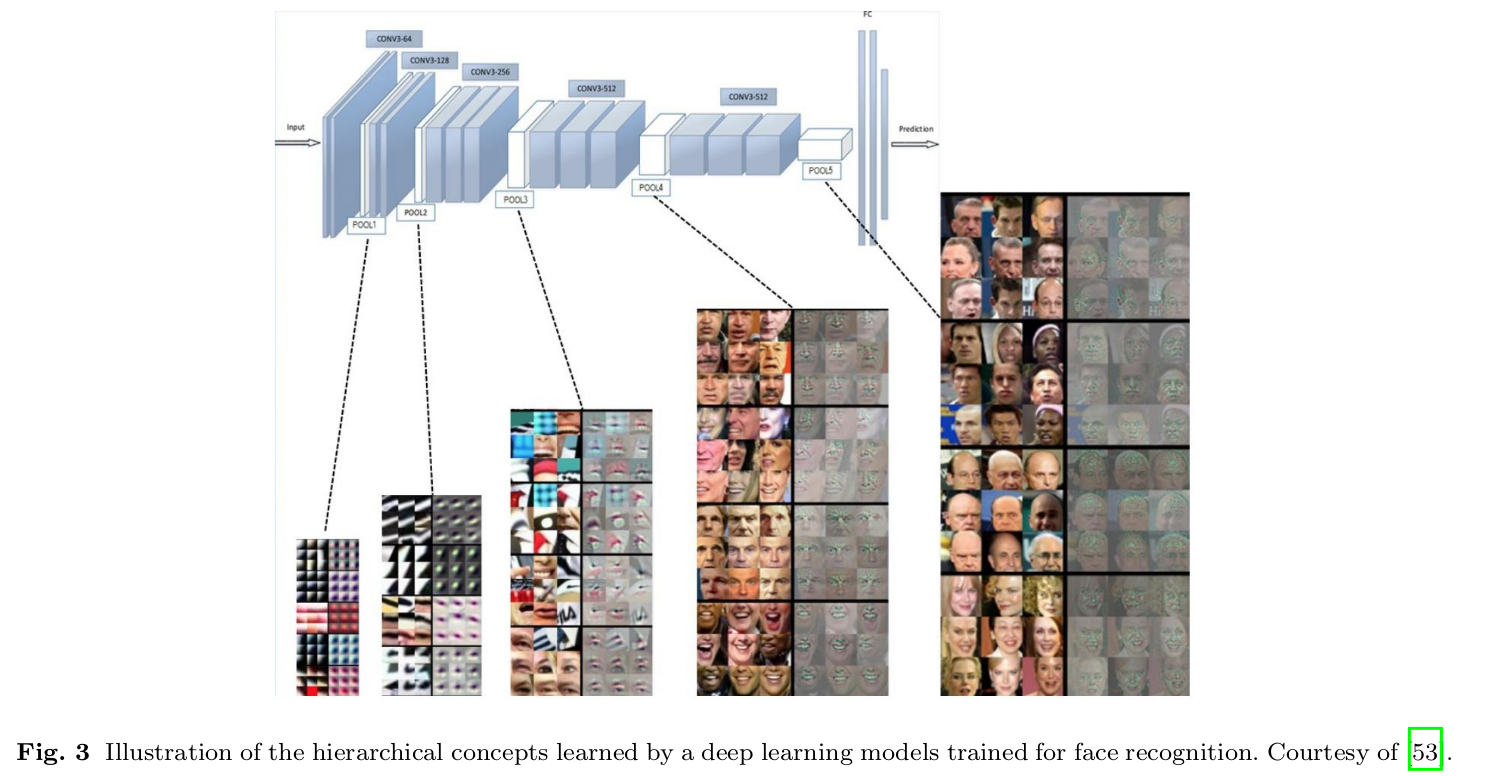
Khuôn mặt và dấu vân tay được cho là đặc điểm sinh trắc sinh lý được sử dụng phổ biến nhất. Vân tay là lâu đời nhất, có từ năm 1893 khi nó được sử dụng để kết tội một nghi phạm giết người ở Argentina [21]. Khuôn mặt có nhiều đặc điểm phân biệt có thể được sử dụng cho các nhiệm vụ nhận dạng [22]. Tuy nhiên, tính nhạy cảm của nó thay đổi do các yếu tố như biểu hiện hoặc lão hóa, có thể là một thách thức [23], [24]. Vân tay bao gồm các đường gờ và rãnh, tạo thành các mẫu duy nhất. Các điểm nhỏ là các phần cục bộ chính của dấu vân tay có thể được sử dụng để xác định tính duy nhất của dấu vân tay, hai trong số những phần quan trọng nhất trong số đó là phần cuối của đường vân tay và đường phân nhánh của đường vân tay [25]. Vân lòng bàn tay là một sự thay thế khác được sử dụng cho mục đích xác thực. Ngoài các đặc điểm nhỏ, bảng màu cũng bao gồm các đặc điểm dựa trên hình học, các điểm đồng bằng, các đường chính và các nếp nhăn [26]. Mống mắt và võng mạc là hai loại sinh trắc học phổ biến nhất có trong mắt và có thể được sử dụng để nhận biết thông qua kết cấu của mống mắt hoặc mô hình mạch máu trong võng mạc. Một điểm thú vị đáng chú ý là ngay cả hai mắt ở cùng một người cũng có những hình thái khác nhau [27]. Tai cũng có thể được sử dụng làm sinh trắc học thông qua hình dạng của các thùy và đường xoắn của chúng, và không giống như hầu hết các tính năng sinh trắc học, không cần sự tương tác trực tiếp của người đó. Tai phải và tai trái đối xứng ở một người trong hầu hết các trường hợp. Tuy nhiên, kích thước của chúng có thể thay đổi theo thời gian [28]. Trong số các đặc trưng hành vi, chữ ký được cho là được sử dụng rộng rãi nhất hiện nay. Các nét trong chữ ký có thể được kiểm tra áp lực của bút trong suốt chữ ký cũng như tốc độ, là một yếu tố tạo nên độ dày của nét [29]. Dáng đi đề cập đến cách đi bộ, đã đạt được chú ý nhiều hơn trong những năm gần đây. Do sự tham gia của nhiều điểm khớp và bộ phận cơ thể trong quá trình đi lại, dáng đi cũng có thể được sử dụng để nhận dạng duy nhất một người từ xa [30]. Các mẫu sinh trắc học khác nhau được thể hiện trong Hình 1.





