TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

──────── \* ───────



**Báo cáo thực tập**

**ĐỀ TÀI :**

TỔNG HỢP TIẾNG VIỆT

Giáo viên hướng dẫn:

**Thầy Dư Thanh Bình**

Sinh viên thực hiện:

**Phí Tùng Lâm 20071670 KTMT - K52**

**Nguyễn Trung Dũng 20070583 KTMT - K52**

Mục lục

[**1.** **Lịch sử** 3](#_Toc316591011)

[**2.** **Khái quát về tổng hợp tiếng nói từ văn bản TTS (Text To Speech)** 5](#_Toc316591012)

[**2.1.** **Các loại âm** 5](#_Toc316591013)

[**2.2.** **Các tham số âm thanh** 6](#_Toc316591014)

[**2.3.** **Các khái niệm ngữ âm** 7](#_Toc316591015)

[**3.** **Tổng hợp tiếng nói từ văn bản** 8](#_Toc316591016)

[**Các quá trình chuyển đổi** 8](#_Toc316591017)

[**4.** **Các phương pháp tổng hợp tiếng nói** 9](#_Toc316591018)

[**4.1.** **Tổng hợp theo cấu âm** 9](#_Toc316591019)

[**4.2.** **Tổng hợp formant theo quy luật** 12](#_Toc316591020)

[**4.3.** **Tổng hợp xích chuỗi** 13](#_Toc316591021)

[**5.** **So sánh các phương pháp tổng hợp tiếng nói.** 18](#_Toc316591022)

1. **Lịch sử**

Tạo tiếng nói bằng máy là một ước mơ của loài người và là mục tiêu thách thức các nhà khoa học trên thế giới nhiều thế kỷ qua. Máy nói bằng cơ khí đầu tiên do Wolfgang Von Kempelen chế tạo năm 1791 mô phỏng bộ máy phát âm của người đã có thể phát ra được một vài âm tố. Ở đầu thế kỷ này, vào năm 1939, máy VODER của nhà khoa học Dudley, phòng thí nghiệm AT&T BELL LABS, đã thành công trong việc mô hình hoá cộng hưởng của ống thanh bằng linh kiện điện tử. Hệ thống điều khiển nhân công 10 bộ cộng hưởng bằng các phím và chiết áp đã có thể tạo ra một số câu nói nghe hiểu được. Từ giữa những năm 1950 trở lại đây, sự ra đời của máy tính điện tử số đã cho phép tạo ra một cách có hệ thống các tham số điều khiển bộ tổng hợp tiếng nói. Cùng với sự phát triển như vũ bão của khoa học và công nghệ ở thập kỷ 80 và 90, kĩ thuật tổng hợp tiếng nói đã có những bước phát triển rất quan trong có sự đóng góp rất lớn của máy tính và xử lý tín hiệu số.

Sự tiến bộ này được thúc đẩy do:

* Khả năng tính toán của máy tính nhanh lên gấp nhiều lần, số lượng dữ liệu được lưu giữ ngày càng lớn và giá thành ngày càng hạ.
* Khả năng xây dựng và truy cập tới các cơ sở dữ liệu âm thanh và văn bản ngày càng dễ dàng.
* Công nghệ nhận dạng tự động tiếng nói được hoàn thiện với độ chính xác ngày càng cao hơn trợ giúp việc tự động phân đoạn âm tiếng nói cho phép tạo ra các cơ sở dữ liệu ngữ âm rất lớn phục vụ cho phân tích tiếng nói.

Nếu tiếng nói cũng như chữ viết là chuỗi các dấu cơ bản được định nghĩa rõ ràng thì các vấn đề về nhận dạng và tổng hợp tiếng nói đã có thể đã được giải quyết từ lâu. Bản chất tiếng nói rất khác biệt, hiệu ứng liên cấu âm quyện các âm tố thành dòng thanh liên tục rất phức tạp mang thông tin ở mọi thành phần của nó nên tổng hợp và nhận dạng tiếng nói trở nên đặc biệt khó khăn.

Việc tổng hợp tiếng nói từ bất kỳ văn bản nào, tổng hợp với số lượng từ không hạn chế của một ngôn ngữ thành tiếng nói trôi chảy, có độ tự nhiên như con người vẫn là vấn đề chưa được giải quyết và đang là một lĩnh vực được quan tâm nghiên cứu nhiều.

Chất lượng âm thanh của các hệ thống tổng hợp tiếng nói từ văn bản đến nay còn rất hạn chế so với tiếng nói tự nhiên của con người. Mặc dù mới chỉ đạt mức độ nghe - hiểu rõ nhưng tổng hợp tiếng nói cũng đã tìm thấy ứng dụng trong nhiều lĩnh vực khác nhau. Chẳng hạn:

* Trợ giúp cho người tàn tật

Hệ thống TTS kết hợp cùng với hệ thống quang học nhận dạng chữ viết

(OCR) có thể đọc các văn bản in cho người khiếm thị.

* Truyền thông tin bằng âm thanh

Một hệ thống TTS rất hữu ích cho việc kiểm tra chính tả các văn bản

trên máy tính vì khi nghe, phát hiện lỗi dễ dàng hơn nhiều so với đọc bằng mắt.

* Ứng dụng trong Viễn thông
  + Các hệ thống thông tin tích hợp (Unified Messaging) có khả năng đọc tự động nội dùng các bức thư điện tử E-Mail và các bức FAX và các thông tin trên Internet qua máy điện thoại cố định và di động thông thường.
  + Truy cập thông tin, cơ sở dữ liệu qua điện thoại và tự động tra cứu bằng máy điện thoại…

Tổng hợp tiếng nói từ văn bản là một ngành khoa học liên quan đến nhiều chuyên ngành như ngôn ngữ, ngữ âm, xử lý tín hiệu số và khoa học máy tính. Hiện nay, với nỗ lực nghiên cứu của các nhà khoa học công nghệ trên thế giới, giao diện giữa người và máy ngày càng thân thiện hơn. Xu thế giao lưu quốc tế đã xuất hiện ngày một nhiều các ứng dụng. Chất lượng âm thanh tổng hợp hiện tại đã vượt qua ngưỡng “nghe hiểu được” và đang tiến dần đến mức độ tự nhiên của tiếng nói con người. Trong một số lĩnh vực ứng dụng ngôn ngữ hẹp, chất lượng âm thanh tổng hợp đã không thể phân biệt được với tiếng nói tự nhiên.

