R

BỘ KHOA HỌC VÀ CÔNG NGHỆ
VIỆN ỨNG DỤNG CÔNG NGHỆ
Trung tâm Công nghệ Vi điện tử và Tin học

---------- 🕮 ----------

BÁO CÁO KHOA HỌC

Đề tài:

Nghiên cứu xây dựng phần mềm "*Tự động đọc văn bản chữ Việt*" bằng phương pháp tổng hợp formant

Chủ nhiệm đề tài: **Lê Hồng Minh**

Hà nội **2004**

Đề tài:

Nghiên cứu xây dựng phần mềm "*Tự động đọc văn bản chữ Việt*" bằng phương pháp tổng hợp formant

- **Cấp quản lý:** Cấp Bộ
- **Thời gian thực hiện:** 24 tháng (1/2002-12/2003)
- **Cơ quan thực hiện:** Trung tâm Công nghệ Vi điện tử và Tin học
- **Cơ quan chủ trì:** Viện Ứng dụng Công nghệ
- **Cơ quan chủ quản:** Bộ Khoa học và Công nghệ
- **Chủ nhiệm đề tài:** ThS. Lê Hồng Minh,
  Trung tâm Công nghệ Vi điện tử và Tin học

Những người tham gia thực hiện:

1. ThS. Trần Cảnh, Trường Đại học Xây dựng
2. ThS. Ngô Minh Dũng, Viện Khoa học Hình sự
3. ThS. Phạm Minh Hoàn, Trung tâm Công nghệ Vi điện tử và Tin học
4. TS. Lê Khánh Hùng, Trung tâm Công nghệ Vi điện tử và Tin học
5. CN. Nguyễn Vĩnh Sơn, Trung tâm Công nghệ Vi điện tử và Tin học
6. CN. Nguyễn Phương Thảo, Trung tâm Công nghệ Vi điện tử và Tin học
7. ThS. Hoàng Minh Thức, Trường Đại học Bách khoa Hà nội
8. CN. Phạm Xuân Tích, Trung tâm Công nghệ Vi điện tử và Tin học
9. CN. Mai Kiều Trang, Trung tâm Công nghệ Vi điện tử và Tin học

MỤC LỤC

DANH SÁCH HÌNH VẼ VÀ BẢNG BIỂU

BẢNG CÁC TỪ VIẾT TẮT VÀ THUẬT NGỮ

American Standard Code for Infomation Interchage (ASCII)	: Bảng mã ASCII
Amplitude	: Biên độ
Articulatory	: Cấu âm
Artificial Neural Networks (ANN)	: Mạng nơ ron nhân tạo
Band Width (BW)	: Dải thông
Coarticulatory	: Liên cấu âm
Concatenation	: Ghép (xích) chuỗi
Diphone	: Âm vị ghép (hai nửa khác nhau)
F0	: Tần số cơ bản
F1, F2, F3, ...	: Các tần số Formant (cộng hưởng)
Formant Frequency	: Tần số cộng hưởng
Frequency Domain PSOLA (FD PSOLA)	: PSOLA miền tần số
Hidden Markov Models (HMM)	: Mô hình Markov ẩn
International Phonetic Alphabet (IPA)	: Bảng chữ cái ngữ âm quốc tế
Interactive Voice Response (IVR)	: Tương tác bằng giọng nói
Linear Predictive PSOLA (LP-PSOLA)	: PSOLA dự đoán tuyến tính
Linear Prediction Code (LPC)	: Mã hoá dự đoán tuyến tính
Phoneme	: Âm vị
Pitch	: Chu kỳ tần số cơ bản
Pitch Synchronous OverLap Add (PSOLA)	: Cộng chồng và đồng bộ Pitch
Pulse Code Modulation (PCM)	: Điều biến mã hoá xung
Pre-Recorded Prompts (PRP)	: Kịch bản được ghi trước
Prosody	: Ngôn điệu
Radiation	: Tán xạ
Short-time	: Thời gian ngắn
Source-Filter-Model	: Mô hình nguồn âm - bộ lọc
Speech Copus	: Kho dữ liệu tiếng nói
Speech Application Programming Interface (SAPI)	: Giao diện lập trình tiếng nói
Spectrum	: Phổ
Spectrogram	: Ảnh thanh phổ
Spoken Text Markup Language (STML)	: Ngôn ngữ đánh dấu văn bản tiếng nói
Speech Synthesis Markup Language (SSML)	: Ngôn ngữ đánh dấu tổng hợp tiếng nói
Speech Synthesis	: Tổng hợp tiếng nói
Syllable	: Âm tiết
Text To Speech (TTS)	: Văn bản thành tiếng nói
Time Domain PSOLA (TD-PSOLA)	: PSOLA miền thời gian
Tone Language	: Ngôn ngữ thanh điệu
Vocal tract	: Tuyến âm

MỞ ĐẦU

Giao tiếp ***Người-Máy*** bằng tiếng nói là mong muốn và mục tiêu phấn đấu từ rất lâu của con người. Một nửa của quá trình giao tiếp là việc Máy tính có thể truyền thông tin cho con người bằng tiếng nói. Bản chất của sự việc này là phải xây dựng được một engine có thể tự động chuyển thành tiếng nói các đoạn văn bản hay một nội dung nào đó (TTS). TTS của các ngôn ngữ chính và của các nước phát triển như tiếng Anh, Pháp... đã có các bước tiến rất xa, có rất nhiều ứng dụng, thậm chí nhiều sản phẩm đã được cứng hoá [SpeakJets]. Đối với tiếng Việt, đây là công việc đặc thù của Việt Nam nên không thể nào chỉ trông đợi từ người ngoài, mà phải do chính người Việt phải chủ động nghiên cứu và phát triển. Những năm gần đây cùng với đà phát triển chung, TTS cũng đã được quan tâm nghiên cứu và có một số kết quả.

Phát triển một hệ TTS tiếng Việt bao gồm tuần tự các bước: dạy máy biết "*nói*" tiếng nói con người; dạy máy biết "*nói tiếng Việt*"; dạy máy biết "*đọc tiếng Việt*" và cuối cùng dạy máy "*đọc có ngữ điệu tiếng Việt*". Các phương pháp khác nhau có các mối quan tâm riêng, nếu sử dụng tiếng nói tự nhiên ghi âm trước thì việc máy "*nói*" và "*nói tiếng Việt*" là vấn đề đơn giản vì chỉ việc phát lại (replay), tuy nhiên lúc này phải quan tâm đến liệu ta có thể thực hiện được các việc "*đọc và đọc có ngữ điệu các văn bản tiếng Việt*" bất kỳ hay không? Hệ TTS không sử dụng tiếng nói tự nhiên ghi âm trước sẽ phải thực hiện tất cả các khâu kể trên, tuy nhiên, ta có thể kế thừa được kết quả nghiên cứu của các ngôn ngữ khác ở bước đầu tiên là dạy máy "*nói*" tiếng người, vì nói ngôn ngữ gì thì cũng là tiếng nói! Điều thuận lợi của công nghệ này là khả năng điều khiển mềm dẻo nên các bước tiếp sau để nâng cao chất lượng sẽ thuận lợi hơn.

Đề tài "*Nghiên cứu xây dựng phần mềm đọc văn bản chữ Việt bằng phương pháp tổng hợp formant*" tiến hành theo giải pháp không sử dụng tiếng nói tự nhiên ghi trước mà bằng tiếng nói tổng hợp được tạo ra dựa theo mô hình và nguyên lý tạo tiếng nói con người, gọi là phương pháp tổng hợp formant. Đề tài đã tiến hành các nội dung nghiên cứu và triển khai liên quan đến các lĩnh vực như ngôn ngữ, ngữ âm

học, xử lý tín hiệu, khoa học máy tính để tạo được một engine phần mềm (đặt tên là ***Vnspeech***). Vnspeech đã tổng hợp được 1 giọng nam từ các thông tin ngữ âm, có thể đọc được văn bản tiếng Việt bất kỳ và cho phép điều khiển mềm dẻo các tham số đặc trưng của tiếng nói, chất lượng tiếng nói tổng hợp tương đối dễ nghe và có thể sử dụng trong nhiều lớp ứng dụng. Các nội dung ***chưa*** triển khai nghiên cứu trong phạm vi đề tài là *đọc văn bản có ngữ điệu* và xây dựng dữ liệu về thông tin ngữ âm của *nhiều giọng*. Để kết quả của đề tài có thể trở thành một sản phẩm dùng chung như một công nghệ cơ bản, có ứng dụng rộng rãi hơn nữa trong nhiều lĩnh vực cần phải tiếp tục nghiên cứu để nâng cao chất lượng tín hiệu, phân tích và tổng hợp ngữ điệu từ văn bản, tăng thêm số lượng giọng nói xây dựng sẵn cũng như các khả năng điều khiển các thông số đặc trưng khác.