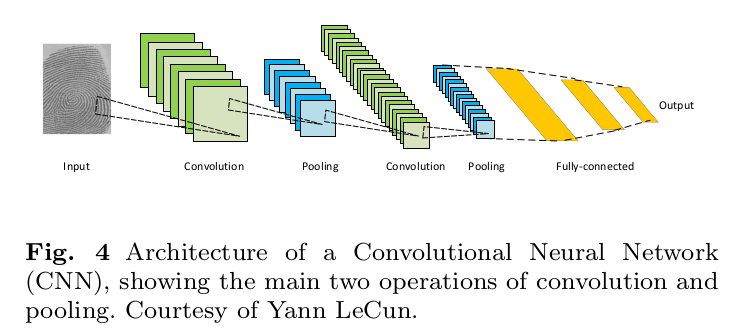
Theo cách truyền thống, quy trình nhận dạng sinh trắc học bao gồm một số bước quan trọng. Hình 2 cho thấy sơ đồ khối của các hệ thống nhận dạng sinh trắc học truyền thống. Thứ nhất, dữ liệu hình ảnh được thu thập thông qua (nhiều) máy ảnh hoặc cảm biến quang học, sau đó được xử lý trước để làm cho thuật toán hoạt động trên nhiều dữ liệu hữu ích nhất có thể. Sau đó, các tính năng được trích xuất từ mỗi hình ảnh. Các công trình nhận dạng sinh trắc học cổ điển chủ yếu dựa trên các tính năng được làm thủ công (do các chuyên gia thị giác máy tính thiết kế) để hoạt động với một loại dữ liệu nhất định [37], [38], [39]. Nhiều tính năng được tạo thủ công dựa trên sự phân bố của các cạnh (SIFT [40], HOG [41]), hoặc bắt nguồn từ miền biến đổi, chẳng hạn như Gabor [42], Fourier [43] và wavelet [44] . Phân tích thành phần chính cũng được sử dụng trong nhiều công trình để giảm kích thước của các đối tượng [45], [46]. Sau khi các tính năng được trích xuất, chúng được đưa vào một bộ phân loại để thực hiện nhận dạng.

Nhiều thách thức nảy sinh trong nhiệm vụ nhận dạng sinh trắc học truyền thống. Ví dụ: các tính năng được làm thủ công phù hợp với một sinh trắc học, sẽ không nhất thiết hoạt động tốt trên những tính năng khác. Do đó, sẽ cần một số lượng lớn các thử nghiệm để tìm và chọn tập hợp các tính năng thủ công thành thạo nhất cho một sinh trắc học nhất định. Ngoài ra, nhiều mô hình cổ điển dựa trên SVM nhiều lớp được huấn luyện theo kiểu một chọi một, sẽ không mở rộng quy mô tốt khi số lượng lớp học lớn.



**2. Toàn cảnh về Deep Neural Network**

**2.1 Convolutional Neural Networks - Mạng Neural Tích chập – CNNs**

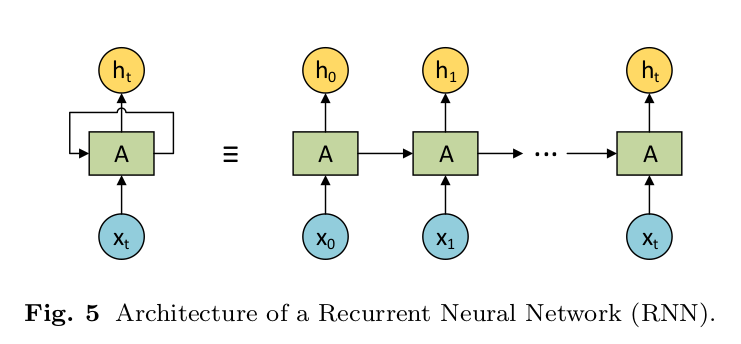


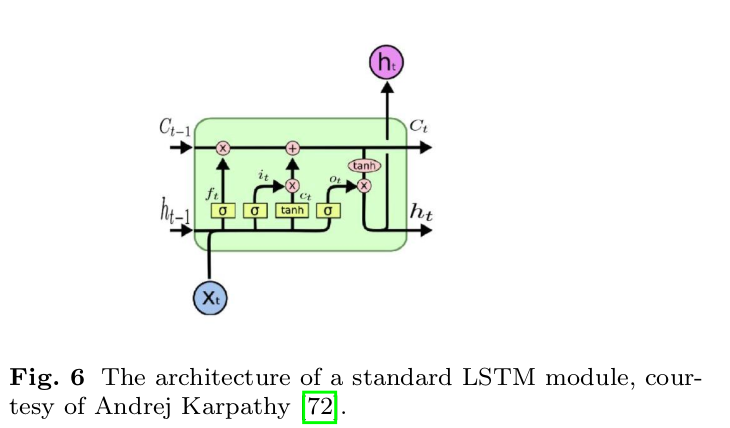
Convolutional Neural Networks - Mạng Neural Tích Chập (CNN) (lấy cảm hứng từ vỏ não thị giác của động vật có vú) là một trong những kiến trúc thành công nhất và được sử dụng rộng rãi trong cộng đồng học sâu - Deep Learning (đặc biệt cho các nhiệm vụ thị giác máy tính). CNN ban đầu được đề xuất bởi Fukushima trong một bài báo có tên là "Neocognitron" [58], dựa trên mô hình hệ thống thị giác của con người do những người đoạt giải Nobel Hubel và Wiesel đề xuất. Sau đó, Yann Lecun và các đồng nghiệp đã phát triển một khung tối ưu hóa (dựa trên sự lan truyền ngược) để học một cách hiệu quả các trọng số mô hình cho một Kiến trúc CNN [55]. Sơ đồ khối của một trong những mô hình CNN đầu tiên được phát triển bởi Lecun et al. được hiển thị trong Hình 4.

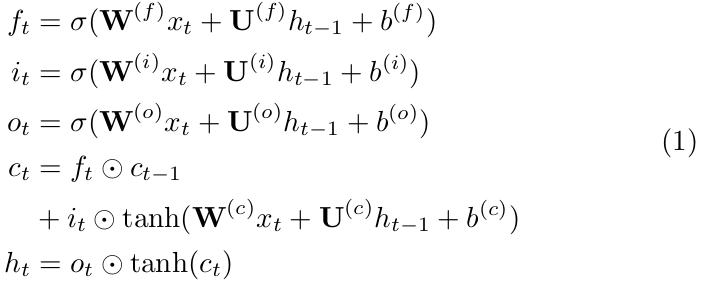
CNNs chủ yếu bao gồm ba loại lớp: lớp tích chập (convolutional layers), trong đó một kernel trượt được áp dụng cho hình ảnh (như trong hoạt động tích chập hình ảnh) để rút trích các đặc trưng; các lớp phi tuyến (nonlinear layers) (thường được áp dụng theo kiểu phần tử khôn ngoan), áp dụng một hàm kích hoạt (activation functions) tập trung vào các đặc trưng để cho phép mô hình hóa các đặc trưng phi tuyến tính của mạng; và lớp gộp (pooling layers), lấy một vùng lân cận nhỏ của bản đồ đặc trung (features map) và thay thế nó bằng một số thông tin thống kê (trung bình, tối đa, v.v.) của vùng lân cận. Các nút trong lớp CNN được kết nối cục bộ; nghĩa là, mỗi đơn vị trong một lớp nhận đầu vào từ một vùng lân cận nhỏ của lớp trước đó (được gọi là trường tiếp nhận). Ưu điểm chính của CNN là cơ chế chia sẻ trọng số thông qua việc sử dụng kernel trượt, nhân này đi qua các hình ảnh và tổng hợp thông tin cục bộ để trích xuất các tính năng. Vì trọng số hạt nhân được chia sẻ trên toàn bộ hình ảnh, CNN có số lượng tham số nhỏ hơn đáng kể so với một mạng neural được kết nối đầy đủ tương tự. Ngoài ra bằng cách xếp chồng nhiều lớp tích chập, các lớp cấp cao hơn sẽ học các đặc trưng từ các trường tiếp nhận ngày càng rộng hơn.