Công nghệ để phát triển các hệ thống tổng hợp tiếng nói từ văn bản có số từ vựng không hạn chế không những rất phức tạp, mà còn phụ thuộc rất nhiều vào đặc trưng của ngôn ngữ tổng hợp. Nói cách khác, mặc dầu về nguyên tắc các phương pháp cơ bản để tổng hợp tiếng nói là tương tự như nhau nhưng mỗi loại ngôn ngữ đều có những đặc trưng ngữ âm riêng biệt, các quy luật cấu âm tạo thành từng âm tố của mỗi ngôn ngữ là không giống nhau. Như vậy, các ngôn ngữ khác nhau sẽ phải có các công nghệ tổng hợp khác nhau, theo các luật tổng hợp khác nhau. Điều này có nghĩa là không thể lấy hệ thống tổng hợp tiếng nước này để tổng hợp tiếng nước khác hoặc ngược lại. Nói cách khác là không thể lấy bộ tổng hợp cho các ngôn ngữ đa âm Âu - Ấn để tổng hợp nên tiếng Việt, một ngôn ngữ đơn âm có thanh điệu.Khái niệm tổng hợp tiếng nói tương đối rộng. Nó có thể bao gồm cả các hệ thống ghép từ đơn giản với số lượng xử lý tính toán thấp, ví dụ như hệ thống thông báo giờ của tổng đài điện thoại.

Tổng hợp theo cách từng từ, có thể tạo ra tiếng nói với chất lượng âm thanh cao nhưng chỉ được sử dụng khi tập bản tin cần tạo ra nhỏ với trật tự các từ tạo thành được định ra trước đó.

1. **Khái quát về tổng hợp tiếng nói từ văn bản TTS (Text To Speech)**

Vấn đề đặt ra khi thiết kế hệ thống tổng hợp tiếng Việt là số lượng cho mỗi một đơn vị âm. Với mỗi đơn vị âm, chúng ta cần bao nhiêu mẫu để đạt được chất lượng “tối ưu” cho một ứng dụng đích. Các tiêu chí quan trọng được cân nhắc “tối ưu” ở đây là:

* Dung lượng bộ nhớ cần thiết để lưu trữ dạng sóng đơn vị âm cũng như các tham số về phổ và ngữ điệu của chúng.
* Thời gian cần thiết để xây dựng hệ thống hay khi phải thiết lập một giọng nói mới. Khác với tổng hợp formant theo quy luật, có thể thay đổi tương đối linh hoạt giọng nói, tổng hợp xích chuỗi phải xây dựng cho mỗi giọng nói một CSDL. Trên lý thuyết, có thể thông qua xử lý tín hiệu để thay đổi giọng nói nhưng chất lượng tiếng nói tổng hợp giảm đi nhiều, vì vậy chưa được áp dụng thực tế.
* Khả năng xử lý thời gian thực và đồng thời nhiều tiến trình tổng hợp trên một máy tính PC. Để tổng hợp tiếng Việt, chúng ta có thể xem xét lựa chọn một số lượng tệp đơn vị tổng hợp với ba khả năng:
  + Thứ nhất với số lượng âm tối thiểu, các đơn vị sẽ là phần phụ âm đầu và phần vần không mang thanh điệu. Tần số cơ bản F0 của các vần là đặc trưng cho thanh điệu có thể thay đổi từ các vần đã có, tổng hợp theo các đường nét đặc trưng tần số cơ bản F0 khác đi thành các vần mang thanh điệu mới được không? Từ đó, chúng ta sẽ tổng hợp được các từ mới.
  + Thứ hai là số lượng đủ cho các vần mang dấu. Khi ấy bộ tổng hợp phải thay đổi các thành phần ngữ điệu như độ dài âm tiết và thanh điệu theo các quy luật biến thanh và quy luật về độ dài của âm tiết.
  + Thứ ba là tổng hợp trên cơ sở số liệu của nhiều mẫu cho một đơn vị âm theo ngữ cảnh khác nhau và các thuật toán tự động để điều khiển được sự xích chuỗi theo ngữ âm mong muốn.
  1. **Các loại âm**

Các nhà ngữ âm học thường phân chia các loại âm của tiếng nói theo vị trí cấu âm và phương thức cấu âm tạo nên chúng. Phân chia theo phương thức cấu âm, tiếng nói được phân làm nguyên âm, phụ âm tắc, phụ âm xát, phụ âm tắc xát và phụ âm mũi…

Nguyên âm khác với phụ âm về độ mở của ống thanh. Nếu ống thanh được mở toàn bộ để cho không khí từ phổi thoát ra hoàn toàn không gặp sự ngăn cản nào thì sẽ tạo ra nguyên âm. Khoang miệng chỉ còn nhiệm vụ để thay đổi âm sắc.

* + Ngược lại với nguyên âm, khi ống thanh bị thu hẹp lại hay tạm thời khép lại, dòng không khí bị hạn chế tạo ra tiếng ồn là phụ âm, khi ấy các khoang âm trở thành cơ quan chính tạo tiếng nói.
  + Trong trường hợp các âm mũi, khi lưỡi gà đóng lại thì khoang mũi trở thành ống dẫn thanh chủ yếu.
  + Âm xát khởi đầu bằng sự thắt lại của ống thanh, hoặc là ở thanh môn, ngạc cứng, răng hay môi.
  + Âm xát vô thanh được tạo ra bởi khí thay đổi bất thường từ thanh môn.
  + Âm xát hữu thanh bao gồm cả phần tuần hoàn và phần không tuần hoàn, có nghĩa là dây thanh có chấn động nhưng thanh môn không được đóng hẳn lại.
  + Âm tắc được hình thành theo ba bước. Đầu tiên, ống thanh được đóng lại ở một điểm nào đó, áp suất không khí trước đó tăng lên và sau đó đột ngột thoát ra tạo thành âm nổ trong một giai đoạn ngắn.
  + Âm tắc hữu thanh khác với âm tắc vô thanh ở chỗ là có hay không có sự dao động của dây thanh trong quá trình nén khí.
  + Âm tắc xát bao gồm âm tắc và tiếp sau đó là âm xát tuy nhiên thời gian tắc rất ngắn.
  1. **Các tham số âm thanh**

Các bộ phân tích tín hiệu xử lý âm thanh tiếng nói lấy ra các đặc trưng của tín hiệu như tần số âm cơ bản F0, tần số formant, cường độ âm hay các tham số từ phân tích phổ như LPC, Cepstrum… phục vụ cho quá trình nghiên cứu về ngữ âm âm thanh, mã hoá, tổng hợp và nhận dạng tiếng nói.