Báo cáo này trình bày các kết quả thu được của quá trình nghiên cứu, triển khai xây dựng phần mềm TTS cho tiếng Việt dựa trên tiếng nói được tổng hợp bằng phương pháp tổng hợp Formant. Báo cáo được bố cục thành 6 chương: chương 1 trình bày về thiết kết của hệ TTS tiếng Việt – Vnspeech, gồm so sánh để lựa chọn phương pháp tổng hợp tiếng nói formant cho triển khai của đề tài; chương 2 trình bày về bộ tổng hợp tiếng nói formant của Klatt, áp dụng để tổng hợp tiếng Việt, đây là phần xử lý tín hiệu số, tạo ra tín hiệu tiếng nói, làm cho máy biết "*nói*" tiếng Việt; các kết quả về nghiên cứu ngữ âm tiếng Việt cho mục đích tổng hợp tiếng nói được trình bày trong chương 3; nội dung của chương 4 là các công việc về chuyển văn bản tiếng Việt thành các tham số điều khiển bộ tổng hợp formant, là đầu vào của bộ tổng hợp, đây là bước "*dạy máy đọc văn bản tiếng Việt*"; chương 5 trình bày một số tiêu chuẩn và cách tiến hành đánh giá chất lượng tiếng nói tổng hợp; cuối cùng là chương 6 giới thiệu sản phẩm của đề tài, kết luận cũng như phương hướng phát triển trong tương lai. Theo bản thuyết minh, nội dung đề tài được chia thành 12 chuyên đề, trong đó có 8 chuyên đề thực hiện các nhiệm vụ lập trình cụ thể, còn lại là về quy trình và dữ liệu. Sự tương ứng giữa trình bày trong báo cáo, sản phẩm phần mềm và các chuyên đề trong bản thuyết minh đề tài như sau:

Chuyên đề 1: sản phẩm phải đạt là quy trình và các nhiệm vụ của hệ TTS được trình bày chủ yếu trong chương 1.

Chuyên đề 3 và 5 là các nghiên cứu về ngữ âm tiếng Việt, được trình bày trong chương 3.

Chuyên đề 4 là nghiên cứu để bước đầu xây dựng Corpus tiếng nói tiếng Việt được thực hiện và trình bày trong chương 3, dữ liệu đã được sử dụng trong quá trình nghiên cứu về ngữ âm tiếng Việt, trường độ các âm vị, âm tiết, dấu cách cũng như sự thay đổi của trường độ trong chương 4 và sử dụng trong phần đánh giá chất lượng của chương 6.

Các chuyên đề 2, 6, 7, 8, 9,1 0, 11, 12 là các công việc về lập trình, các kết quả được thể hiện trong sản phẩm phần mềm cuối cùng Vnspeech, các mô đun chính được giới thiệu trong chương 6.

Nghiên cứu này được thực hiện trong khuôn khổ đề tài cấp Bộ - Bộ Khoa học và Công nghệ (*hợp đồng số 3/HĐ/ĐT- Bộ KHCN&MT, ngày 6/2/2002*) thời gian thực hiện từ 1/2002-12/2003, do Viện Ứng dụng Công nghệ chủ trì. Thông tin giới thiệu, kết quả, phần mềm demo, thư viện lập trình có thể download tại http://www.freewebs.com/vnspeech.

Chương 1

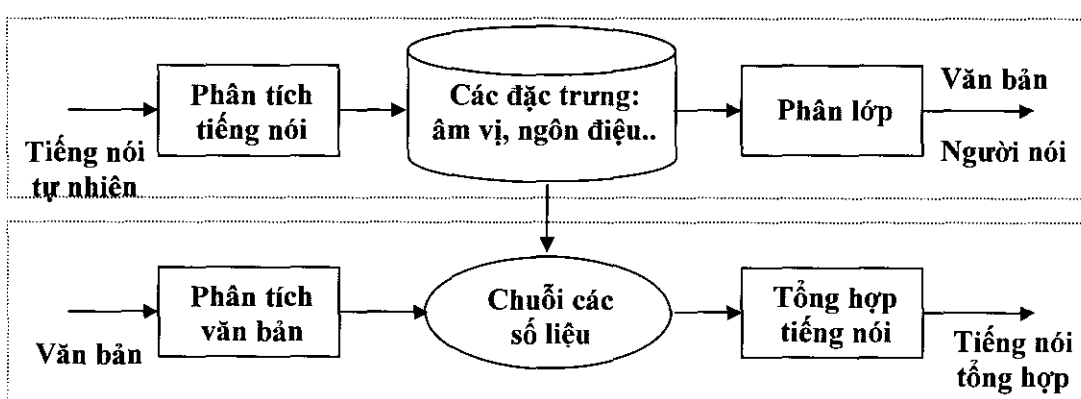
THIẾT KẾ HỆ TTS TIẾNG VIỆT- VNSPEECH

Phần này giới thiệu một số nét chung về xử lý tiếng nói, chuyển văn bản thành tiếng nói để lựa chọn giải pháp tổng hợp tiếng nói cho xây dựng hệ chuyển văn bản thành tiếng nói cho tiếng Việt.

Hầu hết các hệ TTS ngày nay sử dụng một trong hai công nghệ là tổng hợp formant hoặc tổng hợp xích chuỗi (ghép nối) để tạo tín hiệu tiếng nói [Klatt87, Keller02, Tuấn00c]. Mỗi công nghệ đều có các ưu điểm riêng và đây là phần khác biệt khi xây dựng một hệ TTS. Các bước phân tích chuẩn hoá văn bản, xác định thông tin ngữ điệu là công việc chung nhưng phát sinh các tham số điều khiển từ văn bản sẽ phụ thuộc vào công nghệ tổng hợp được lựa chọn.

1.1. Xử lý tiếng nói

Xử lý tiếng nói là thuật ngữ chỉ các nghiên cứu về phân tích tiếng nói, tổng hợp tiếng nói và nhận dạng tiếng nói, người nói. Hình 1.1 là sơ đồ về các công việc chính và mối liên hệ giữa chúng trong nghiên cứu, triển khai xử lý tiếng nói.



Hình 1.1. Các công việc chính của lĩnh vực xử lý tiếng nói

Từ tiếng nói tự nhiên, phân tích để xác định các đặc trưng ngữ âm, qua quá trình phân lớp, nếu để xác định đó là nội dung gì thì công việc này gọi là "*nhận dạng tiếng nói*", nếu để xác định người nói thì đó là "*nhận dạng hay giám định người nói*". Nếu đầu vào là văn bản, căn cứ vào các thông tin dữ liệu về ngữ âm, tạo ra tiếng nói tổng hợp tương ứng với nội dung này thì đó là quá trình "*chuyển văn bản thành tiếng nói*". Ta thấy, phân tích tiếng nói để xác định các thông tin ngữ âm đặc trưng là công việc trung tâm của xử lý tiếng nói.

1.2. Tổng hợp tiếng nói

Tổng hợp tiếng nói là quá trình tạo tiếng nói không phải bằng bộ máy phát âm của con người. Theo hình 1.1, tổng hợp tiếng nói là một trong các nhiệm vụ chính của xử lý tiếng nói. Về chi tiết có thể có nhiều phương pháp, mô hình khác nhau để tạo tiếng nói, nhưng nói chung có thể chia thành 2 loại chính:

1.2.1. Phương pháp trên cơ sở hệ thống

Phương pháp này được gọi là tổng hợp **Articulatory**, là phương pháp tổng hợp trên nguyên tắc tạo một hệ thống (vật lý hay mô phỏng) giống như bộ máy phát âm con người về vị trí, hình dáng cũng như sự dịch chuyển các bộ phận khi cấu âm. Hiện tại phương pháp này mới đạt được một số kết quả ban đầu trong phòng thí nghiệm, tuy nhiên, nó được xem như là một hướng đi tiềm năng để tạo được tiếng nói tổng hợp chất lượng cao. Hiện có một nghiên cứu về tổng hợp Articulatory gián tiếp (mô phỏng) các nguyên âm tiếng Việt đang được tiến hành trong khuôn khổ luận án NCS [Thắng00].

1.2.2. Các phương pháp trên cơ sở tín hiệu

Các phương pháp này dựa trên nguyên tắc tiếng nói là một loại tín hiệu, do vậy nó quan tâm đến việc làm thế nào để sinh ra các tín hiệu giống tiếng nói tự nhiên của con người về các đặc tính đặc trưng như sóng, phổ, năng lượng,

spectrogram, tần số cơ bản, tần số cắt không.... Phương pháp này có nhiều hướng tiếp cận khác nhau.

1.2.2.1. Tổng hợp Concatenation

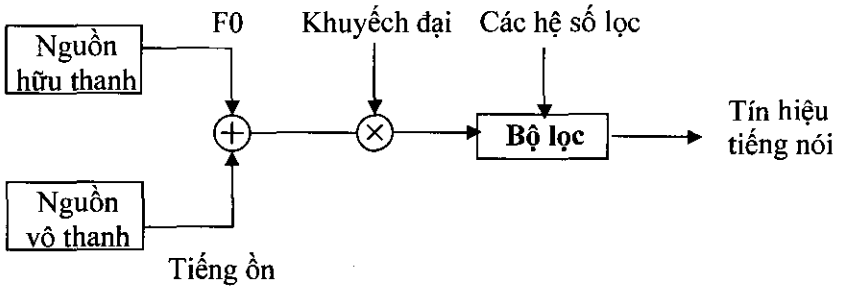
Tổng hợp Concatenation là phương pháp tạo tiếng nói bằng cách phát lại các ghép đoạn tiếng nói tự nhiên ghi trước. Phân loại phương pháp dựa theo chiều dài của các đoạn tiếng nói được ghi, tuy nhiên, nếu đoạn tiếng nói ghi trước là các ngữ đoạn có nghĩa hoặc kịch bản được ghi trước (PRP) thì đó không gọi là tổng hợp mà là hệ thống phát thông báo, do đó đơn vị của phương pháp này thường nhỏ hơn mức từ. Các phương pháp dựa trên Concatenation thường sử dụng kỹ thuật PSOLA (**P**itch **S**ynchronous **O**ver**L**ap **A**dd) để làm trơn điểm ghép nối và thay đổi trường độ cũng như chu kỳ Pitch, thông dụng nhất là TD-PSOLA (PSOLA miền thời gian), ngoài ra còn có FD-PSOLA (PSOLA miền tần số), LP-PSOLA (PSOLA dự đoán tuyến tính).

