SỰ PHÁT TRIỂN CỦA NHẬN DẠNG GIỌNG NÓI

1. **GIỚI THIỆU**

**II. KHÁI QUÁT VỀ NGÀNH**

Trong [khoa học máy tính](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_m%C3%A1y_t%C3%ADnh), trí tuệ nhân tạo hay AI ("ây-ai", [tiếng Anh](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh): *Artificial Intelligence*), đôi khi được gọi là trí thông minh nhân tạo, là [trí thông minh](https://vi.wikipedia.org/wiki/Tr%C3%AD_th%C3%B4ng_minh) được thể hiện bằng [máy móc](https://vi.wikipedia.org/wiki/M%C3%A1y_m%C3%B3c), trái ngược với trí thông minh tự nhiên của con người. Thông thường, thuật ngữ "trí tuệ nhân tạo" thường được sử dụng để mô tả các máy móc (hoặc máy tính) có khả năng bắt chước các chức năng "nhận thức" mà con người thường phải liên kết với [tâm trí](https://vi.wikipedia.org/wiki/T%C3%A2m_tr%C3%AD), như "học tập" và "giải quyết vấn đề".

Khi máy móc ngày càng tăng khả năng, các nhiệm vụ được coi là cần "trí thông minh" thường bị loại bỏ khỏi định nghĩa về AI, một hiện tượng được gọi là [hiệu ứng AI](https://vi.wikipedia.org/w/index.php?title=Hi%E1%BB%87u_%E1%BB%A9ng_AI&action=edit&redlink=1). Một câu châm ngôn trong Định lý của Tesler nói rằng "AI là bất cứ điều gì chưa được thực hiện" .Ví dụ, [nhận dạng ký tự quang học](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_k%C3%BD_t%E1%BB%B1_quang_h%E1%BB%8Dc) thường bị loại trừ khỏi những thứ được coi là AI, đã trở thành một công nghệ thông thường.[[4]](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o#cite_note-4) khả năng máy hiện đại thường được phân loại như AI bao gồm thành công hiểu lời nói của con người, cạnh tranh ở mức cao nhất trong [trò chơi chiến lược](https://vi.wikipedia.org/wiki/Tr%C3%B2_ch%C6%A1i_chi%E1%BA%BFn_l%C6%B0%E1%BB%A3c) (chẳng hạn như [cờ vua](https://vi.wikipedia.org/wiki/C%E1%BB%9D_vua) và [Go](https://vi.wikipedia.org/wiki/C%E1%BB%9D_v%C3%A2y)),  xe hoạt động độc lập, định tuyến thông minh trong mạng phân phối nội dung, và mô phỏng quân sự.

Trí tuệ nhân tạo có thể được phân thành ba loại hệ thống khác nhau: trí tuệ nhân tạo phân tích, lấy cảm hứng từ con người và nhân tạo. AI phân tích chỉ có các đặc điểm phù hợp với [trí tuệ nhận thức](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_th%E1%BB%A9c); tạo ra một đại diện nhận thức về thế giới và sử dụng học tập dựa trên kinh nghiệm trong quá khứ để thông báo các quyết định trong tương lai. AI lấy cảm hứng từ con người có các yếu tố từ [trí tuệ](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_x%C3%BAc_c%E1%BA%A3m) nhận thức và [cảm xúc](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_x%C3%BAc_c%E1%BA%A3m); hiểu cảm xúc của con người, ngoài các yếu tố nhận thức và xem xét chúng trong việc [ra quyết định](https://vi.wikipedia.org/wiki/Ra_quy%E1%BA%BFt_%C4%91%E1%BB%8Bnh). AI nhân cách hóa cho thấy các đặc điểm của tất cả các loại năng lực (nghĩa là trí tuệ nhận thức, cảm xúc và xã hội), có khả năng tự ý thức và tự nhận thức được trong các tương tác.

Trí tuệ nhân tạo được thành lập như một môn học thuật vào năm 1956, và trong những năm sau đó đã trải qua nhiều làn sóng lạc quan, sau đó là sự thất vọng và mất kinh phí (được gọi là " mùa đông AI "),tiếp theo là cách tiếp cận mới, thành công và tài trợ mới. Trong phần lớn lịch sử của mình, nghiên cứu AI đã được chia thành các trường con thường không liên lạc được với nhau.[[12]](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o#cite_note-Fragmentation_of_AI-12) Các trường con này dựa trên các cân nhắc kỹ thuật, chẳng hạn như các mục tiêu cụ thể (ví dụ: " [robot học](https://vi.wikipedia.org/wiki/Robot_h%E1%BB%8Dc) " hoặc "học máy"), việc sử dụng các công cụ cụ thể ("logic" hoặc [mạng lưới thần kinh nhân tạo](https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_n%C6%A1-ron_nh%C3%A2n_t%E1%BA%A1o)) hoặc sự khác biệt triết học sâu sắc. Các ngành con cũng được dựa trên các yếu tố xã hội (các tổ chức cụ thể hoặc công việc của các nhà nghiên cứu cụ thể).