Khi quan sát dạng sóng tín hiệu âm thanh tiếng nói ta thấy có các đoạn tín hiệu biên độ lớn, tương đối tuần hoàn. Đó là phần âm hữu thanh của tiếng nói. Còn các âm đoạn như nhiễu băng rộng có biên độ thấp là phần tín hiệu vô thanh. Ở phần hữu thanh, tần số âm cơ bản F0 thay đổi chậm và có giá trị ở khoảng từ 120Hz đến 200Hz cho nam, 150Hz đến 400Hz cho nữ.

* 1. **Các khái niệm ngữ âm**

Âm vị là một đơn vị trừu tượng nhỏ nhất của cơ cấu âm thanh ngôn ngữ, được định nghĩa làm chức năng phân biệt hình thức ngữ âm của những đơn vị có nghĩa như từ và hình vị.

Âm tố: Một âm thanh vật lý được tạo ra khi cấu âm một âm vị được gọi là âm tố. Âm tố là đơn vị âm thanh nhỏ nhất của tiếng nói có thể tách ra về mặt cấu âm - thính giác.

Âm tiết là đơn vị phát âm nhỏ nhất được phân định tự nhiên trong tiếng nói của con người. Biến thể âm vị hình thành ra do có nhiều dạng cấu âm để phát thành âm cho một âm vị. Biến thể tự do là những biến thể không mang thông tin về ngữ nghĩa (như tiếng địa phương). Biến thể kết hợp do ảnh hưởng bởi động học của ống thanh khi tạo ra các âm xung quanh. Đó là các hiện tượng giảm âm, nuốt âm, thay đổi dạng âm từ hữu thanh sang vô thanh và ngược lại…

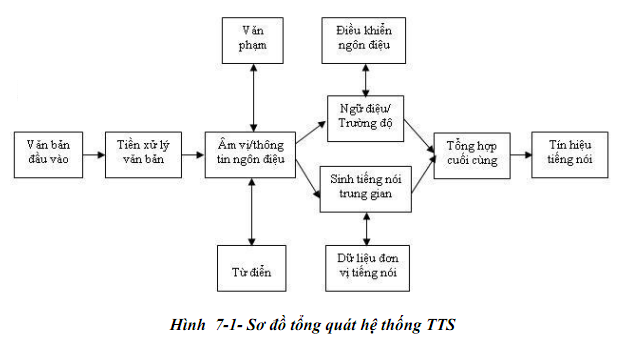
Liên cấu âm: Hiện tượng liên cấu âm xảy ra khi các cơ quan phát âm chuyển liên tục ở các vị trí để cấu âm từ âm vị này đến âm vị khác. Mức độ liên cấu âm phụ thuộc vào tốc độ nói và kiểu nói. Trong tiếng nói của con người, sự chuyển tiếp các động tác cấu âm từ vị trí này sang vị trí khác tạo ra độ tự nhiên cho tiếng nói.

Ngữ điệu: Tiếng nói không chỉ bao gồm chuỗi các động tác cấu âm mà còn kèm cả những điểm nghỉ lấy hơi, nâng cao hay hạ thấp giọng, tăng hay giảm cường độ âm thanh. Các biến đổi này đều có chứa thông tin, mang ý nghĩa biểu cảm nhất định.

Thanh điệu: Việc tăng, giảm, uốn và nhấn trên một âm vị tạo ra thông tin mới cho từ trong quá trình diễn đạt ngôn ngữ ta gọi là thanh điệu. Tiếng Việt, Trung Quốc, Thái Lan,… là những ngôn ngữ có thanh điệu. Thanh điệu được tạo ra từ sự thay đổi tần số cơ bản F0.

Vần: Tiếng nói có thanh điệu thường được liên kết một nguyên âm với một phụ âm cuối hay độc lập nguyên âm, bán âm cuối, nguyên âm ghép cộng với thanh tạo nên vần. Khi nghiên cứu ngôn ngữ có thanh điệu, vần được xem là một cấu thành rất quan trong, có nhiều ý nghĩa trong lĩnh vực tổng hợp tiếng nói.

1. **Tổng hợp tiếng nói từ văn bản**

****

TTS bao gồm việc chuyển đổi các mô tả về ngữ âm và ngữ điệu một văn bản ở đầu vào thành dạng sóng tiếng nói, sử dụng các thuật toán xử lý tín hiệu số. Các số liệu về tiếng nói có thể được mã hoá hay tham số hoá theo một dạng nào đó và được lưu giữ trước đó trong bộ nhớ. Văn bản có thể được nhập vào máy tính bằng bàn phím, bằng bộ nhận dạng tự động chữ viết, truy xuất từ cớ sở dữ liệu hay truyền về từ mạng Internet.

Một hệ thống TTS bao gồm hai quá trình tương đối độc lập có liên quan đến nhau là phần xử lý văn bản và phần xử lý tín hiệu số. Đầu tiên, hệ thống phải phân tích văn bản đầu vào tạo ra mô tả về ngữ âm và ngữ điệu của văn bản. Sau đó khối xử lý tín hiệu phải chuyển đổi các thể hiện ngôn ngữ dạng trừu tượng này thành dạng sóng tiếng nói theo kĩ thuật tổng hợp tiếng nói.

**Các quá trình chuyển đổi**

Theo sơ đồ tổng quát, để thực hiện việc chuyển đổi văn bản thành tiếng nói phải thực hiện bốn bước sau đây:

**1.** Chuẩn hoá văn bản (Text Normalization): Xác định biên của câu và từ, chuyển chữ số trong văn bản thành dạng chữ, chuyển đổi các chữ viết tắt và ghi nhận một số dấu quan trọng cho phần xử lý ngữ điệu.