1.2.2.2. Tổng hợp Formant

Lý thuyết âm học của quá trình tạo tiếng nói con người xem bộ máy phát âm của con người là hệ thống gồm: nguồn âm là đôi dây thanh điều khiển dòng khí thoát ra từ phổi; tuyến âm là các khoang cộng hưởng gồm khoang hầu, khoang miệng và khoang mũi, lưỡi thay đổi vị trí làm thay đổi hình dáng tuyến âm; hình dáng và vị trí đôi môi, sự cho phép hay không cho phép dòng khí thoát qua đường mũi khi nói, cách thoát hơi qua miệng thể hiện đặc tính tán xạ của mô hình. Tuyến âm được mô tả theo hai cách: mô hình tuyến âm nối tiếp - các bộ cộng hưởng được ghép nối tiếp; và mô hình tuyến âm song song - sự cộng hưởng để thể hiện các tần số formant được diễn ra đồng thời.

Tổng hợp formant là phương pháp dựa trên lý thuyết âm học của quá trình tạo tiếng nói [Klatt87, Styger94]. Mô hình bộ tổng hợp là một hệ thống nguồn gồm nguồn âm và các bộ lọc (Hình 1.3). Các tần số formant và các tham số đặc trưng khác là tham số điều khiển mô hình này. Phương pháp này mềm dẻo, tạo được số

lượng âm không hạn chế, yêu cầu dữ liệu lưu trữ nhỏ nhưng độ tự nhiên của tiếng nói tổng hợp chưa cao.



Hình 1.3. Mô hình nguồn âm-bộ lọc (source-filter model)

Ngoài ra, còn các phương pháp khác được nảy sinh từ các kỹ thuật xử lý tín hiệu như: phương pháp dự đoán tuyến tính LPC, phương pháp hình sin, phương pháp dựa trên mã hoá-giải mã phổ...[Minh02a].

Có một số cách phân loại khác như phương pháp tổng hợp trên cơ sở luật và phương pháp trên cơ sở tiếng nói tự nhiên ghi trước nhưng kết luận là cũng vẫn chỉ bao gồm 3 phương pháp chính được nêu trên.

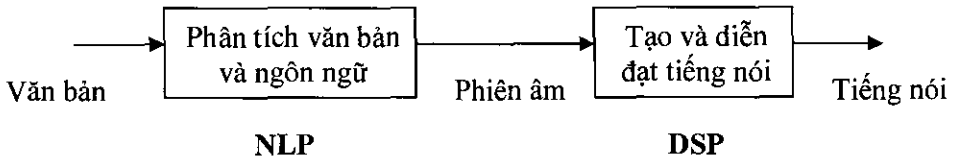
Mặc dù tồn tại nhiều phương pháp tạo tiếng nói tổng hợp khác nhau nhưng hiện chỉ phương pháp Concatenation và phương pháp tổng hợp Formant là được sử dụng trong các hệ TTS hiện nay.

1.3. Chuyển văn bản thành tiếng nói

Chuyển văn bản thành tiếng nói (Text To Speech - TTS) là ứng dụng tự động đọc thành tiếng văn bản sử dụng tiếng nói tổng hợp, đôi khi nó còn thường được hiểu bao gồm cả chuyển khái niệm thành tiếng nói (Concept To Speech - CTS). Như hình 1.1, xây dựng một hệ TTS là một quá trình cần không chỉ tổng hợp tiếng nói mà còn gồm cả phân tích tiếng nói.

Một quá trình TTS thường được chia thành hai giai đoạn: 1) Phân tích văn bản, chuyển văn bản đầu vào thành dãy các phiên âm hoặc một sự biểu diễn ngôn ngữ nào đó; và 2) Tạo tín hiệu tiếng nói (tổng hợp tiếng nói), âm thanh tiếng nói đầu

ra được tạo ra từ thông tin về phiên âm và ngữ điệu của giai đoạn trước. Hai giai đoạn này còn được gọi ***Xử lý ngôn ngữ tự nhiên (NLP)*** và ***Xử lý tín hiệu số (DSP)***, có thể minh hoạ bằng sơ đồ hình 1.2.



Hình 1.2. Mô hình văn bản thành tiếng nói

Văn bản đầu vào có thể từ các chương trình xử lý văn bản, trang web, thư điện tử hoặc các nguồn có thể chuyển thành chuỗi ký tự. Chuỗi ký tự sau đó được phân tích chuẩn hoá thành biểu diễn ngữ âm duy nhất, thường là một chuỗi các âm vị với các thông tin như ngữ điệu, trường độ và độ nhấn mạnh. Bộ tổng hợp tiếng nói trực tiếp tạo ra âm thanh tiếng nói từ thông tin cung cấp từ phần xử lý văn bản, các phương pháp tổng hợp tiếng nói chính đã được giới thiệu chi tiết tại phần trên.

1.3.1. Xử lý ngôn ngữ tự nhiên

Nhiệm vụ đầu tiên của bất kỳ hệ thống *"Chuyển văn bản thành tiếng nói"* nào là chuyển đổi văn bản đầu vào thành dạng biểu diễn về ngữ âm. Quá trình này phụ thuộc vào từng ngôn ngữ cụ thể. Với các ngôn ngữ mà văn bản được viết gần như tương ứng với cách phát âm thì sự chuyển đổi khá đơn giản, chẳng hạn như tiếng Việt. Một số ngôn ngữ như tiếng Anh do cách viết khác với cách đọc nên sự chuyển đổi phức tạp hơn. Để chuyển đổi, bao giờ cũng cần một tập các quy tắc chuyển đổi và những ngoại lệ. Sự phức tạp chính ở phần ngoại lệ và khả năng có thể dùng các quy tắc đơn giản để biểu diễn quy luật và có thể mô tả hết các tình huống của ngôn ngữ hay không. Sự chuyển đổi có thể chia thành ba bước là tiền xử lý văn bản, tạo dữ liệu ngôn ngữ cho phát âm đúng và phân tích của những đặc tính diễn đạt cho đúng về ngữ điệu nhấn mạnh và khoảng thời gian.

1.3.1.1. Tiền xử lý văn bản

Xử lý trước văn bản là công việc chuẩn hoá, xác định các ký tự đọc hay không, các ký tự ngắt nghỉ, đưa về dạng viết đầy đủ của các dạng khác như: biểu thức số, ngày-tháng, chữ viết tắt, tên riêng, từ lạ... Có nhiều tình huống có thể gây nhập nhằng, muốn xác định được chính xác cần phải tiến hành phân tích văn phạm, ngữ pháp, hiểu văn bản. Chẳng hạn, số 8695484 sẽ đọc là "tám sáu chín năm bốn tám bốn" nếu là số điện thoại, còn sẽ đọc là "tám triệu sáu trăm chín mươi nhăm ngàn bốn trăm tám mươi tư" nếu là một số. Phân số và ngày tháng có thể gây nhầm lẫn, 1/6 có thể là "một phần sáu" (phân số) hoặc "ngày mồng một tháng sáu" (ngày-tháng). Các chữ số La mã cũng có thể gây nhập nhằng như "I" có thể là số 1 hoặc chữ i, hoặc nhầm lẫn với một số viết tắt phổ biến như MCM. Ta có thể chọn mở rộng viết tắt thành từ đầy đủ hoặc đọc kiểu đánh vần từng ký tự. Biểu thức "1-3" có thể được đọc như "một trừ ba" hoặc "một ba" (tỷ số) hay "một đến ba" (liệt kê).

1.3.1.2. Biểu diễn ngữ âm

Từ dãy thuần ký tự của một ngôn ngữ, cần phải chuyển thành biểu diễn duy nhất về ngữ âm. Luôn có hiện tượng một ký tự có thể biểu diễn vài âm vị khác nhau và một âm vị có thể được viết bằng một số ký tự khác nhau. Bảng chữ cái ngữ âm Quốc tế (IPA) là một trong các giải pháp để giải quyết vấn đề này.

1.3.1.3. Ngôn điệu

Xác định đúng ngữ điệu, nhấn mạnh, độ kéo dài phần phát âm và phần nghỉ từ văn bản viết là vấn đề cần phải quan tâm nhất trong tất cả các hệ thống TTS để tăng chất lượng [Keller02]. Những đặc tính này được gọi chung là ngôn điệu, là cách diễn đạt hay các đặc tính siêu đoạn và có thể được xem như giai điệu, nhịp điệu và nhấn mạnh của tiếng nói tại mức cảm thụ. Ngữ điệu có nghĩa là đường nét của Pitch hoặc tần số cơ bản thay đổi như thế nào trong khi nói. Cách diễn đạt của tiếng nói liên tục phụ thuộc vào một số khía cạnh như nghĩa của câu, đặc trưng và cảm xúc

của người nói.

Xác định trường độ tại mức câu hoặc nhóm các từ thành cụm từ cho chính xác là vấn đề khó vì sự phân đoạn ngôn điệu không phải thường xuyên được đánh dấu bằng dấu chấm câu trong văn bản, và sự nhấn mạnh cụm từ cũng không được đánh dấu rõ ràng. Nếu không có những sự tạm dừng hơi trong lúc nói hoặc dừng sai chỗ, tiếng nói nghe sẽ không tự nhiên hoặc thậm chí nghĩa của câu có thể bị hiểu sai. Trong tiếng Việt, sự thay đổi cao độ và trường độ các âm vị trong một âm tiết còn diễn tả một âm tiết khác (thanh điệu).