CNN đã được áp dụng cho các nhiệm vụ thị giác máy tính khác nhau như: phân đoạn ngữ nghĩa (semantic segmentation)[59], phân đoạn hình ảnh y tế (medical image segmentation) [60], phát hiện đối tượng (object detection)[61], siêu phân giải (super resolution)[62], nâng cao hình ảnh (image enhancement)[63], tạo phụ đề cho hình ảnh và video (caption generation for image and videos)[ 64], và nhiều hơn nữa. Một số kiến trúc CNN nổi tiếng nhất bao gồm AlexNet [47], ZFNet [65], VGGNet [66], ResNet [67], GoogLenet [68], MobileNet [69] và DenseNet [70].

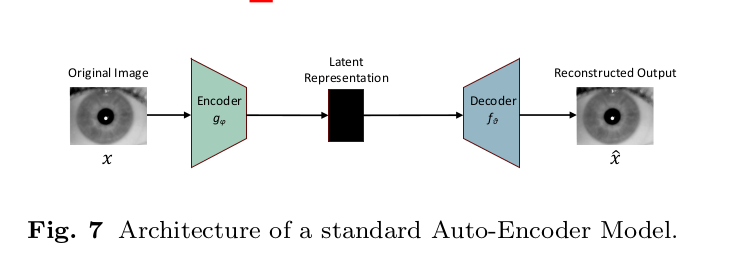
**2.2 Recurrent Neural Networks - Mạng hồi quy RNN và LSTM (Long Short Term Memory)**



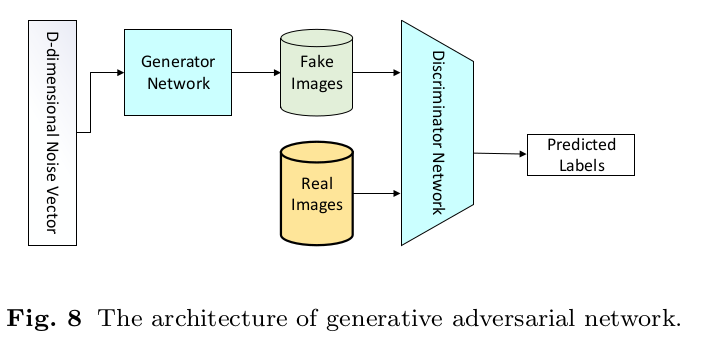




**2.3 Auto-Encoders**



**2.4 Generative Adversarial Networks (GAN)**



**2.5 Tiếp cận bằng Transfer Learning**

**3. Deep Learning dựa trên sinh trắc học (nhân trắc học)**

3.1 Face Recognition - Nhận dạng khuôn mặt

3.1.1 Face Datasets

- Yale and Yale Face Database B

- CMU Multi-PIE

- Labeled Face in The Wild (LFW)

- PolyU NIR Face Database

- YouTube Faces

- VGGFace2

- CASIA-WebFace

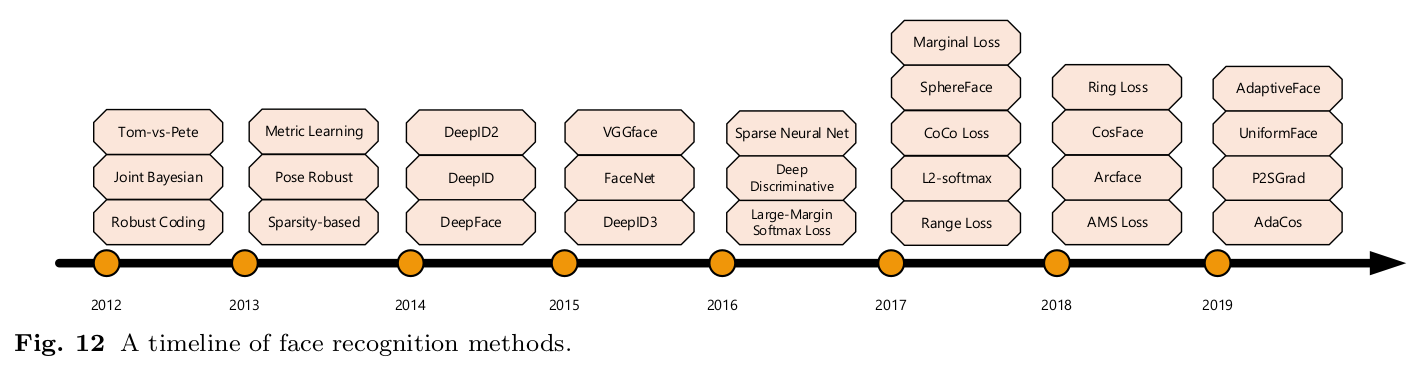
- MS-Celeb

- CelebA

- IJB-C

- MegaFace

3.1.2 Deep Learning Works on Face Recognition



3.2 Fingerprint Recognition - Nhận dạng dấu vân tay

3.2.1 Fingerprint Datasets

- FVC Fingerprint Database

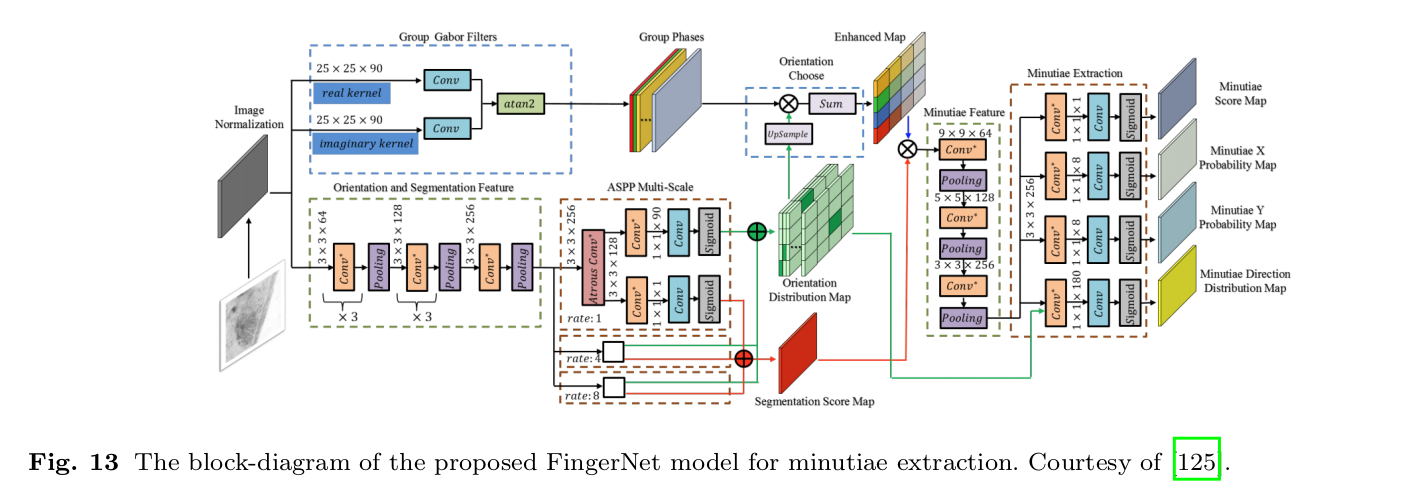
- PolyU High-resolution Fingerprint Database