**2.** Chuyển đổi từ chữ sang âm vị (Grapheme to Phoneme) dựa trên từ điển phát âm hay theo quy luật ngôn ngữ. Quá trình chuyển đổi “phiên âm” này rất quản trọng, chiếm tới hơn 30% khối lượng cộng việc cho phần lớn các ngôn ngữ và chữ viết không phải là loại chữ viết ghi âm. Chữ viết của tiếng Việt là loại chữ viết ghi âm nên việc phát triển một hệ thống TTS tiếng Việt giảm được rất nhiều.

**3.** Dự báo ngữ điệu từ văn bản: độ dài, cao độ tần số cơ bản của các ngữ đoạn tổng hợp phải phù hợp với các quy luật về ngữ điệu của

một ngôn ngữ. Đó là việc xác định vị trí trọng âm của từ được phát âm. Lên xuống giọng ở các vị trí khác nhau trong câu, như đầu câu, cuối câu vv… và xác định các biến thể khác nhau của các âm phụ thuộc vào ngữ cảnh khi được phát âm trong một ngữ lưu liên tục. Bước này bao gồm cả việc xác định điểm dừng nghỉ lấy hơi khi phát âm, nói hoặc đọc một văn bản. Từ đó tạo ra các thông tin để điều khiển ngữ điệu thích hợp cho bộ tổng hợp tiếng nói.

**4.** Tổng hợp tiếng nói là giai đoạn cuối cùng thực hiện việc giải mã các thông tin từ chuỗi mô tả ngữ âm nhận được từ khối xử lý văn bản, tạo ra chuỗi các tham số cho bộ tổng hợp (phụ thuộc vào phương pháp tổng hợp). Bộ tổng hợp tạo ra tín hiệu tiếng nói dạng sóng.

Tổng hợp tiếng nói có thể đạt được cơ bản theo hai phương pháp:

* + Phương pháp thứ nhất được thực hiện dưới dạng một loạt các quy tắc mô tả một cách chính thức các âm vị, ảnh hưởng lẫn nhau giữa các âm vị khi phát ra một âm.
  + Phương pháp thứ hai là lưu giữ những mẫu âm vị, biến đổi âm vị và đồng phát âm vào cơ sở dữ liệu tiếng nói và sự dụng chúng như chúng vốn có là các đơn vị âm học cơ bản (ở vị trí của âm vị).

Hai hệ TTS đã xuất hiện từ hai phương pháp này. Hai phương pháp này nhanh chóng trở thành hai triết lý tổng hợp với sự khác biệt thể hiện trong nội dung và mục tiêu của chúng. Đó là: tổng hợp bằng quy luật và tổng hợp xích chuỗi.

Ta có thể chia phuơng pháp tổng hợp tiếng nói trong các hệ thống TTS thành ba loại là tổng hợp cấu âm (articulatory synthesis), tổng hợp formant theo quy luật (rule based formant synthesis) và tổng hợp xích chuỗi (concatenation synthesis).

1. **Các phương pháp tổng hợp tiếng nói** 
   1. **Tổng hợp theo cấu âm**

Tổng hợp theo cấu âm sử dụng mô hình phỏng sinh học. Trong mô hình này, tiếng nói được đặc trưng không chỉ bằng tín hiệu âm thanh mà cả bằng vị trí của các cơ quan cấu âm, sự tham gia của khoang mũi, áp suất không khí trong phổi, lực căng của dây thanh. Nó bao gồm hoạt động tương tác phi tuyến giữa nguồn kích (chấn động dây thanh hoặc hỗn loạn không khí) và trở kháng đầu vào thay đổi theo thời gian của ống thanh. Phương pháp này còn được gọi là tổng hợp theo mô hình hệ thống.

Để xây dựng đặc trưng âm thanh của cơ quan phát âm, hệ thống phải tham số hoá được hình khối ống thanh và ống mũi, phải lựa chọn được một mô hình để mô tả sự lan truyền của sóng trong ống thanh và mô hình hoá nguồn âm (chấn động dây thanh, dòng không khí hỗn loạn) và sự tương tác của nó với ống thanh. Phương pháp này hiện nay chỉ hình thành trên lý thuyết, thực tế ngành khoa học phỏng sinh chưa có đủ thông tin đầu đủ để mô phỏng theo cấu âm.

**Mô hình ống thanh**

Cho các âm không phải là âm mũi ta chỉ cần xem xét ống thanh có thiết diện biến thiên từ thanh môn cho tới môi. Ống thanh thay đổi liên tục, hình dáng của chúng phải được xác định trong khoảng thời gian vài mini giây một lần. Hình không gian của ống thanh được mô tả theo vị trí của các cơ quan phát âm, lưỡi, môi, hàm…

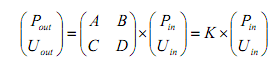
Mô hình được ứng dụng nhiều nhất là mô hình của Mermelstein với các tham số mô tả vị trí cơ quan cấu âm. Giá trị của các tham số được liệt kê chi tiết ở bảng 7.1. Từ mô tả cấu âm này, có thể tính được các số liệu cho hàm mặt cắt ngang ống thanh A(x) là hàm số phụ thuộc khoảng cách dọc theo ống thanh từ thanh môn cho tới môi

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên | Ý nghĩa | Ngưỡng dưới | Ngưỡng trên |
| 1  2  3  4  5  6  7  8  9  10  11 | Rtp  Tejp  Xcp  Stp  Xclp  Xhp  Ycp  Telp  Cylp  Yhp  Vel | Bán kính lưới (cm)  Góc mở hàm (độ)  Trung tâm lưỡi trục x (cm)  Chiều dài lưỡi (cm)  Nhô ra của môi (cm)  Vị trí xương trục x (cm)  Trung tâm lưỡi trục y (cm)  Độ nâng lưỡi (deg)  Chiều cao môi y (cm)  Vị trí xương trục y (cm)  Độ mở vòm miệng (cm2) |  |  |

***Bảng 7-1- Các tham số của mô hình cấu âm Mermelstein***

**Lan truyền sóng trong ống thanh**

Trong nhiều cách tính đặc trưng truyền sóng trong ống thanh thì phương pháp ma trận chuỗi được áp dụng nhiều nhất. Ma trận chuỗi (còn gọi là ma trận ABCD) thể hiện ống thanh ở miền tần số. Với bất kỳ phần nào của ống thanh hay của ống mũi ta có:



Ma trận chuỗi K có thể thể hiện bất kỳ phần nào của bất kỳ độ dài hay mặt phẳng cắt ngang nào của ống thanh. Tính các phần tử của K rất đơn giản nên ống có mặt cắt không đổi.