Lĩnh vực này được thành lập dựa trên tuyên bố rằng [trí thông minh của con người](https://vi.wikipedia.org/wiki/Tr%C3%AD_th%C3%B4ng_minh_c%E1%BB%A7a_con_ng%C6%B0%E1%BB%9Di) "có thể được mô tả chính xác đến mức một cỗ máy có thể được chế tạo để mô phỏng nó". Điều này làm dấy lên những tranh luận triết học về bản chất của [tâm trí](https://vi.wikipedia.org/wiki/T%C3%A2m_tr%C3%AD) và đạo đức khi tạo ra những sinh vật nhân tạo có trí thông minh giống con người, đó là những vấn đề đã được [thần thoại](https://vi.wikipedia.org/wiki/L%E1%BB%8Bch_s%E1%BB%AD_ng%C3%A0nh_tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o), viễn tưởng và [triết học](https://vi.wikipedia.org/wiki/Tri%E1%BA%BFt_h%E1%BB%8Dc_v%E1%BB%81_tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o) từ [thời cổ đại](https://vi.wikipedia.org/wiki/Th%E1%BB%9Di_k%E1%BB%B3_c%E1%BB%95_%C4%91%E1%BA%A1i) đề cập tới. Một số người cũng coi AI là [mối nguy hiểm cho nhân loại](https://vi.wikipedia.org/wiki/%C4%90i%E1%BB%83m_k%E1%BB%B3_d%E1%BB%8B_c%C3%B4ng_ngh%E1%BB%87) nếu tiến triển của nó không suy giảm. Những người khác tin rằng AI, không giống như các cuộc cách mạng công nghệ trước đây, sẽ tạo ra [nguy cơ thất nghiệp hàng loạt](https://vi.wikipedia.org/w/index.php?title=Th%E1%BA%A5t_nghi%E1%BB%87p_c%C3%B4ng_ngh%E1%BB%87&action=edit&redlink=1).

Trong thế kỷ 21, các kỹ thuật AI đã trải qua sự hồi sinh sau những tiến bộ đồng thời về sức mạnh máy tính, [dữ liệu lớn](https://vi.wikipedia.org/wiki/D%E1%BB%AF_li%E1%BB%87u_l%E1%BB%9Bn) và hiểu biết lý thuyết; và kỹ thuật AI đã trở thành một phần thiết yếu của [ngành công nghệ](https://vi.wikipedia.org/w/index.php?title=C%C3%B4ng_nghi%E1%BB%87p_c%C3%B4ng_ngh%E1%BB%87&action=edit&redlink=1), giúp giải quyết nhiều vấn đề thách thức trong học máy, [công nghệ phần mềm](https://vi.wikipedia.org/wiki/C%C3%B4ng_ngh%E1%BB%87_ph%E1%BA%A7n_m%E1%BB%81m) và [nghiên cứu vận hành](https://vi.wikipedia.org/wiki/V%E1%BA%ADn_tr%C3%B9_h%E1%BB%8Dc).

1. **KHÁI QUÁT VỀ ĐỀ TÀI**

Nhận dạng giọng nói là quá trình máy tính (hoặc loại máy khác) nhận dạng giọng nói từ ngữ. Về cơ bản, nó có nghĩa là nói chuyện với máy tính của bạn và để nó nhận ra chính xác bạn đang nói gì.

Nhận dạng giọng nói. Công nghệ này không còn là một khái niệm mới mẻ, tất cả những ông lớn ngành công nghệ đều đã đang lấn sân vào “cuộc chơi” này. Đó là trợ lý ảo Siri của Apple, Cortana của Microsoft, Alexa của Amazon, đến cả Samsung cũng chập chững cho thai nghen trợ lý Bixby của riêng mình hay không thể không kể đến Google Assistant của Google.