- CASIA Fingerprint Dataset

- NIST Fingerprint Dataset

- NIST Fingerprint Dataset

3.2.2 Deep Learning Works on Fingerprint Recognition



3.3 Iris Recognition - Nhận dạng mắt

3.3.1 Iris Datasets

- CASIA-Iris-1000 Database

- UBIRIS Dataset

- IIT Delhi Iris Dataset

- ND Datasets

- MICHE Dataset

3.3.2 Deep Learning Works on Iris Recognition

3.4 Palmprint Recognition - Nhận dạng lòng bàn tay

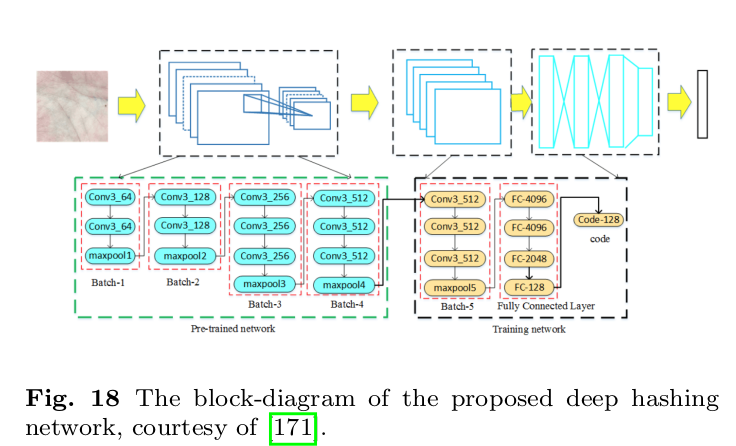
3.4.1 Palmprint Datasets

- PolyU Multispectral Palmprint Dataset

- CASIA Palmprint Database

- IIT Delhi Touchless Palmprint Database

3.4.2 Deep Learning Works on Palmprint Recognition



3.5 Ear Recognition - Nhận dạng tai

3.5.1 Ear Datasets

- IIT Ear Database

- AWE Ear Dataset

- Multi-PIE Ear Dataset

- USTB Ear Database

- UERC Ear Dataset

- AMI Ear Dataset

- CP Ear Dataset

- WPUT Ear Dataset

3.5.2 Deep Learning Works on Ear Recognition

**3.6 Voice Recognition - Nhận dạng giọng nói**

Nhận dạng giọng nói (còn được gọi là nhận dạng người nói) là tác vụ xác định ID của một người bằng cách sử dụng các đặc điểm của giọng nói của một người. Theo một cách nào đó, nhận dạng người nói bao gồm cả các đặc điểm hành vi và sinh lý, chẳng hạn như trọng âm và cao độ tương ứng. Việc sử dụng các cách tự động để thực hiện nhận dạng người nói có từ những năm 1960 khi Phòng thí nghiệm Bell được các cơ quan thực thi pháp luật tiếp cận về khả năng xác định những người gọi đã đe dọa đánh bom bằng lời nói qua điện thoại [194]. Trong những năm qua, các nhà nghiên cứu đã phát triển nhiều mô hình có thể thực hiện nhiệm vụ này một cách hiệu quả, đặc biệt là với sự trợ giúp của học sâu. Ngoài các ứng dụng bảo mật, nó cũng đang được sử dụng trong các trợ lý cá nhân ảo, chẳng hạn như Google Assistant, để họ có thể nhận ra và phân biệt giọng nói của chủ sở hữu điện thoại với những người khác [195].

Nhận dạng người nói có thể phân loại thành nhận biết người nói và xác minh người nói. Nhận biết người nói là quá trình xác định ID của một người từ một tập hợp giọng nói được đăng ký bằng cách sử dụng một nói nhất định [196], trong khi xác minh người nói là quá trình chấp nhận hoặc từ chối danh tính đề xuất được xác nhận cho một người nói [197]. Vì hai nhiệm vụ này thường chia sẻ cùng một quy trình đánh giá theo các thước đo thường được sử dụng, các thuật ngữ đôi khi được sử dụng thay thế cho nhau trong các bài báo tham chiếu. Việc nhận dạng người nói cũng liên quan chặt chẽ đến vấn đề speaker diarization, trong đó luồng âm thanh đầu vào được phân chia thành các đoạn đồng nhất theo nhận biết người nói [198].

**3.6.1 Voice Datasets**

- NIST SRE – Speakers Recognition Evaluation (SRE) Datasets: Bắt đầu từ năm 1996, Viện Tiêu chuẩn và Công nghệ Quốc gia (NIST - National Institute of Standards and Technology) đã tổ chức một loạt các đánh giá cho nghiên cứu nhận dạng người nói [199]. Do đó, bộ dữ liệu Đánh giá nhận dạng người nói (SRE) do NIST biên soạn đã trở thành tập dữ liệu được sử dụng rộng rãi nhất để đánh giá hệ thống nhận dạng người nói. Chúng được thu thập theo kiểu ngày càng phát triển và mỗi kế hoạch đánh giá có trọng tâm hơi khác nhau. Các bộ dữ liệu đánh giá này khác nhau về độ dài âm thanh [200], thiết bị ghi âm (điện thoại, thiết bị cầm tay và video) [201], nguồn gốc dữ liệu (ở Bắc Mỹ hoặc bên ngoài) 202 và các tình huống khớp / không khớp. Trong những năm gần đây, SRE 2016 và SRE 2018 là những bộ dữ liệu phổ biến nhất trong lĩnh vực này.

- SITW - Speakers in the Wild (SITW) tập dữ liệu được thu thập thông qua các điều kiện không bị giới hạn [203]. Không giống như các bộ dữ liệu SRE, dữ liệu này không được thu thập trong các điều kiện được kiểm soát và do đó chứa tiếng ồn và âm vang thực. Cơ sở dữ liệu bao gồm các bản ghi âm của 299 người nói, với trung bình tám phiên khác nhau cho mỗi người.