1.3.2. Ứng dụng của TTS

TTS được ứng dụng trong nhiều lĩnh vực khác nhau. Trong tương tác người-máy bằng tiếng nói (IVR): TTS giúp máy đưa ra các thông báo cho người dùng bằng tiếng nói thay vì hiển thị văn bản hoặc các đèn hiệu, khả năng này sẽ rất có ích trong các tình huống mắt người dùng đang bận phải quan sát như đang lái xe.

Trong truyền thông: tích hợp vào các hệ thống truyền thông thông điệp hợp nhất, lúc này kể cả thư điện tử có thể được đọc cho người nhận qua đường thoại thay vì phải trực tiếp mở và đọc bằng máy tính.

Trợ giúp người khuyết tật: Một hệ thống gồm phần mềm gồm máy quét, phần mềm nhận dạng ký tự (có thể gồm phần mềm dịch tự động) và sau đó chuyển văn bản thành tiếng nói sẽ rất có ích cho các người bị khiếm thị. Các thiết bị tích hợp nhỏ gọn có thể dùng cho từng cá nhân, các hệ thống nhiều tính năng sẽ rất có ý nghĩa trong các phòng đọc hay thư viện lớn, phục vụ nhiều đối tượng. Ngoài ra, ta còn có thể nghĩ đến một thiết bị tích hợp chuyển văn bản, nội dung thành tiếng nói có thể giúp người câm giao tiếp bằng tiếng nói (tổng hợp) thay vì ngôn ngữ cử chỉ hay chữ viết.

Các phần mềm ứng dụng: Tích hợp TTS sẽ tạo cho các phần mềm phong phú hơn khi cần thông báo với người dùng, thay vì chỉ thuần tuý đưa ra các thông báo, kết quả dạng văn bản, nay có thể thêm tích năng tiếng nói. Ngoài ra, có thể thiết kế các phần mềm, tính năng đọc thành tiếng là một ưu điểm quan trọng để làm việc tốt

như: phần mềm soát lỗi chính tả, lỗi sẽ dễ được phát hiện hơn khi nghe so với người tự đọc bằng mắt.

Loại ứng dụng người dùng có thể truy cập từ xa qua đường thoại như các ứng dụng khai thác, truy vấn cơ sở dữ liệu.

Các hệ thống cần phát ra các thông báo bằng tiếng nói với nội dung thay đổi, lượng từ vựng lớn.

Khi tiếng nói tổng hợp có chất lượng cao, nó có thể ứng dụng rộng rãi trong ngành giáo dục như tạo các phần mềm công cụ dạy đọc, dạy nói cũng như để giảng bài, đọc bài.

1.3.3. Một số hệ TTS

- **DECtalk** của **Fonix Corporation**

Là sản phẩm TTS nổi tiếng nhất, được kế thừa từ các hệ MITalk và Klattalk [Klatt87]. Phiên bản Fonix DECtalk 5.0 cho các ứng dụng nhúng vào các thiết bị cầm tay được giới thiệu 1/10/2003, chạy trên các hệ điều hành Linux and Pocket PC.
Công nghệ: tổng hợp Formant theo mô hình của Klatt
Gồm 9 giọng nói và 6 ngôn ngữ (U.S. and UK English, Castilian and Latin American Spanish, German and French).
Công nghệ TTS của Fonix DECtalk được Eintech Co. Ltd. (Korea) tích hợp cho thiết bị Eintech's Magic Talker's Personal Bilingual Assistant (model EK-D8800) 1/10/2002. [SpeechTech]

- **ETI-Eloquence SF** của **SpeechWorks**

Công nghệ: tổng hợp formant theo mô hình của Klatt.
Ngôn ngữ: 13 ngôn ngữ là: U.S. English, UK English, Continental French, Castilian Spanish, German, Japanese, Brazilian Portuguese, Mexican Spanish, Canadian French, Finnish, Italian, Korean và Mandarin Chinese.
Tích hợp ngầm định vào phần mềm trợ giúp người khiếm thị JAWS.[Freedom]
Engine TTS để tích hợp các thiết bị cầm tay giới thiệu 22/4/2003.[SpeechTech]

- **SenSyn** của **Sensimetrics Corporation**

Ngôn ngữ: tiếng Anh;

Công nghệ: trên cơ sở bộ tổng hợp formant của Klatt. [Sensimetrics]

- **SVTTS** của **SoftVoice**

Công nghệ: tổng hợp formant. Sản phẩm Talk It! có hai ngôn ngữ là tiếng Anh và tiếng Tây ban nha với 20 giọng được xây dựng sẵn và có thể điều khiển mềm dẻo các tham số của tiếng nói.[SoftVoice]

- **Infovox** của **Telia Promotor AB**

Công nghệ: dựa trên tổng hợp formant, Infovox 230, gồm tiếng Anh-Anh, Anh-Mỹ, Đan mạch, Phần lan, Pháp, Đức, Ai xơ len, Ytalia, Na uy, Tây ban nha, Thụy Điển, và Hà lan.

- **Bell Labs Text-to-Speech** của **AT&T (Lucent Technologies)**

Công nghệ: ghép nối các đoạn diphone, triphone, có thể được mã hoá dưới dạng các tham số LPC.

Ngôn ngữ: gồm tiếng Anh, Pháp, Tây ban nha, Ytalia, Đức, Nga, Rumani, Trung quốc và Nhật...

- **ProVerbe** của **CNET** (Centre National d'Etudes Telecommunications) - France Telecom.

Công nghệ: ghép nối các diphone, sử dụng kỹ thuật PSOLA.

Ngôn ngữ: tiếng Anh, Anh-Mỹ, Pháp, Đức, và Tây ban nha.

- **TTS3000/M** của **Lernout & Hauspie**

Công nghệ: ghép nối các đoạn diphone, triphone và tetraphone.

Ngôn ngữ: tiếng Anh – Mỹ, Đức, Hà lan, Tây ban nha, Ytalia và Triều tiên

- **Festival**

Phát triển tại CSTR - University of Edinburgh bởi Alan Black và Paul Taylor, hợp

tác với CHATR ATR Nhật Bản.

Công nghệ: ghép nối, sử dụng các kỹ thuật như kích thích dư LPC, PSOLA và MBROLA. Ngôn ngữ: tiếng Anh-Mỹ, Anh, Tây ban nha và Welsh.

Được cung cấp miễn phí như thư viện phần mềm để phát triển tổng hợp ngôn ngữ bất kỳ [Festival]

- **Whistler** của **Microsoft**

Là một hệ thống TTS có thể dạy được, hiện được tích hợp vào Windows XP, 2000 và engine TTS được tích hợp trong thư viện lập trình SAPI.

Công nghệ: ghép nối [Acero], huấn luyện dựa vào mô hình Markov ẩn (HMM) [Donovan96].

Ngôn ngữ: tiếng Anh

- **VTalk**

Phát triển tại Viện Khoa học Kỹ thuật Bưu điện [Tuấn00c].

Công nghệ: ghép nối, đơn vị là phụ âm đầu và vần, sử dụng kỹ thuật PSOLA.

Ngôn ngữ: tiếng Việt, 1 giọng nam;

- **VnVoice**

Phát triển tại Viện Công nghệ Thông tin – Trung tâm Khoa học Tự nhiên và Công nghệ Quốc gia.

Công nghệ: ghép nối, đơn vị là phụ âm đầu và vần, sử dụng kỹ thuật PSOLA.

Ngôn ngữ: tiếng Việt.

1.4. Phương án xây dựng hệ TTS tiếng Việt- VnSpeech

Phần này phân tích các ưu điểm và nhược điểm của hai công nghệ là tổng hợp formant và tổng hợp ghép nối được áp dụng phổ biến trong các hệ TTS, trên cơ sở các kết luận được rút ra từ quá trình nghiên cứu xây dựng các hệ TTS cho các ngôn ngữ khác, khả năng áp dụng cho tiếng Việt và lý do đề tài chọn hướng tiếp cận theo công nghệ tổng hợp formant.

1.4.1. TTS dựa trên tổng hợp xích chuỗi

Ưu điểm được nhắc đến khi nói đến tổng hợp xích chuỗi là phương án này dễ triển khai, có tiếng nói tự nhiên và vấn đề phải quan tâm là quyết định độ dài các đoạn tiếng nói tự nhiên để làm đơn vị ghép nối. Nhược điểm của phương pháp này là dữ liệu lớn, khó thay đổi giọng nói, khả năng điều khiển các tham số ngữ điệu hạn chế.

Tổng hợp xích chuỗi đối với các ngôn ngữ như tiếng Anh, Pháp vấn đề lựa chọn độ dài đoạn tiếng nói tự nhiên làm đơn vị âm lưu trữ khá phức tạp. Nếu chọn đơn vị là từ thì số lượng sẽ rất lớn, nếu chọn đơn vị là âm tiết thì như tiếng Anh, cũng có đến 10000 âm tiết và đồng thời ghép nối giữa các âm tiết rất phức tạp vì hiệu ứng liên cấu âm và các sự biến âm, luyến âm, nuốt âm; nếu chọn đơn vị là âm vị thì có số lượng đơn vị nhỏ nhưng sự ghép nối các âm vị rất phức tạp như lý do ghép nối âm tiết. Do vậy, hầu hết các hệ TTS cho tiếng Anh, Pháp đều chọn đơn vị lưu trữ là diphone (hai nửa âm vị liền nhau) để chứa sẵn hiệu ứng liên cấu âm khi ghép nối. Kỹ thuật phổ biến nhất được sử dụng để làm trơn điểm ghép nối và thay đổi các yếu tố cao độ, trường độ là TD-PSOLA.