**Mô hình nguồn kích thanh**

Nguồn kích thanh cho tổng hợp cấu âm sử dụng mô hình hai khối lượng của Ishizaka và Flanagan.

**Nguyên lý của tổng hợp cấu âm**

Thuật toán tổng hợp theo mô hình cấu âm gồm các bước như nhau. Từ đặc tả về các tham số cấu âm, ta tính được ma trận chuỗi từ thanh môn đến môi. Sau đó tính hàm truyền và trở kháng đầu vào từ các phần tử A,B,C,D của ma trận. Chuyển đổi Fourier ngược hàm truyền H và tỉ số trở kháng ta nhận được các giá trị tương ứng ở miền thời gian. Sau đó tính áp suất Δp = p 5-p1. Giá trị áp suất này tác động vào bộ cộng hưởng điều khiển chuyển động dây thanh của mô hình hai khối lượng tạo ra âm thanh tiếng nói.

Như vậy, ngoài việc mô hình hoá các cơ quan phát âm, tổng hợp theo cấu âm cần phải xây dựng hai loại quy luật ngôn ngữ và luật vật lý. Luật vật lý mô tả như ở trên xác định ánh xạ cấu hình ống thanh cụ thể cho tín hiệu âm thanh, xác định mối quan hệ giữa hoạt động cấu âm và tín hiệu âm thanh tiếng nói.

Còn luật ngôn ngữ xác định mối quan hệ giữa mô tả ngữ âm và các hoạt động cấu âm. Thông thường, luật ngôn ngữ được thực hiện cho từng âm vị. Khi cấu âm cho một đơn vị, luật này xác định thời điểm các cơ quan chuyển động, tốc độ chuyển động và mối quan hệ giữa cơ quan này và cơ quan khác. Vì phát ra một âm vị không nhất thiết phải có sự chuyển động của tất cả các cơ quan nên trong cùng một thời điểm cơ quan phát âm có thể được xác định cho âm vị khác. Theo cách này thì tổng hợp theo cấu âm có thể tạo ra các giải pháp hoàn hảo cho các trường hợp ngữ âm khó như cụm phụ âm, nguyên âm ba. Mô hình cấu âm và hai luật ngôn ngữ, vật lý này quyết định chất lượng tiếng nói tổng hợp. Hiện nay, chúng ta chưa có kiến thức đầy đủ về từng chi tiết của các loại luật, mô hình cấu âm cũng còn đơn giản vì vậy chất lượng tiếng nói tổng hợp theo cấu âm còn rất thấp. Phương pháp tổng hợp cấu âm hiện nay chỉ dừng ở trong phòng thí nghiệm hay sản phẩm mẫu mà chưa được ứng dụng thực tế. Các bộ tổng hợp theo cấu âm là công cụ lý tưởng để nghiên cứu về cấu âm của tiếng nói và hứa hẹn về lâu dài là giải pháp hoàn thiện nhất để tổng hợp ra âm thanh tiếng nói giống như tiếng nói tự nhiên của con người.

* 1. **Tổng hợp formant theo quy luật**

Quy trình để xây dựng bộ tổng hợp formant theo luật gồm hai quá trình riêng biệt là quá trình phân tích âm thanh tiếng nói để tìm ra các quy luật và tổng hợp lại tiếng nói dựa vào các quy luật này.

**Phân tích tìm quy luật**

Quá trình phân tích được tiến hành trên cơ sở dữ liệu tiếng nói tự nhiên chủ yếu bao gồm các âm tiết có dạng C-V (phụ âm – nguyên âm) hay C-V-C (phụ âm – nguyên âm – phụ âm) của nhiều giọng nói. Cơ sở dữ liệu này tốt nhất phải bao phủ được tất cả các hình thái ngữ âm của một ngôn ngữ.

Thuật toán tiên đoán tuyến tính LPC, được sử dụng trong quá trình phân tích phổ để xác định mẫu của tần số formant, băng thông formant, tần số âm cơ bản F0 của tín hiệu tiếng nói. Tần số từ F1 đến F5 và băng thông W1 đến W5 ở phần ổn định của các âm tố được lưu giữ dưới dạng bảng. Đó là các giá trị “đích” để tạo ra một âm vị.

Trong tiếng nói, phần mang thông tin được quan tâm nghiên cứu nhiều hơn là các phần chuyển tiếp giữa các âm vị từ nguyên âm sang phụ âm và ngược lại (C-V và V-C). Tần số formant chuyển tiếp từ âm vị nọ sang âm vị kia được mô hình hoá và tạo thành các quy luật tổng hợp cho kĩ thuật tổng hợp formant. Như vậy về cơ bản các quá trình phân tích tìm luật để mô tả sự thay đổi liên tục phổ của một ngữ lưu bằng các formant và sự chuyển tiếp của các formant từ âm vị này sang âm vị khác. Hệ thống quy luật bao gồm:

* Luật để xác định vị trí đích các formant của mỗi âm vị trong một ngữ lưu cũng như thời gian tồn tại của các vị trí đó.
* Luật nhằm làm trơn các formant đích.
* Luật để chuyển đổi các hàm thời giam thành dạng sóng thanh.

**Tổng hợp formant**

Hệ thống tổng hợp formant sẽ căn cứ vào các tham số điều khiển từ bảng tra cứu và cùng với thuật toán được xây dựng cho sự chuyển tiếp là các tín hiệu điều khiển bộ tổng hợp để tạo các âm đoạn tính.

Thiết kế của bộ tổng hợp formant được dựa trên mô hình tương tự đầu cuối tạo tiếng nói do Frant đề xuất.

Tổng hợp formant được phân loại theo cấu hình mắc các bộ cộng hưởng song song hay nối tiếp. Hàm truyền của bộ tổng hợp khi mắc nối tiếp bộ cộng hưởng tương tự như hàm truyền ống thanh vì vậy bộ tổng hợp nối tiếp mô hình hoá rất tốt cho các âm hữu thanh. Còn bộ tổng hợp với các bộ cộng hưởng mắc song song sẽ tạo ra cho các âm tắc, âm xát và âm mũi một chất lượng âm thanh tốt hơn.