- VoxCeleb: VoxCeleb dataset [204] và Vox-Celeb2 dataset là các tập dữ liệu công khai được tổng hợp từ các video phỏng vấn được tải lên YouTube để nhấn mạnh việc thiếu dữ liệu quy mô lớn không bị giới hạn để nhận dạng người nói. Những dữ liệu này được thu thập bằng cách sử dụng một đường ống hoàn chỉnh. CNN đồng bộ hóa hai luồng được sử dụng để ước tính mối tương quan giữa đoạn âm thanh và chuyển động miệng của video, sau đó các kỹ thuật nhận dạng khuôn mặt dựa trên CNN được sử dụng để xác định người nói cho chú thích giọng nói. VoxCeleb1 có hơn 100.000 câu nói cho 1.251 người nổi tiếng và VoxCeleb2 chứa hơn một triệu câu nói cho 6.112 danh tính.

Ngoài các bộ dữ liệu được thiết kế hoàn toàn cho các tác vụ nhận dạng người nói, nhiều bộ dữ liệu được thu thập để nhận dạng giọng nói tự động cũng có thể được sử dụng để đào tạo hoặc đánh giá hệ thống nhận dạng người nói. Ví dụ, tập dữ liệu Switchboard [206] và Fisher Corpus [207], ban đầu được thu thập cho các nhiệm vụ nhận dạng giọng nói, cũng được sử dụng để đào tạo mô hình trong NIST Speaker Recognition Evaluations. Mặt khác, các nhà nghiên cứu có thể sử dụng bộ dữ liệu nhận dạng giọng nói hiện có để chuẩn bị bộ dữ liệu đánh giá nhận dạng người nói của riêng họ nhằm chứng minh tính hiệu quả của nghiên cứu của họ. Ví dụ: tập dữ liệu Librispeech [208] và tập dữ liệu TIMIT [209] được tác giả xử lý trước trong [210] để dùng làm tập đánh giá cho tác vụ nhận dạng người nói.

**3.6.2 Phương thức hoạt động của học sâu trong nhận dạng giọng nói (Deep learning work on Voice Recognition)**

Trước thời đại Deep Learning, hầu hết các hệ thống nhận dạng người nói hiện đại đều được xây dựng với phương pháp tiếp cận i-vectors [211], sử dụng phân tích dữ kiện để xác định không gian chiều thấp mô hình hóa cả sự biến đổi của loa và kênh. Trong những năm gần đây, việc khám phá các phương pháp học sâu để nhận dạng người nói ngày càng trở nên phổ biến. Một trong những cách tiếp cận đầu tiên trong số những nỗ lực này là kết hợp các mô hình âm học dựa trên DNN vào khung i-vector [212]. Phương pháp này sử dụng mô hình âm thanh DNN được đào tạo cho Nhận dạng giọng nói tự động (ASR) để thu thập số liệu thống kê của người nói cho đào tạo mô hình i-vector. Nó đã được chỉ ra rằng cải tiến này dẫn đến giảm 30% tỷ lệ lỗi tương đối.

Cùng lúc đó, d-vector đã được đề xuất trong [213] để giải quyết việc nhận dạng văn bản của người nói phụ thuộc vào mạng neural chúng ta. Trong cách tiếp cận này, một DNN được huấn luyện để phân loại người nói ở cấp khung. Trong quá trình đăng ký và kiểm tra, DNN được huấn luyện được sử dụng để rút trích các đặc trưng cụ thể của người nói từ lớp ẩn cuối cùng. Sau đó, “d-vector” được tính toán bằng cách lấy trung bình các đặc trưng này và được sử dụng như speaker embeddings để nhận dạng. Phương pháp này cho thấy sự cải thiện tương đối 14% và 25% so với hệ thống i-vector trong điều kiện sạch và ồn tương đối.

Trong [214], mạng neural có độ trễ thời gian được huấn luyện để rút trích “x-vector” ở mức phân đoạn để nhận dạng giọng nói độc lập với văn bản (text-independent speech recognition). Mạng này nhận các đặc trưng của phân đoạn giọng nói và chuyển chúng qua một vài lớp phi tuyến tính, tiếp theo là lớp tổng hợp để phân loại người nói ở cấp độ phân đoạn. X-Vector sau đó được rút trích từ lớp gộp để đăng ký và kiểm tra. Nó chỉ ra rằng hệ thống x-vector có thể đạt được hiệu suất nhận dạng người nói tốt hơn so với cách tiếp cận i-vector truyền thống, với sự trợ giúp của việc tăng cường dữ liệu.

Các phương pháp tiếp cận end-to-end dựa trên mạng neural cũng được khám phá trong nhiều bài báo khác nhau. Trong [215] và [216], mạng neural được thiết kế để nhận các cặp phân đoạn giọng nói và được huấn luyện để phân loại các mục tiêu khớp/ không khớp. Một bộ ba hàm mất mát (loss function) được thiết kế đặc biệt được đề xuất trong [217] để thay thế một hàm mất mát phân loại nhị phân. Hàm mất mát tổng quát end-to-end (GE2E), tương tự như bộ ba hàm mất mát, được đề xuất trong [218] để nhận dạng người nói phụ thuộc vào văn bản trên tập dữ liệu nội bộ.

Trong [219], một mục tiêu tối ưu hóa bổ sung được gọi là mất mát trong lớp được đề xuất để cải thiện deep speaker embeddings đã được huấn luyện được với bộ ba hàm mất mát. Bài báo chỉ ra rằng các mô hình được huấn luyện sử dụng mất mát trong lớp có thể mang lại mức giảm tương đối không đáng kể là 30% trong tỷ lệ lỗi tương đương EER) so với bộ ba mất mát ban đầu. Hiệu quả được đánh giá trên cả tập dữ liệu VoxCeleb và VoxForge.

Trong [210], các tác giả đã đề xuất một phương pháp học speaker embeddings từ dạng sóng thô bằng cách tối đa hóa thông tin lẫn nhau. Cách tiếp cận này sử dụng kiến trúc bộ encoder-discriminator tương tự như kiến trúc của Generative Adversarial Networks (GANs) để tối ưu hóa thông tin lẫn nhau một cách ngầm định. Các tác giả cho thấy rằng cách tiếp cận này học một cách hiệu quả các cách trình bày hữu ích của người nói, dẫn đến hiệu suất vượt trội trên tập dữ liệu VoxCeleb khi so sánh với hệ thống bộ ba mất mát dựa trên i-vector và CNN.

Trong [220], các tác giả kết hợp một bộ rút trích đặc trưng tích chập sâu (deep convolutional feature extractor), các chức năng tổng hợp và hàm mất mát biên độ lớn vào bộ nhận dạng người nói sâu từ đầu đến cuối của họ (end-to-end deep speaker recognizers). Các mô hình riêng lẻ và tập thể từ phương pháp này đã đạt được hiệu suất hiện đại trên VoxCeleb với mức cải thiện tương đối lần lượt là 70% và 82% so với kết quả được báo cáo tốt nhất. Các tác giả cũng đề xuất sử dụng mạng neural để bổ sung bộ phân loại PLDA, cho phép họ nhận được kết quả hiện đại nhất trên tập dữ liệu NIST-SRE 2016

[195] David B Yoffie, Liang Wu, Jodie Sweitzer, Denzil Eden, and Karan Ahuja. Voice war: Hey google vs. alexa vs.siri. 2018.