Do đặc điểm tiếng Việt là ngôn ngữ đơn âm tiết, không có sự luyến âm, nuốt âm khi cấu âm và một âm tiết tiếng Việt có thể chia thành 3 thành phần có mối liên kết lỏng kẻo là phụ âm đầu, vần và dấu thanh (về mặt này, tiếng Việt thuận tiện hơn tiếng Thái, cũng là một ngôn ngữ thanh điệu [Pradit00]) nên xử lý hiệu ứng liên cấu âm không quá phức tạp nếu các đơn vị được lựa chọn đồng đều. Đồng thời, số lượng âm tiết hay dùng trong tiếng Việt cũng không quá lớn, số lượng vần hay âm vị càng nhỏ hơn (bảng 1.1) cho nên, đối với tiếng Việt, dường như các khó khăn về số lượng các đơn vị cũng như hiệu ứng cấu âm khi ghép nối như với tiếng Anh, Pháp không gặp phải, lựa chọn đơn vị chủ yếu liên quan đến độ mềm dẻo của chương trình. Bảng 1.1 liệt kê một số phương án có thể sử dụng khi xây dựng hệ TTS dựa trên ghép nối cho tiếng Việt.

Bảng 1.1. Một số phương án lựa chọn đơn vị cho tổng hợp xích chuỗi tiếng Việt

Stt	Đơn vị lưu trữ	Số lượng	Ghi chú
1	Âm tiết	< 7000	Các âm tiết hay dùng nhất
2	Phụ âm đầu và vần có dấu	< 800	22 phụ âm đầu, ~700 vần có dấu
3	Phụ âm đầu và vần không dấu	< 200	22 phụ âm đầu. 155 vần
4	Âm vị	39	Toàn bộ âm vị tiếng Việt
5	Diphone	~ 1600	Bán âm vị và ngữ cảnh

Đã có một số nghiên cứu cũng như sản phẩm TTS trên cơ sở ghép nối cho tiếng Việt. Các phương án 1,2,3 đều đã được triển khai, trong đó phương án 1 có nhiều thử nghiệm nhất (do dễ triển khai), tuy nhiên các hệ thương phẩm hầu hết chọn phương án 2,3 hay kết hợp các phương án. Các hệ thống cho tiếng Việt cũng sử dụng công nghệ TD-PSOLA để làm trơn điểm ghép nối và biến đổi các tham số cao độ và trường độ [Hùng03, Tuấn00b, Tuấn00c].

1.4.2. TTS trên cơ sở tổng hợp formant

Khuyết điểm hay được nhắc đến khi bàn về mô hình tổng hợp formant dựa trên mô hình nguồn âm-bộ lọc là tiếng nói tạo ra nghe "robot" vì mô hình này mô tả tốt cho âm hữu thanh và các tần số formant nhưng không có các đặc trưng vật lý của tuyến âm. Ưu điểm của tổng hợp formant là dữ liệu và chương trình rất nhỏ, đặc biệt có thể điều khiển mềm dẻo các thông số đặc trưng của tiếng nói, điều này rất quan trọng khi xây dựng các hệ TTS chất lượng cao. Mô hình tổng hợp tiếng nói formant tiêu biểu nhất là mô hình của Klatt, đã có các sản phẩm thương mại nổi tiếng như DECtalk (tiền thân là MITALK) thành công với mô hình này, hệ này thường đạt điểm cao trong các đánh giá chất lượng [Klatt87] và vẫn được tiếp tục phát triển và sử dụng rất nhiều. Nhiều hệ TTS cũng như các nhà nghiên cứu về tổng hợp tiếng nói cho các ngôn ngữ khác nhau đã sử dụng mô hình này [Bangayan97]. Một trong các kết quả ấn tượng là thí nghiệm về sao chép tiếng nói: các tham số đặc trưng được phân tích ra từ tiếng nói tự nhiên, sau đó được điều chỉnh bằng tay cho bộ tổng hợp formant của Klatt, kết quả tạo được tiếng nói tổng hợp không thể phân

biệt được với tiếng nói tự nhiên. Điều này nói lên là có thể tổng hợp được tiếng nói với chất lượng rất cao, khi tạo được các tham số điều khiển thích hợp [Klatt87].

Tổng hợp formant cũng chính là nghiên cứu phân tích ngữ âm của một ngôn ngữ, các thông số đặc trưng chỉ thực sự là đúng đắn khi có thể sử dụng để tổng hợp lại được. Như đã chỉ ra trong hình 1.1, phân tích là một nhiệm vụ trung tâm của xử lý tiếng nói, như vậy, nghiên cứu tổng hợp formant là nội dung không thể bỏ qua khi đặt ra vấn đề nghiên cứu xử lý tiếng nói một cách cơ bản, toàn diện.

Tuy nhiên, triển khai xây dựng bộ tổng hợp formant cho một ngôn ngữ cũng như hệ TTS dựa trên tổng hợp formant không phải là công vệc dễ dàng, như trong [Tuấn00] đã nhận xét:

> "Đối với các ngôn ngữ khác, kể cả tiếng Việt, thu thập được đủ số liệu, đủ kiến thức để phân tích và xây dựng được hệ thống các quy luật tổng hợp bằng formant không những chỉ là khối lượng công việc khổng lồ mà còn cần có kiến thức rất sâu rộng về ngữ âm. Hệ thống tổng hợp MITALK cho tiếng Anh-Mỹ là một ví dụ, có nền xuất phát rất cao vì đã thừa hưởng được nhiều kết quả của các nhà khoa học trước đó, cũng còn cần hơn 10 năm của cả tập thể nghiên cứu của trường đại học MIT trước khi sản phẩm có thể chuyển giao thương mại hoá. Vì vậy, cho tới ngày nay, phương pháp này cũng chỉ thành công cho một số ít các ngôn ngữ có nền tảng khoa học công nghệ tiên tiến."

và đến nay, chưa xuất hiện bất kỳ sản phẩm nào có thể tổng hợp formant tiếng Việt. Nhận xét trên hoàn toàn đúng, tuy nhiên, không có nghĩa là không thể tổng hợp formant tiếng Việt, đồng thời tổng hợp formant đạt được chất lượng cao cho tiếng Anh thì cũng sẽ tổng hợp đạt được chất lượng cao cho tiếng Việt, vì đây là mô hình của bộ máy phát âm của con người về tín hiệu. Điều cần phải nghiên cứu là làm sao có được các tham số điều khiển thích hợp với các âm vị cũng như ngữ âm của tiếng Việt. Mặc dù có thể thừa hưởng được nhiều kết quả nghiên cứu cho các ngôn ngữ khác, nhưng nghiên cứu xử lý tiếng Việt còn rất nhiều việc phải làm, đặc biệt các nghiên cứu về ngữ âm tiếng Việt.

1.4.3. Lựa chọn phương án

Còn rất nhiều việc phải làm với các sản phẩm TTS [Sproat99], một trong các tâm điểm hiện nay của các nhà nghiên cứu về xử lý tiếng nói và tổng hợp tiếng nói từ văn bản, cũng như để nâng cao chất lượng các hệ TTS là ngữ điệu. Ngữ điệu là vấn đề thách thức và quan tâm nhất của các nhà nghiên cứu TTS hiện nay [Keller02, VanSanten97], có hai vấn đề cần phải giả quyết là: 1) Xác định các thông tin ngữ điệu từ văn bản; và 2) Tổng hợp các thông tin diễn tả ngữ điệu. Một nhược điểm của phương án xích chuỗi chưa được nhấn mạnh khi xem xét các hệ loại này ở mức bình thường là khả năng biến đổi các thông số như cao độ, trường độ, năng lượng để thể hiện ngữ điệu rất hạn chế. Hiện đang có nhiều nghiên cứu tìm kỹ thuật khác thay thế PSOLA để ghép nối và điều chỉnh tín hiệu [Acero]. Cho nên, nói chính xác là tổng hợp xích chuỗi cũng gặp phải vấn đề về tính tự nhiên ở mức câu, trong khi đó phương án tổng hợp formant có thể điều khiển các tham số này rất mềm dẻo và hiển nhiên.

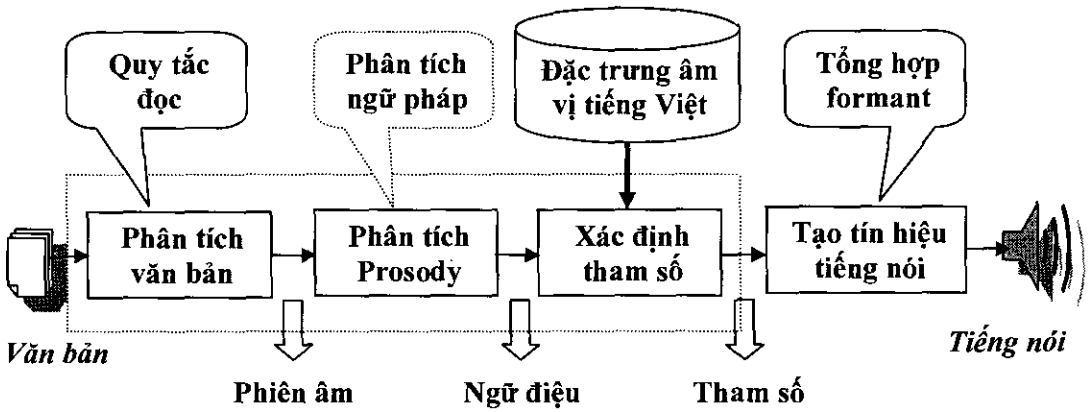
TTS trên cơ sở tổng hợp formant, ngoài mục tiêu chuyển văn bản (nội dung) thành tiếng nói, nó còn là nghiên cứu cơ bản có ý nghĩa quan trọng trong lĩnh vực nghiên cứu ngữ âm của một ngôn ngữ. Tổng hợp là quá trình ngược của phân tích, nên đây là một trong các phương pháp phân tích: ***phân tích bằng tổng hợp***, và phương pháp phân tích này còn chưa có điều kiện sử dụng đối với tiếng Việt.