Nhận dạng tiếng nói là một quá trình [nhận dạng mẫu](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_m%E1%BA%ABu), với mục đích là phân lớp (classify) thông tin đầu vào là tín hiệu tiếng nói thành một dãy tuần tự các mẫu đã được học trước đó và lưu trữ trong bộ nhớ. Các mẫu là các đơn vị nhận dạng, chúng có thể là các từ, hoặc các [âm vị](https://vi.wikipedia.org/wiki/%C3%82m_v%E1%BB%8B). Nếu các mẫu này là bất biến và không thay đổi thì công việc nhận dạng tiếng nói trở nên đơn giản bằng cách so sánh dữ liệu tiếng nói cần nhận dạng với các mẫu đã được học và lưu trữ trong bộ nhớ. Khó khăn cơ bản của nhận dạng tiếng nói đó là tiếng nói luôn biến thiên theo thời gian và có sự khác biệt lớn giữa tiếng nói của những người nói khác nhau, tốc độ nói, ngữ cảnh và môi trường âm học khác nhau. Xác định những thông tin biến thiên nào của tiếng nói là có ích và những thông tin nào là không có ích đối với nhận dạng tiếng nói là rất quan trọng. Đây là một nhiệm vụ rất khó khăn mà ngay cả với các kỹ thuật xác suất thống kê mạnh cũng khó khăn trong việc tổng quát hoá từ các mẫu tiếng nói những biến thiên quan trọng cần thiết trong nhận dạng tiếng nói.

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

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

Cách tiếp cận nhận dạng tiếng nói bằng thống kê bao gồm: sử dụng [mô hình Markov ẩn](https://vi.wikipedia.org/wiki/M%C3%B4_h%C3%ACnh_Markov_%E1%BA%A9n), [mạng nơ-ron](https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_n%C6%A1-ron), sử dụng [cơ sở tri thức](https://vi.wikipedia.org/wiki/C%C6%A1_s%E1%BB%9F_tri_th%E1%BB%A9c), v.v..

Các định nghĩa sau đây là những điều cơ bản cần thiết để hiểu công nghệ nhận dạng giọng nói:

Phát âm là sự phát âm (nói) của một từ hoặc những từ biểu thị một ý nghĩa duy nhất cho máy vi tính. Các lỗi sai có thể là một từ đơn, một vài từ, một câu hoặc thậm chí nhiều câu.

Sự phụ thuộc của người nói: Các hệ thống phụ thuộc vào loa được thiết kế xung quanh một loa cụ thể. Chúng thường nhiều hơn chính xác đối với người nói chính xác, nhưng kém chính xác hơn nhiều đối với người nói khác. Họ cho rằng người nói sẽ nói với một giọng nói và nhịp độ nhất quán. Hệ thống loa độc lập được thiết kế cho nhiều loại của người nói. Hệ thống thích ứng thường bắt đầu như hệ thống độc lập với loa và sử dụng đào tạo kỹ thuật thích ứng với người nói để tăng độ chính xác nhận dạng của họ.

Từ vựng: Từ vựng (hoặc từ điển) là danh sách các từ hoặc cách phát biểu có thể được SR nhận dạng hệ thống. Nói chung, các từ vựng nhỏ hơn sẽ dễ dàng hơn để máy tính nhận ra, trong khi các từ vựng lớn hơn từ vựng khó hơn. Không giống như các từ điển thông thường, mỗi mục nhập không nhất thiết phải là một từ ngữ. Chúng có thể dài bằng một hoặc hai câu. Các từ vựng nhỏ hơn có thể chỉ có 1 hoặc 2 các từ được công nhận (ví dụ: "Thức dậy"), trong khi các từ vựng rất lớn có thể có hàng trăm nghìn hoặc nhiều hơn!

Cộng dồn: Khả năng của một trình nhận dạng có thể được kiểm tra bằng cách đo độ chính xác của nó - hoặc mức độ nhận biết của nó lời nói. Điều này không chỉ bao gồm việc xác định chính xác một câu nói mà còn xác định xem cách nói không có trong từ vựng của nó. Hệ thống ASR tốt có độ chính xác từ 98% trở lên! Các độ chính xác chấp nhận được của một hệ thống thực sự phụ thuộc vào ứng dụng.

Đào tạo

Một số công cụ nhận dạng giọng nói có khả năng thích ứng với người nói. Khi hệ thống có khả năng này, nó có thể cho phép đào tạo diễn ra. Hệ thống ASR được đào tạo bằng cách có tiêu chuẩn lặp lại người nói hoặc các cụm từ phổ biến và điều chỉnh thuật toán so sánh của nó để phù hợp với người nói cụ thể đó.

Đào tạo trình nhận dạng thường cải thiện độ chính xác của nó. Những người nói khó nói hoặc phát âm một số từ nhất định cũng có thể áp dụng phương pháp đào tạo. Miễn là người nói có thể lặp lại một cách nhất quán một câu nói, hệ thống ASR có đào tạo phải có khả năng thích ứng.

1. **KHÁI NIỆM/ ĐỊNH NGHĨA**
2. PHÁT TRIỂN CỦA
3. CẤU TRÚC