[196] Imran Naseem, Roberto Togneri, and Mohammed Bennamoun. Sparse representation for speaker identification. In 2010 20th International Conference on Pattern Recognition, pages 4460–4463. IEEE, 2010.

[197] S. Furui. Speaker recognition. Scholarpedia, 3(4):3715,2008. revision #64889.198. Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree. Speaker diarization using deep neural network embeddings. In International Conference on Acoustics, Speech and Signal Processing,

pages 4930–4934. IEEE, 2017.

[199] Alvin F Martin and Mark A Przybocki. The nist speaker recognition evaluations: 1996-2001. In 2001: A Speaker Odyssey-The Speaker Recognition Workshop, 2001.

[200] The 2010 nist speaker recognition evaluation. 2010.

[201] The 2018 nist speaker recognition evaluation. 2018.

[202] The 2016 nist speaker recognition evaluation. 2016.

[203] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson. The speakers in the wild (sitw) speaker recognition database. In Interspeech, pages 818–822, 2016.

[204] Arsha Nagrani, Joon Son Chung, and Andrew Zis- serman. Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612, 2017.

[205] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. ArXiv preprint arXiv:1806.05622, 2018.

[206] J Godfrey and E Holliman. Switchboard-1 release 2: Linguistic data consortium. SWITCHBOARD: A User’s Manual, 1997.

[207] Christopher Cieri, David Miller, and Kevin Walker. The fisher corpus: a resource for the next generations of speech-to-text. In LREC, volume 4, pages 69–71, 2004.

[208] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210. IEEE, 2015.

[209] Victor Zue, Stephanie Seneff, and James Glass. Speech database development at mit: Timit and beyond. Speech communication, 9(4):351–356, 1990.

[210] Mirco Ravanelli and Yoshua Bengio. Learning speaker representations with mutual information. ArXiv preprint arXiv:1812.00271, 2018.

[211] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 19(4):788–798, 2010.

[212] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In International Conference on Acoustics, Speech and Signal Processing, pages 1695–1699. IEEE, 2014.

[213] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In International Conference on Acoustics,

Speech and Signal Processing (ICASSP). IEEE, 2014.