Căn cứ vào lý do trên, cùng với sự cần thiết phải tiến hành nghiên cứu sâu sắc hơn nữa về xử lý tiếng nói tiếng Việt, hướng có thể phát triển tiếp tục của nội dung nghiên cứu, tiếp tục các kết quả đã đạt được của nghiên cứu khảo sát giai đoạn trước [Minh02a], đề tài đã chọn theo hướng xây dựng hệ TTS dựa trên tổng hợp formant, sử dụng mô hình tổng hợp của Klatt.

1.5. Mô hình hệ TTS tiếng Việt - VnSpeech

Tuân theo cấu trúc chung, hệ VnSpeech - chuyển văn bản tiếng Việt thành tiếng nói trên cơ sở tổng hợp formant được xây dựng theo sơ đồ hình 1.4. Trong mô hình này, phần xử lý văn bản để cung cấp thông tin cho bộ tổng hợp được chia

thành 3 phần: đó là phân tích, chuẩn hoá văn bản sau đó phân tích xác định các đặc tính về ngữ điệu và phân tích xác định các tham số ngữ âm đặc trưng. Phần xử lý tín hiệu số, tạo tín hiệu tiếng nói là bộ tổng hợp tiếng nói formant hỗn hợp của Klatt.



Hình 1.4. Mô hình hệ VnSpeech

Trình bày chi tiết về quá trình xây dựng hệ chuyển từ văn bản thành tiếng nói tiếng Việt - Vnspeech sẽ được trình bày trong các chương tiếp sau.

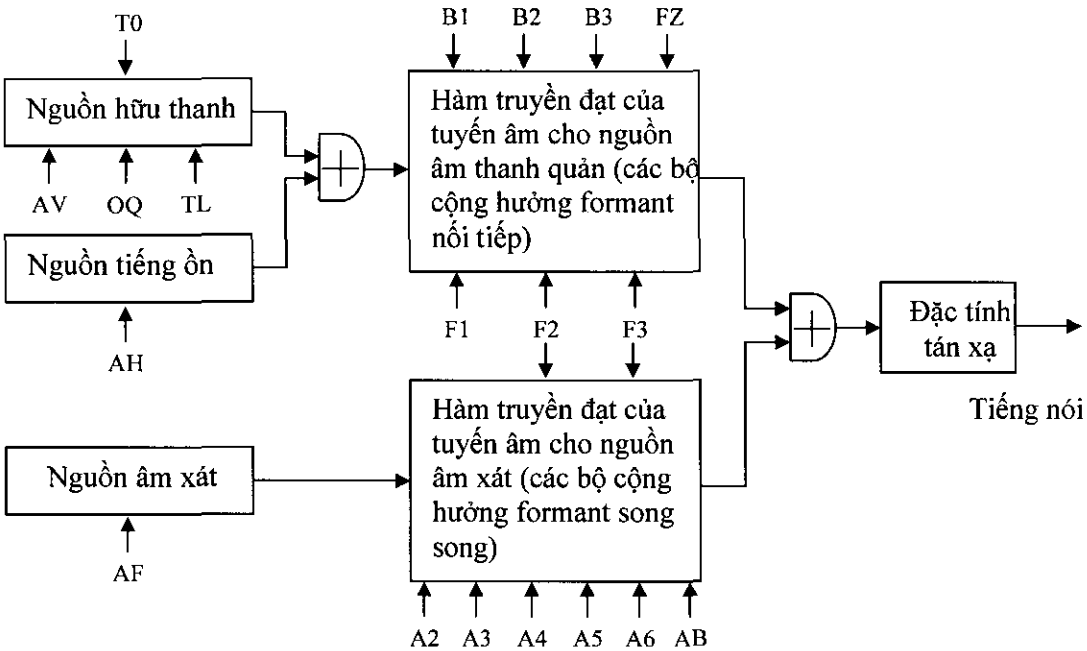
Chương 2

BỘ TỔNG HỢP TIẾNG NÓI FORMANT

Bộ tổng hợp tiếng nói là phần xử lý tín hiệu số, tạo tín hiệu tiếng nói từ các tham số điều khiển. Hệ TTS tiếng Việt - VnSpeech được xây dựng dựa trên mô hình tổng hợp formant hỗn hợp của Klatt. Như đã biết, khả năng của mô hình của Klatt là có thể tạo được bất kỳ âm thanh nào giống bộ máy phát âm của con người, nhiệm vụ ở đây là "***dạy nói***" các âm vị tiếng Việt và sau đó là "***dạy đọc***" văn bản tiếng Việt. Phần này trình bày về mô hình tổng hợp của Klatt và giải pháp để nó có thể "*nói*" tiếng Việt. Các phần sau sẽ lần lượt trình bày các nghiên cứu để "*dạy đọc*" văn bản tiếng Việt.

2.1. Mô hình tổng hợp của Klatt

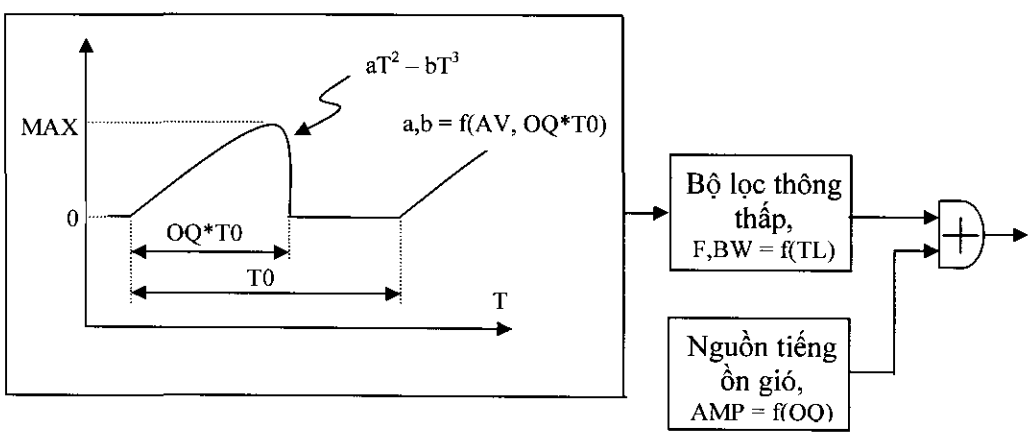
Mô hình tổng hợp tiếng nói của Klatt [Klatt87, Styger94] mô phỏng quá trình tạo tiếng nói của con người dựa trên nguyên lý nguồn âm-bộ lọc của quá trình tạo tiếng nói, đây là mô hình tổng hợp formant hỗn hợp bao gồm cả tuyến âm nối tiếp và song song với nguồn kích phức hợp. Sơ đồ khối bộ tổng hợp được trình bày trong hình 2.1 với các biến tham số quan trọng nhất để điều khiển nguồn âm và tuyến âm, các tham số điều khiển khác thường được gán giá trị ngầm định và không trình bày ở đây.



Hình 2.1. Sơ đồ khối bộ tổng hợp của Klatt

2.1.1. Nguồn kích thích

Nguồn kích thích gồm nguồn hữu thanh được tạo bởi các tín hiệu tuần hoàn và nguồn vô thanh được sinh ra từ các tín hiệu ngẫu nhiên. Nguồn hữu thanh của bộ tổng hợp formant của Klatt được trình bày trong hình 2.2.



Hình 2.2. Nguồn hữu thanh

Nguồn kích hữu thanh được điều khiển bởi 4 tham số là: OQ, TL, AV và T0 trong đó:

OQ : hệ số mở

TL : độ nghiêng phổ

AV : biên độ hữu thanh

T0 : chu kỳ lấy mẫu cơ bản (= 1/F0 : tần số cơ bản)

Hàm sóng hữu thanh cơ sở có dạng: $aT^2 - bT^3$

Trong đó:

T : biến thời gian

các hệ số a, b là hàm của AV và OQ*T0

Ưu điểm của nguồn hữu thanh này là tốc độ âm lượng sóng cửa hầu được định nghĩa tốt tại các thời điểm đóng, mở với hình dáng không đều, tốc độ đóng nhanh hơn tốc độ mở. Tốc độ âm lượng sóng hữu thanh tuân theo hàm trên trong suốt pha mở của chu kỳ và bằng 0 trong thời gian còn lại. Phổ của nguồn tự nhiên một số chỗ không đồng nhất với một điểm 0 yếu tại khoảng 600 Hz. Có thể điều chỉnh để phổ nghiêng hơn, sử dụng hoặc OQ hoặc TL để phỏng theo hiệu ứng đóng cửa hầu không hoàn toàn và sự làm tròn góc của sóng âm vào lúc kết thúc.

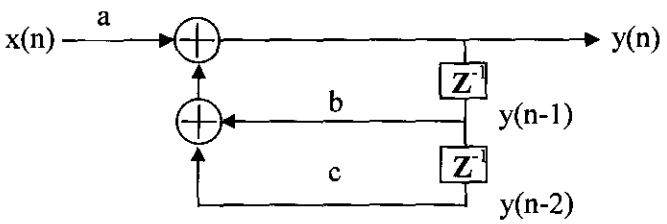
Khuyết điểm của sóng nguồn tự nhiên là độ lớn phổ một số chỗ không đều do vậy formant sẽ hơi mỏng bớt khi nó gần tần số 600 Hz (vị trí điểm 0 thực sự phụ thuộc vào OQ). Biên độ formant này thay đổi giống như xuất hiện trong tiếng nói tự nhiên.