* 1. **Tổng hợp xích chuỗi**

Phương pháp xích chuỗi nhằm vượt qua sự không thoả đáng của các hệ thống sử dụng quy luật bằng cách loại bớt sự cần thiết của một số quy luật thậm chí tất cả các quy luật. Để thực hiện điều này, cách hiệu quả nhất là lưu trữ trong kho dữ liệu không phải hệ thống quy luật mô tả các đơn vị âm hay quy luật của các tham số để tạo ra đơn vị âm mà trực tiếp từng âm đoạn của âm thanh tiếng nói.

Trong các phương pháp nêu trên, tổng hợp theo cấu âm, tổng hợp formant theo quy luật và tổng hợp xích chuỗi thì tổng hợp xích chuỗi mang nhiều tính công nghệ, được quyết định bởi sự phát triển của máy tính. So sánh về chất lượng trong ba phương pháp, thì tổng hợp xích chuỗi hiện nay đã thành công nhất trong việc tạo ra tiếng nói không chỉ nghe hiểu rõ mà đạt được độ tự nhiên gần với tiếng nói của con người hơn cả. Phần lớn các hệ thống TTS chất lượng cao cho các ngôn ngữ có trên thị trường hiện nay đều sử dụng phương pháp tổng hợp xích chuỗi này.

Khi xây dựng một hệ thống tổng hợp xích chuỗi, tiếng nói chứa tập hợp các đơn vị âm thanh cơ bản của một ngôn ngữ, thông thường tương ứng với một chuỗi ngắn các âm vị có lựa chọn được đọc, ghi âm và lưu giữ trong CSDL. Các đơn vị âm được lựa chọn, phân tích các tham số đặc trưng và sau đó có thể được mã hoá bằng một phương pháp mã hoá tiếng nói nào đó tạo thành các mẫu cho một đơn vị âm thanh. Các mẫu hoặc các tham số phân tích của cúng được lưu giữ lại trong kho dữ liệu đơn vị tổng hợp.

Khi tổng hợp một phát ngôn mới, cho trước mô tả ngữ âm của phát ngôn đó, hệ thống sử dụng các quy luật để định vị đơn vị thích hợp, truy xuất chúng ra khỏi kho dữ liệu và xích chuỗi chúng lại với nhau.

Thông thường, một hệ thống tổng hợp xích chuỗi có ít nhất là một hoặc hai mẫu cho mỗi một đơn vị cơ bản. Do các tham số về cao độ tần số cơ bản, độ dài và cường độ của các đơn vị này rất khác nhau (chúng được trích ra từ tín hiệu tiếng nói trong ngữ cảnh khác nhau) cho nên, khi tổng hợp, hệ thống theo phương pháp tổng hợp xích chuỗi phải thực hiện hai công việc chính.

* Một là phải thay đổi các tham số siêu đoạn tính của các đơn vị để tiếng nói tổng hợp thể hiện được ngữ điệu thích hợp như mong muốn.
* Hai là phải thực hiện việc lựa chọn hay thay đổi các đặc tính của âm đoạn đơn vị tại các biên ghép nối của chúng sao cho việc ghép nối là mịn nhất để đảm bảo chất lượng âm thanh tổng hợp tạo ra gân với tiếng nói tự nhiên.

Như vậy chất lượng của một hệ thống TTS sử dụng phương pháp tổng hợp xích chuỗi được quyết định bởi các yêu tố sau:

* Sự lựa chọn tập âm đơn vị cơ bản. Trong đó ngữ âm của ngôn ngữ tổng hợp là yếu tố chính quyết định loại đơn vị cơ bản.
* Chất lượng âm thanh và mức độ bao phủ các tổ hợp âm đoạn cần thiết cho một ngôn ngữ của CSDL âm.
* Khả năng cung cấp các tham số về ngữ điệu như đường nét tần số cơ bản cường độ âm thanh và độ dài các âm cho bộ tổng hợp. Các tham số này được tiên đoán trong phần xử lý ngôn ngữ của bộ tổng hợp từ văn bản.
* Khả năng thay đổi tần số cơ bản F0 và độ dài các đơn vị âm cơ bản phù hợp với ngữ điệu được miêu tả ở đầu vào của bộ tổng hợp mà không ảnh hưởng tới chất lượng cảm thụ của âm đó.
* Thuật toán xích chuỗi và mô hình tín hiệu cho phép dễ dàng làm trơn hai điểm ghép nối của các âm đoạn được xích chuỗi.

Chính vì hai yếu tố sau nên một số hệ thống tổng hợp xích chuỗi thực hiện mã hoá tham số tiếng nói. Với cách thể hiện bằng các tham số, đặc trưng phổ và ngữ điệu của tiếng nói được thay đổi dễ dàng hơn mà không thay đổi đặc tính nhận dạng âm đơn vị cơ bản.

**Kĩ thuật cộng chồng đồng bộ cao độ tần số cơ bản – PSOLA**

Phương pháp thay đổi cao độ tần số cơ bản và độ dài tiếng nói ở miền thời gian có lợi thế hơn hẳn về chất lượng âm thanh tổng hợp và khả năng thực hiện thay đổi thời gian thực. Như chúng ta đã biết, tổng hợp xích chuỗi bao gồm hai quá trình là quá trình giải mã và xích chuỗi**.**

Quá trình giải mã: tái tạo lại dạng sóng của tín hiệu tiếng nói từ dạng tham số trong kho dữ liệu.

Xích chuỗi: một chuỗi các đơn vị âm được xích chuỗi sau khi đã thay đổi các đặc trưng siêu đoạn tính của chúng phù hợp với miêu tả ngữ âm.

Xích chuỗi dùng mã tiên đoán tuyến tính LPC kết hợp hai quá trình nay làm một. Thay đổi thanh điệu được thực hiện bằng cách thay đổi tần số cơ bản F0, nguồn kích thanh trước khi đưa vào bộ lọc tổng hợp. độ dài được thay đổi bằng cách cập nhật các tham số cho bộ tổng hợp theo tần suất khác với tần suất phân tích.

Kĩ thuật xử lý tín hiệu miền thời gian để thay đổi cao độ tần số cơ bản và độ dài tiếng nói trực tiếp trên dạng sóng tiếng nói là cộng chồng đồng bộ chu kỳ cao độ tần số cơ bản (PSOLA).