3.7 Signature Recognition - Nhận dạng chữ ký

3.7.1 Signature Datasets

- ICDAR 2009 SVC

- SVC 2004

- Offline GPDS-960 Corpus

3.7.2 Deep Learning Works on Signature Recognition

3.8 Gait Recognition - Nhận dạng dáng đi

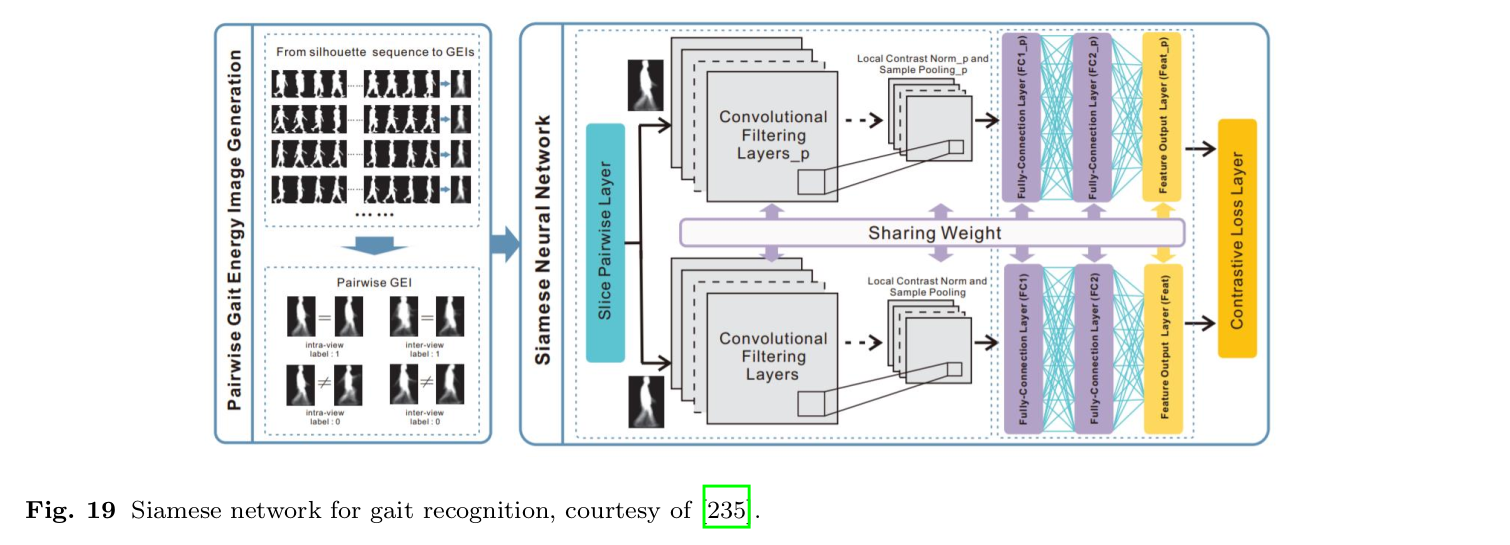
3.8.1 Gait Datasets

- CASIA Gait Database

- Osaka Treadmill Dataset

- Osaka University Large Population (OULP) Dataset

3.8.2 Deep Learning Works on Gait Recognition



4. Performance of Different Models on Different Datasets

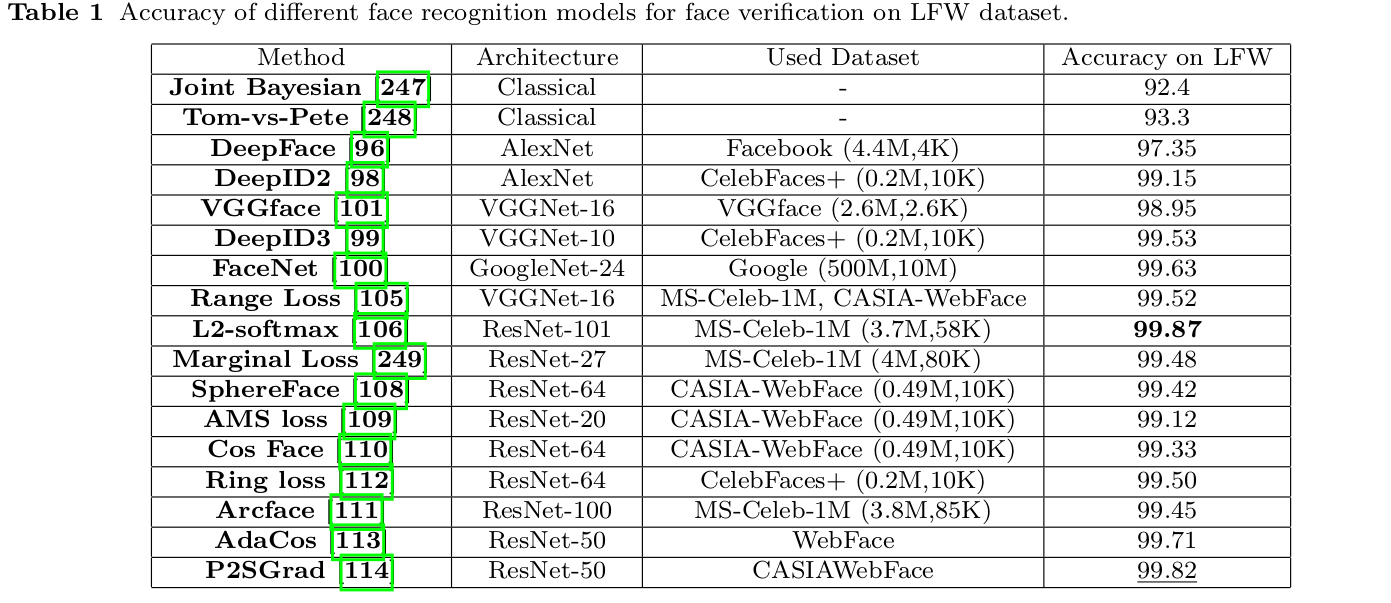
4.1 Popular Metrics For Evaluating Biometrics Recognition Systems

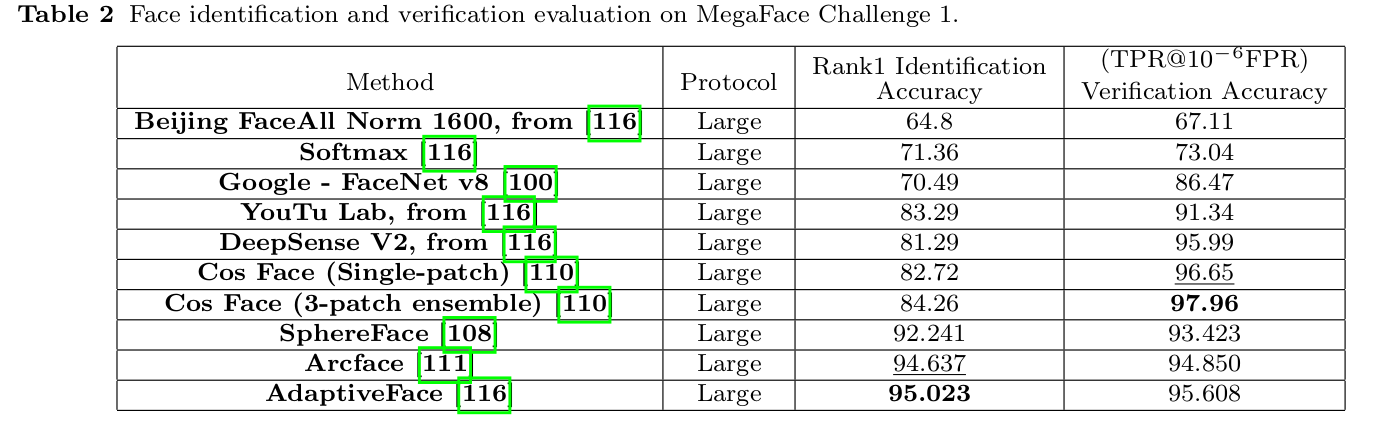
\* Biometric verification

\* Closed-set identification

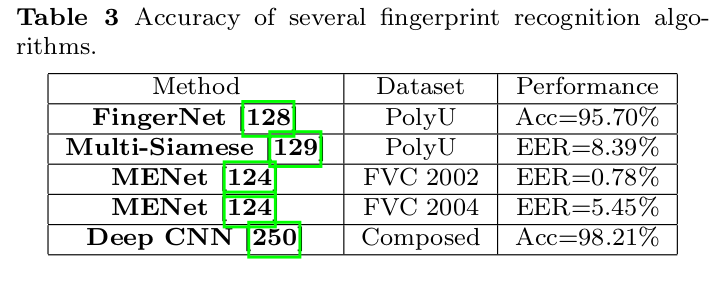
\* Open-set identification

\* Performance of Models for Face Recognition

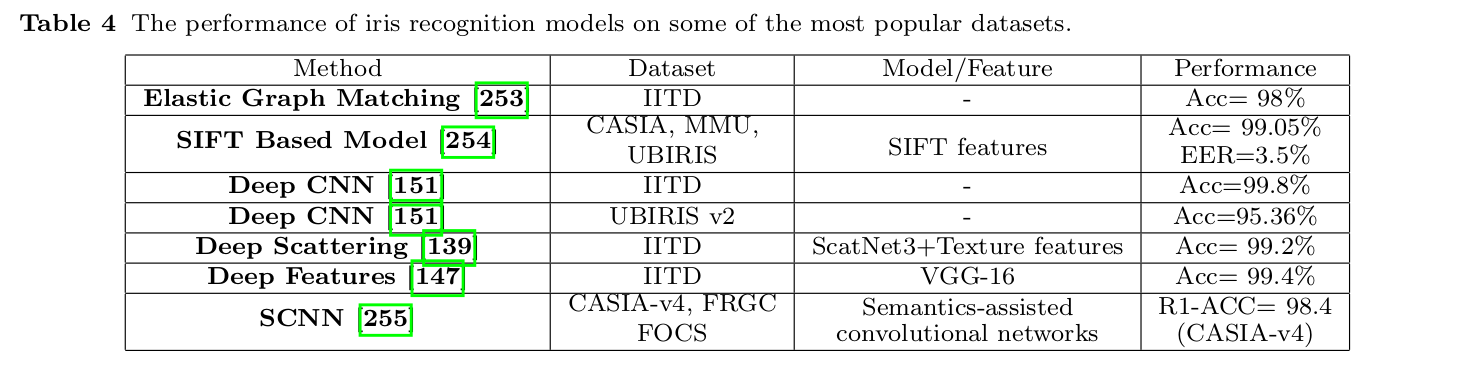




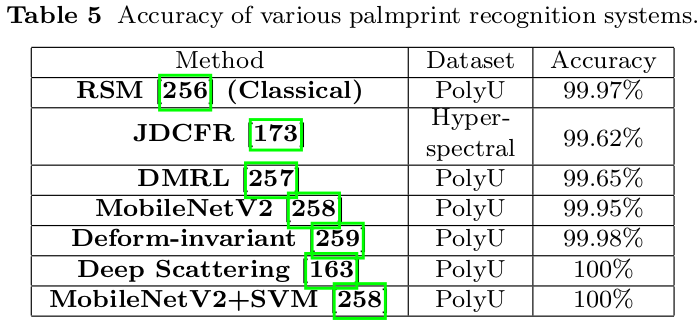
\* Performance of Models for Fingerprint Recognition