Nguồn vô thanh để mô tả kích thích khi tạo các âm vô thanh là bộ sinh số ngẫu nhiên. Trong pha mở của đôi dây thanh, kích thích từ nguồn hữu thanh được kết hợp với tín hiệu từ nguồn tiếng ồn ngẫu nhiên để mô tả kích thích cho các âm bật hơi.

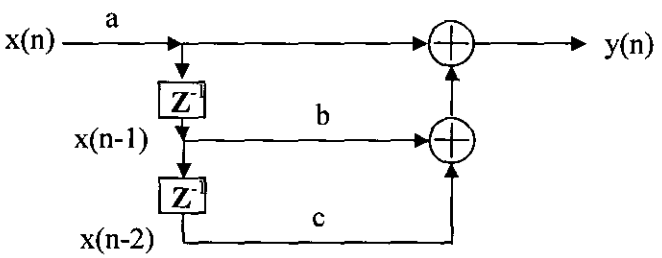
Sự tán xạ được thể hiện trong nguồn kích thích hữu thanh bằng cách cộng với đạo hàm bậc nhất của giá trị quá khứ.

2.1.2. Tuyến âm

Tuyến âm gồm 2 nhánh: nhánh nối tiếp và nhánh song song. Mỗi nhánh gồm các bộ lọc bậc 2 diễn tả tần tần số cộng hưởng và phản cộng hưởng của tín hiệu tiếng nói (hình 2.3).



(a) Hệ truy hồi bậc hai



(b) Hệ không truy hồi bậc hai

Hình 2.3. Sơ đồ các bộ lọc bậc hai

Hình 2.3(a) biểu diễn bộ lọc thông dải - bandpass (còn gọi là hệ truy hồi bậc hai) mô tả tần số cộng hưởng (điểm cực). Bộ lọc được điều khiển bằng các thông số là tần số lấy mẫu, tần số cộng hưởng và độ rộng dải thông của nó. Các hệ số a, b, c được thiết lập theo các hệ thức sau:

r = exp((-PI*BW)/Fs)

c = -(r*r)

b = 2*r *cos(2*PI*BW*f/Fs)

a = 1 - b - c

Trong đó:

Fs : tần số lấy mẫu (= 1/Ts: Chu kỳ lấy mẫu)

f : tần số cộng hưởng

BW : độ rộng dải thông

PI : hệ số pi (~ 3,1415927)

Tín hiệu ra y(n) được lọc từ tín hiệu vào x(n) theo phương trình:

y(n) = a*x(n) + b*y(n-1) + c*y(n-2)

Hình 2.3(b) biểu diễn bộ lọc chắn dải – bandstop (còn gọi là hệ không truy hồi bậc hai) mô tả tần số phản cộng hưởng (điểm không). Bộ lọc cũng được điều khiển bằng các thông số là tần số lấy mẫu, tần số phản cộng hưởng và băng thông của nó. Các hệ số a, b, c được thiết lập như với bộ lọc thông dải với một số thay đổi sau:

f = -f

a = 1/a

b = 1/b

c = 1/c

Tín hiệu ra y(n) được lọc từ tín hiệu vào x(n) theo quan hệ:

y(n) = a.x(n) + b.x(n-1) + c.x(n-2)

Chú ý: các giá trị x(0), x(-1), y(0), y(-1) được khởi tạo bằng 0. Biên độ cộng hưởng A được mô tả bằng cách nhân với hệ số a. (A*a)

Nhánh nối tiếp của tuyến âm gồm 8 bộ cộng hưởng mô tả 8 tần số formant, 1 bộ mô tả điểm cực cho âm mũi và 1 bộ phản cộng hưởng mô tả điểm 0 âm mũi. Nhánh song song gồm 6 bộ cộng hưởng cho 6 tần số formant và 1 bộ cho điểm cực âm mũi. Ngoài ra còn 1 bộ lọc xung cửa hầu, 1 bộ lọc thông thấp cho nguồn tiếng ồn ngẫu nhiên, 1 bộ lọc thể hiện sự tán xạ âm qua miệng và mũi.

2.1.3. Đặc tính tán xạ

Đặc tính tán xạ được mô tả bằng bộ lọc thông cao, diễn tả sự tán xạ của âm ra ngoài qua mũi hoặc miệng. Để thể hiện điều này trong tính toán thực tế, đặc tính

tán xạ được tích hợp vào nguồn kích và trong quá trình cộng hưởng bằng cách cộng thêm đạo hàm bậc nhất của tín hiệu trước đó.

2.2. Các tham số điều khiển

Mô hình tổng hợp formant hoạt động để tạo các tín hiệu tiếng nói khác nhau bằng các tham số điều khiển. Các tham số điều khiển được chia thành hai loại: các hằng số, thiết lập giá trị cho toàn phiên làm việc và các biến số, nhận các giá trị thay đổi theo mỗi khoảng cập nhật. Mỗi hằng số hay biến số được định nghĩa một khoảng giá trị (cực tiểu, cực đại) và một giá trị ngầm định khi khởi tạo.

2.2.1. Các hằng số

Có 3 hằng số được thiết lập cho mỗi phiên tổng hợp, Vnspeech cho phép điều chỉnh trực quan các giá trị này.

1. Tần số lấy mẫu (sr): là số mẫu cần phải tạo ra ứng với 1 giây tiếng nói tổng hợp. Giá trị ngầm định là 10000 mẫu/giây (Hz). Nếu sr tăng, phổ của tiếng nói tổng hợp sẽ nghiêng hơn, sử dụng một bộ lọc thông thấp chống trùm phổ với tần số cắt khoảng 4500 - 4800 Hz cho giá trị 10000 Hz. Do vậy, nếu sr thay đổi, cần sử dụng bộ lọc với tần số cắt thích hợp.

2. Khoảng cách cập nhật (ui): là số ms của sóng âm được tạo giữa các lần cập nhật các biến tham số (đoạn đủ ngắn để các thuộc tính của tiếng nói được coi là tuyến tính, bất biến). Giá trị 5 ms là có thể phản ánh hầu hết các thay đổi nhanh chóng của các tham số tiếng nói, tuy nhiên thực tế chỉ cần sử dụng giá trị 10 ms đã là đủ.

Các tham số liên quan đến tạo nguồn kích thích cửa hầu như (F0, AV, OQ, TILT, SKEW) là không thay đổi chính xác tại thời điểm cập nhật chỉ ra bởi 'ui' mà thay đổi tại mẫu sóng tiếp theo mà tại đó cửa hầu mở. Giá trị tần số cơ bản thấp có thể làm trễ sự thay đổi đến 10 ms (trung bình là 5 ms khi F0 là 100 Hz, 2,5 ms khi F0 là 200 Hz)

3. Số lượng formant trong tuyến âm nối tiếp (nf): Là số lượng formant tính từ F1 đến tối đa F8 thực sự có trong tuyến âm nối tiếp.

Giá trị ngầm định là 5 ứng với tần số lấy mẫu 10000 mẫu/giây và người nói có chiều dài tuyến âm là 17 cm (nghĩa là khoảng cách trung bình giữa các formant là 1000 Hz). Muốn mô hình tuyến âm có chiều dài khác 17 cm hoặc tần số lấy mẫu thay đổi thì cần phải thay đổi 'nf'. Ví dụ, để mô hình giọng nữ, thường tuyến âm ngắn hơn chiều dài trung bình tuyến âm của nam 20% thì 'nf' cần phải thiết lập là 4.

Nếu tần số lấy mẫu là 16000 mẫu/giây thì giọng nam cần phải có 8 formant trong khoảng từ 0 - 8000 Hz, như vậy 'nf' sẽ thiết đặt là 8. Chỉ 6 formant thấp có tần số và dải thông là được thiết lập bởi người dùng, formant thứ 7 và 8 có tần số và dải thông được cố định tại F7 = 6500, B7 = 500, F8 = 7500, B8 = 600. Tuyến âm song song chỉ có 6 formant, do vậy sẽ phải tăng F6 để phổ tiếng ồn với điểm tập trung trên giá trị ngầm định là F6 = 4990 Hz khi 'sr' tăng.

Tuy nhiên 'nf' chỉ xấp xỉ rất sơ bộ chiều dài tuyến âm. Nếu ví dụ người nói có chiều dài tuyến âm ngắn hơn 10% so bình thường, ta có thể chỉ sử dụng 5 formant trong nhánh liên tiếp, thiết lập các formant cao hơn thích hợp và sử dụng tham số nghiêng phổ TILT để đạt được sự phù hợp độ nghiêng phổ cho giọng nói này.

2.2.2. Các biến số

Có 40 biến số để điều khiển bộ tổng hợp, mỗi bộ giá trị của các biến số được gọi là một frame, mỗi bộ này sẽ điều khiển để tạo ra một đoạn theo thiết lập của hằng số 'ui'.

1. F0: Tần số cơ bản của mỗi giọng nói (pitch), ở đây giá trị này được sử dụng theo thang chia 0.1 Hz, nghĩa là 100Hz sẽ được biểu diễn bằng giá trị 1000.
2. AV: Biên độ của các âm hữu thanh của nhánh nối tiếp, đơn vị tính là dB. Khoảng giá trị từ 0-70, thường chọn 60 cho nguyên âm.
3. F1: Tần số formant (cực) thứ nhất, trong khoảng 200-1300 Hz.
4. B1: Băng thông của formant thứ nhất nhánh nối tiếp trong khoảng 40-1000 Hz.
5. F2: Tần số formant thứ hai, trong khoảng 550 - 3000 Hz.