**Tổng hợp theo mô hình số liệu corpus-based**

***Nguyên lý tổng hợp mô hình số liệu***

Như trình bày ở trên, nếu xây dựng hệ thống tổng hợp tiếng nói với số lượng đơn vị âm tối thiểu và đủ ta phải dựa vào các công nghệ xử lý tín hiệu số để thay đổi các tham số ngữ điệu của câu tổng hợp theo quy luật thay đổi ngữ âm để tăng độ tự nhiên của tiếng nói. Kĩ thuật điều khiển trực tiếp dạng sóng tiếng nói trên cơ sở cộng chồng đồng bộ cao độ tần số cơ bản PSOLA.

PSOLA cho phép thay đổi dễ dàng cao độ tần số cơ bản và đô dài của tiếng nói, không làm ảnh hưởng đáng kể đến chất lượng âm thanh so với kĩ thuật tham số LPC.

Tuy nhiên, với bất kỳ một kĩ thuật xử lý tín hiệu số nào, chất lượng tiếng nói cũng bị giảm đi. Như vậy, để tổng hợp ra tiếng nói có độ tự nhiên gần với tiếng nói con người, chúng ta cần phải có thật nhiều các đơn vị âm nguyên bản phù hợp cho mọi ngữ cảnh để giảm thiểu sự tác động để thay đổi tín hiệu làm giảm chất lượng âm thanh tổng hợp. Đây chính là điểm xuất phát của ý tưởng xây dựng phương pháp số liệu corpus-based cho tổng hợp tiếng nói.

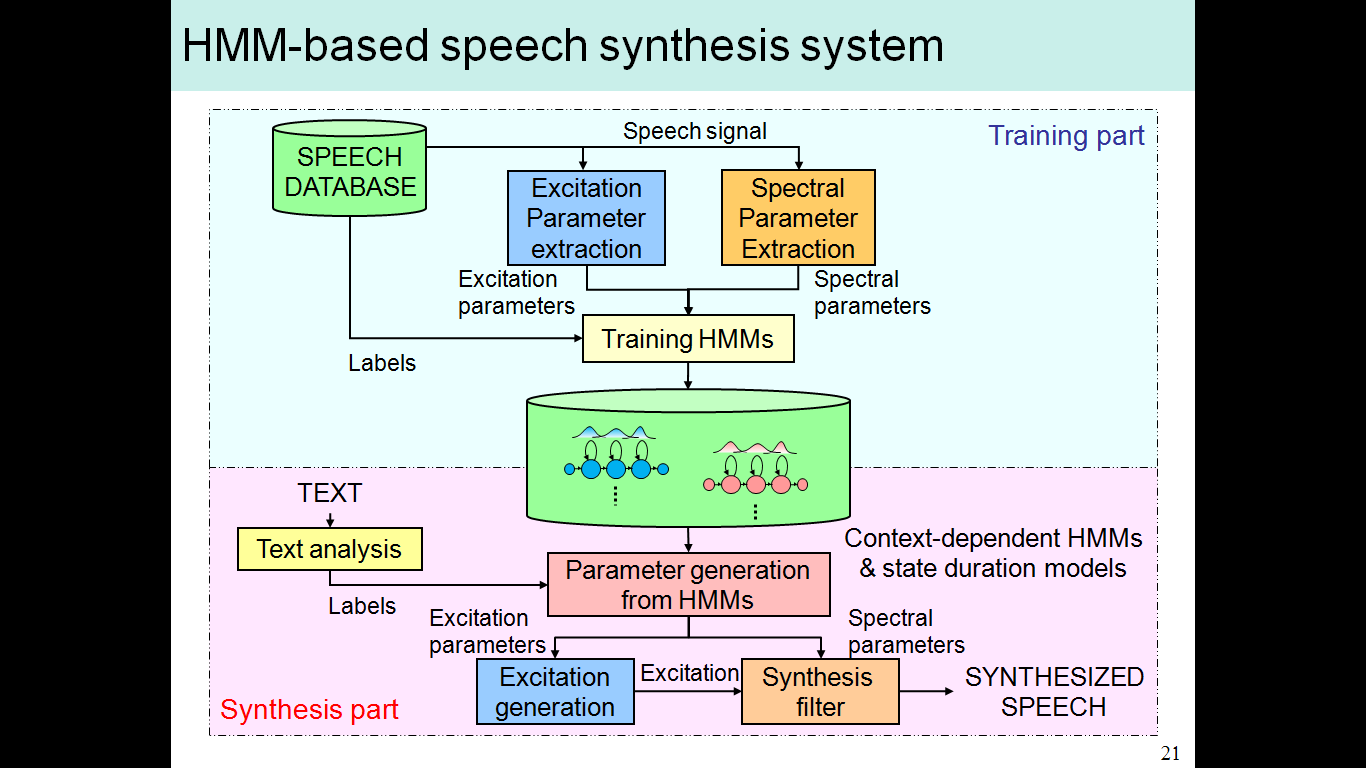
Khi dựa trên cơ sở dữ liệu lớn có nhiều biến thể âm thì chúng ta không tạo ra mà chỉ lựa chọn và sắp xếp lại các đơn vị âm một cách thích hợp. Các âm này được thiết lập chỉ số theo đặc tả ngữ âm và ngữ điệu được lưu trữ trong kho dữ liệu trước đó. Khi lựa chọn phải giảm thiểu sự gián đoạn về phổ và tăng tối đa tiệm cận về các tham số ngữ điệu giữa đơn vị âm trong cơ sở dữ liệu tới ngữ đoạn đích cần tổng hợp. Ngữ đoạn đích được mô tả với các thuộc tính về ngữ âm và ngữ điệu của nó được xử lý tạo lập ở khối xử lý ngôn ngữ.