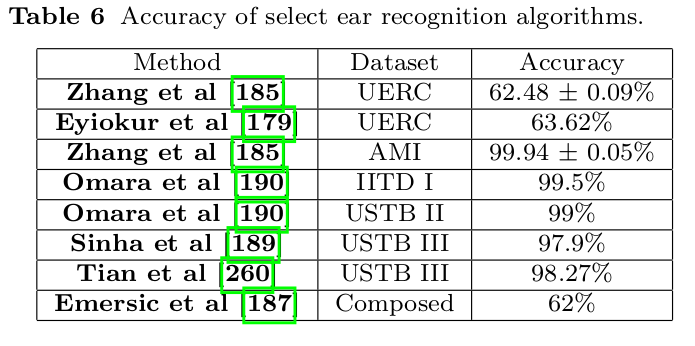
\* Performance of Models for Iris Recognition



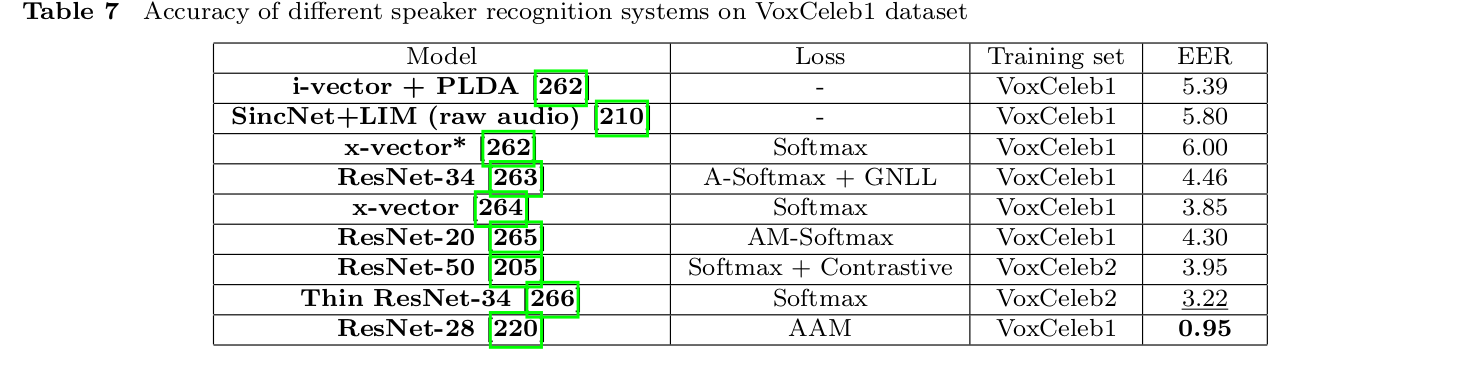
\* Performance of Models for Palmprint Recognition



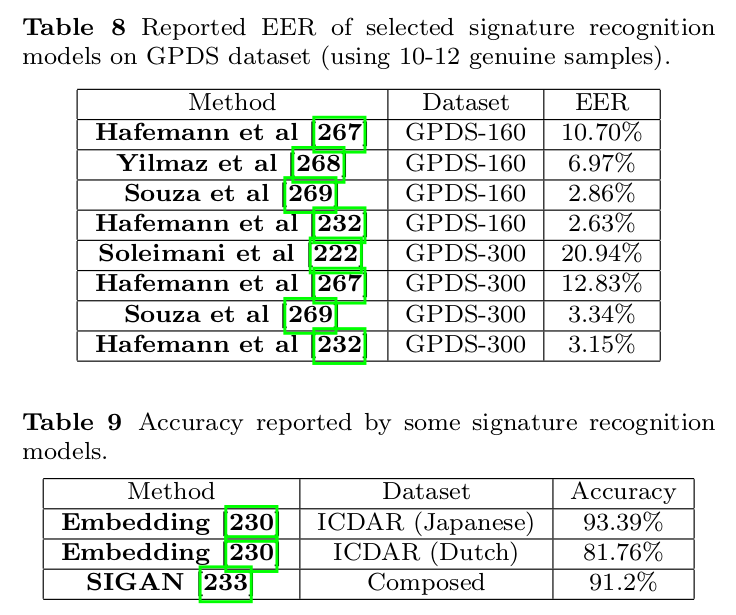
\* Performance of Models for Ear Recognition



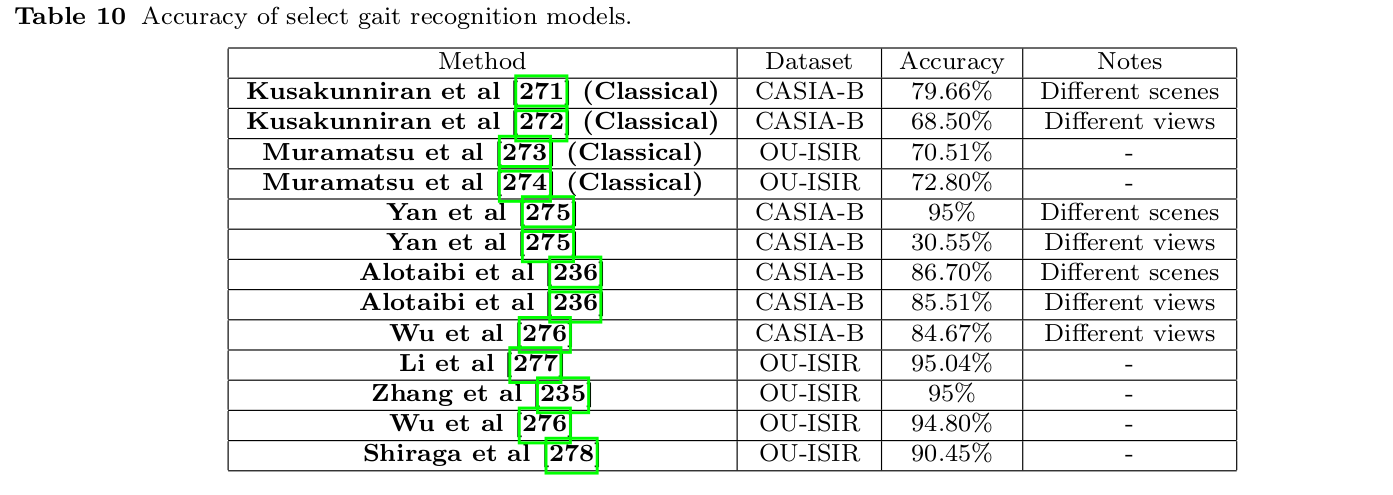
\* Performance of Models for Voice Recognition



\* Performance of Models for Signature Recognition



\* Performance of Models for Gait Recognition



5. Challenges and Opportunities

5.1 More Challenging Datasets

5.2 Interpretable Deep Models

5.3 Few Shot Learning, and Self-Supervised Learning

5.4 Biometric Fusion

5.5 Realtime Models for Various Applications

5.6 Memory Efficient Models

5.7 Security and Privacy Issues

6. Conclusion

7. Reference papers