Với phương thức tiếp cận này, ta thấy rằng, bất kỳ một khối dữ liệu tiếng nói nào cũng có thể được sử dụng làm nguồn cho đơn vị tổng hợp, chỉ cần có điều kiện là đã có bản ghi âm và mô ta chính tả của tiếng nói. Như vậy, ta có thể tổng hợp được tiếng nói có cùng giọng và thậm chí giống cả cách của người nói tạo ra nguồn âm đó. Chất lượng tiếng nói tổng hợp ở phương pháp này phụ thuộc vào sự cân bằng ngữ âm của số liệu và vào độ sẵn sàng của các đơn vị đại diện trong khối dữ liệu. Nguồn âm càng giàu, độ tự nhiên càng cao; nguồn âm nghèo thì tiếng nói bị gián đoạn và chất lượng không ổn định. Số lượng lớn đơn vị âm đặt ra các yêu cầu mới về quản trị các âm đơn vị này trong cơ sở dữ liệu như thế nào, qua trình lựa chọn được tiến hành ra sao để đảm bảo tốc độ và khả năng ứng dụng thời gian thực. Bộ tổng hợp dựa trên cơ sở số liệu được xây dựng theo ba giai đoạn chính là: Tự động tạo lập chỉ số cho tiếng nói tự nhiên bao gồm việc định vị trí của các đơn vị âm tổ hợp và xác định ngữ cảnh của chúng ví dụ như vị trí trong ngữ đoạn, loại âm tiết, thanh điệu của các âm tiết ngay trước và sau vv… Như vậy, trong giai đoạn này, các quy luật thay đổi của các âm phụ thuộc vào ngữ cảnh trong quá trình tạo tiếng nói được phát hiện và tích hợp vào trong chỉ số tham số của các đơn vị âm.

Xây dựng cơ sở dữ liệu đơn vị âm - Việc tạo lập CSDL đơn vị tổng hợp được thực hiện một lần cho một khối dữ liệu mới. CSDL chứa vector đặc tính về phổ và ngữ điệu cũng như các trọng số của đơn vị âm theo chỉ số đã được thiết lập ở giai đoạn trước. CSDL tổng hợp được được tái sử dụng để thực hiện vô số các tổng hợp khác nhau.

Bộ tổng hợp thực hiện chức năng lựa chọn đơn vị âm trên cơ sở so sánh đặc tả đích được chuẩn bị ở phần xử lý văn bản với các đặc tính đơn vị âm chứa trong CSDL, truy xuất khỏi CSDL, xích chuỗi chúng lại với nhau.

**Tổng hợp theo mô hình markov ẩn**



Xuất hiện từ đầu những năm 90 nhưng tổng hợp tiếng nói theo mô hình markov ẩn đã dần trở thành phổ biến bởi những ưu điểm của nó.

Mô hình Markov ẩn là một mô hình thống kê có khả năng ghi nhớ các biến đổi theo thời gian của các đại lượng không ổn định, thường được áp dụng trong lĩnh vực nhận dạng tiếng nói. Mô hình Markov ẩn (HMM) được xây dựng trên cơ sở lý thuyết của chuỗi Markov và các tiến trình Markov.

Tổng quan về mô hình markov ẩn đối với tổng hợp tiếng nói có thể nhìn hình trên,một tập cơ sở dữ liệu về giọng nói được training qua mô hình HMM để ra được một tập các tham số.Khi có được tập tham số này rồi,điều còn lại là chúng ta phải xử lý đoạn văn bản đưa vào để tạo ra các tập tham số tương ứng với tập đã training và đi ngược lại để tổng hợp tiếng nói

Những ưu điểm của mô hình markov ẩn là:

-Chất lượng giọng nói tốt,ổn định

-Dễ dàng thay đổi giọng nói

1. **So sánh các phương pháp tổng hợp tiếng nói.**

Sau khi giới thiệu những đặc điểm cơ bản nhất của các phương pháp tổng hợp tiếng nói ta có thể rút ra một số nhận xét về các phương pháp này. Các nhận xét này nhằm mục đính đưa ra đánh giá về ba phương pháp dựa trên chất lượng tiếng nói tổng hợp, chi phí tính toán và kích thước dữ liệu.

* *Về chất lượng của tiếng nói tổng hợp*: Trong các phương pháp nói trên thì phương pháp mô phỏng bộ máy phát âm về nguyên tắc sẽ cho chất lượng tốt nhất. Để đạt được điều này thì vấn đề quan trọng là làm sao mô phỏng chính xác bộ máy phát âm của con người. Công việc này hoàn toàn không đơn giản, mặc dù đã có sự trợ giúp của mày tính nhưng do cấu trúc phức tạp của bộ máy phát âm nên chi phí tính toán sẽ rất lớn. Trong hai phương pháp còn lại thì thực tế cho thấy phương pháp ghép nối thường cho chất lượng tốt hơn.
* *Về hiệu quả tính toán*: Rõ ràng là phương pháp mô phỏng bộ máy phát âm đòi hỏi chi phí tính toán lớn nhất vì phải mô phỏng một cách chính xác nhất bộ máy phát âm phức tạp của con người. Hai phương pháp còn lại có chi phí tính toán thấp hơn do đặc điểm các thuật toán được sử dụng.
* *Về kích thước dữ liệu*: Phương pháp ghép nối có kích thước dữ liệu lớn nhất do số lượng từ vựng là rất lớn. Các phương pháp còn lại do không phải lưu trữ các mẫu nên có kích thước dữ liệu nhỏ hơn.

Qua những nhận xét trên thì khó khăn lớn nhất của phương pháp mô phỏng bộ máy phát âm là làm sao để mô phỏng chính xác bộ máy phát âm của con người. Với phương pháp tổng hợp bằng formant thì vấn đề cần giải quyết là chất lượng tiếng nói tổng hợp. Còn với phương pháp tổng hợp ghép nối thì có ưu điểm là chi phí tính toán không cao và chất lượng khá tốt, khó khăn lớn nhất là giảm kích thước dữ liệu. Khó khăn này, như đã trình bày, có thể khắc phục bằng cách tổng hợp tiếng nói từ những đơn vị nhỏ hơn từ như âm vị, diphone.Còn phương pháp tổng hợp tiếng việt sử dụng HMM thì giải quyết được các yếu điểm của các phương pháp trên : giảm kích thước dữ liệu và cải thiện chất lượng tiếng nói

Với mục đích nghiên cứu việc tổng hợp tiếng Việt và dựa trên những đặc điểm của các phương pháp tổng hợp, luận văn này sẽ sử dụng phương pháp tổng hợp sử dụng mô hình markov ẩn. Trong số những phương pháp dùng để tổng hợp bằng ghép nối thì HMM là phương pháp được sử dụng rộng rãi nhất với ưu điểm là chi phí tính toán thấp và giữ nguyên được nhiều thông tin trong tiếng nói do thao tác trực tiếp với tín hiệu trên miền thời gian,đồng thời việc chuyển giọng nói tổng hợp được thực hiện nhanh chóng và dễ dàng.