6. B2: Băng thông của formant thứ hai nhánh nối tiếp trong khoảng 40-1000 Hz.
7. F3: Tần số formant thứ ba, trong khoảng 1200-4999 Hz.
8. B3: Băng thông của formant thứ ba nhánh nối tiếp trong khoảng 40-1000 Hz.
9. F4: Tần số formant thứ tư, trong khoảng 1200-4999 Hz.
10. B4: Băng thông của formant thứ tư nhánh nối tiếp trong khoảng 40-1000 Hz.
11. F5: Tần số formant thứ năm, trong khoảng 1200-4999 Hz.
12. B5: Băng thông của formant thứ năm nhánh nối tiếp trong khoảng 40-1000 Hz.
13. F6: Tần số formant thứ sáu, trong khoảng 1200-4999 Hz.
14. B6: Băng thông của formant thứ sáu nhánh nối tiếp trong khoảng 40-2000 Hz.
15. FNZ: Tần số của điểm không âm mũi trong khoảng 248-528 Hz (chỉ nhánh nối tiếp).
16. BNZ: Băng thông của điểm không âm mũi trong khoảng 40-1000 Hz (chỉ nhánh nối tiếp)
17. FNP: Tần số điểm cực âm mũi, trong khoảng 248-528 Hz
18. BNP: Băng thông của điểm cực âm mũi trong khoảng 40-1000 Hz
19. ASP: Biên độ âm bật hơi, trong khoảng 0-70 dB.
20. KOPEN: Hệ số mở của sóng âm, khoảng từ 0-60, thường là 30. Nó ảnh hưởng đến chất lượng của giọng nói như trầm khó nghe hoặc mềm mại nhẹ nhàng. Nó chỉ có tác dụng khi kích thích là xung hay mô phỏng tự nhiên còn với kích thích là sự lấy mẫu sóng âm thực thì hệ số này là cố định.
21. ATURB: Biên độ của độ ồn của giọng nói, trong khoảng từ 0-80 dB, thường sử dụng giá trị là 40 dB. Có thể dùng tham số này để mô phỏng chất lượng giọng khoẻ/yếu.
22. TILT: Độ nghiêng của phổ bằng dB, trong khoảng 0-24. Làm nghiêng phổ phát ra. Tăng giá trị này nhấn mạnh tần số thấp và nhẹ bớt tần số cao của tiếng nói.
23. AF: Biên độ âm xát, bằng dB, trong khoảng 0-80 (nhánh song song)

24. SKEW: Xiên phổ – chu kỳ thay đổi độ xiên, trong khoảng 0-40
25. A1: Biên độ formant thứ nhất của nhánh song song, trong khoảng 0-80 dB.
26. B1P: Băng thông của formant thứ nhất trong nhánh song song, bằng Hz.
27. A2: Biên độ formant thứ hai của nhánh song song.
28. B2P: Băng thông của formant thứ hai trong nhánh song song.
29. A3: Biên độ formant thứ ba của nhánh song song.
30. B3P: Băng thông của formant thứ ba trong nhánh song song.
31. A4: Biên độ formant thứ tư của nhánh song song
32. B4P: Băng thông của formant thứ tư trong nhánh song song.
33. A5: Biên độ formant thứ năm của nhánh song song.
34. B5P: Băng thông của formant thứ năm trong nhánh song song.
35. A6: Biên độ formant thứ sáu của nhánh song song.
36. B6P: Băng thông của formant thứ sáu trong nhánh song song.
37. ANP: Biên độ tần số cho âm mũi trong nhánh song song.
38. AB: Biên độ phần chuyển thẳng cho âm xát, bằng dB, từ 0-80.
39. AVP: Biên độ âm hữu thanh cho nhánh song song, trong khoảng 0-70 dB.
40. GAIN: Khuyếch đại chung, bằng dB, trong khoảng 0-80.

2.3. Tổng hợp tiếng Việt bằng mô hình tổng hợp formant

Để bộ tổng hợp formant trên "*nói*" được tiếng Việt, cần thiết lập các tham số tổng hợp tương thích với cách phát âm tiếng Việt và đặc trưng của hệ thống âm vị tiếng Việt.

Qua thực nghiệm xây dựng hệ Vnspeech nhận thấy, các hằng số cho toàn phiên làm việc thích hợp nhất được thiết lập như sau:

- Tần số lấy mẫu 'sr', có thể thay đổi nhưng chỉ cần 10000 Hz là đủ để nghe rõ các âm tiếng Việt.
- Khoảng cách cập nhật các tham số 'ui', thiết lập là 10 ms là đủ để mô tả sự thay đổi.
- Số bộ cộng hưởng trong nhánh nối tiếp, chọn 'nf' = 5 là đủ tốt

Các biến số được thiết lập từ các tham số đặc trưng của hệ thống âm vị tiếng Việt (được trình bày chi tiết trong phần ngữ âm tiếng Việt) và căn cứ vào cấu tạo âm tiết tiếng Việt. Số bộ các tham số phụ thuộc vào trường độ âm vị và giá trị 'ui'. Hiện tại, chỉ khoảng 20 tham số điều khiển tuyến âm là cần thay đổi giá trị theo từng frame, các tham số điều khiển nguồn âm và tán xạ cũng như tham số của các tần số formant cao được sử dụng giá trị ngầm định. Riêng tham số F0, được thiết lập tổng thể cho cả đoạn và từng âm tiết, do vậy, với mỗi frame, F0 sẽ nhận giá trị thích hợp để thể hiện đường nét chung.

Chương 3

MỘT SỐ KẾT QUẢ PHÂN TÍCH NGỮ ÂM TIẾNG VIỆT

Phần này giới thiệu một số quan điểm chung về ngữ âm tiếng Việt được được công nhận rộng rãi [Bảng02, Cẩn97, Chử00, Giáp01, Hạo98, Quỳnh01, Thuật99] và các kết quả của thu được của quá trình phân tích ngữ âm để xây dựng hệ TTS tiếng Việt. Phân tích ngữ âm, xác định các thông số đặc trưng các âm vị, các đặc điểm về ngữ âm, ngữ điệu của tiếng Việt là nội dung quan trọng, không thể thiếu khi xây dựng hệ tổng hợp tiếng Việt trên cơ sở formant cũng như các nghiên cứu khác về tiếng Việt.

3.1. Tiếng nói con người

Căn cứ vào cách cấu tạo âm của bộ máy phát âm, cách thoát ra của luồng không khí, các âm vị được phân thành 2 nhóm chính là nguyên âm (vowel) và phụ âm (consonant). Khi dây thanh dao động có chu kỳ, dòng khí được thoát ra ngoài tự do tạo thành nguyên âm. Ngược lại, luồng không khí từ phổi đi ra nếu bị cản trở tại một điểm nào đó như: đôi dây thanh đóng hoặc mở, khép chặt hai môi, tiếp xúc đầu lưỡi với lợi ... sẽ tạo nên các phụ âm. Ngoài hai loại âm vị chủ yếu trên còn có loại âm vị mang tính chất trung gian được gọi là bán nguyên âm hay bán phụ âm.
Sau đây là một số tiêu chí phân loại các nguyên âm tiếng Việt:

Theo vị trí của lưỡi:

- Nguyên âm dòng trước: khi phát âm các nguyên âm này, đầu lưỡi đưa về phía trước, ví dụ /i/, /e/, /ɛ/, /a/, /i̯e/
- Nguyên âm dòng giữa: khi phát âm các nguyên âm này, phần giữa của lưỡi nâng lên phía ngạc. Tiếng Việt không có nguyên âm dòng giữa.

- Nguyên âm dòng sau: khi phát âm các nguyên âm này, phần sau của lưỡi nâng lên phía ngạc mềm, ví dụ /u/, /o/, /ɔ/, /ɯ/, /ɤ/, /ɯ̯ɤ/, /u̯o/.

Theo độ mở của miệng:

- Nguyên âm có độ mở rộng: /a/
- Nguyên âm có độ mở hơi rộng: /ɛ/, /ɔ/
- Nguyên âm có độ mở hơi hẹp: /e/, /o/, /ɤ/, /i̯e/, /ɯ̯ɤ/, /u̯o/
- Nguyên âm có độ mở hẹp: /i/, /u/, /ɯ/

Theo hình dáng đôi môi:

- Nguyên âm tròn môi: /u/, /o/, /ɔ/, /u̯o/
- Nguyên âm không tròn môi:/i/, /e/, /i̯e/, /ɛ/, /a/, /ɯ/, /ɤ/, /ɯ̯ɤ/

Ngoài ra còn có một số tiêu chí khác như trường độ: nguyên âm dài hay ngắn, theo tính mũi hoá. IPA mô tả các nguyên âm theo một hình thang nguyên âm (Hình 3.2 dưới), trong hình này, 3 vạch đứng thể hiện 3 dòng nguyên âm (trước, giữa, sau); bên trái mỗi vạch là các nguyên âm không tròn môi, bên phải là các nguyên âm tròn môi; theo chiều từ trên xuống dưới độ mở của miệng rộng dần.

Miêu tả và phân loại các phụ âm: phụ âm thường được phân loại và miêu tả căn cứ vào hai tiêu chuẩn chính là phương thức cấu âm và vị trí cấu âm.

Phương thức cấu âm:

- Các âm bật: khi không khí đi ra ngoài bị cản trở hoàn toàn, phải phá vỡ sự cản trở để ra ngoài gây tiếng nổ nhẹ, ví dụ /p/, /t/, /k/
- Các âm xát: không khí không bị chặn hoàn toàn, phải đi qua một khe nhỏ giữa hai cơ quan cấu âm, gây nên tiếng xát nhẹ, ví dụ /v/, /f/, /s/.
- Các âm rung: lưỡi con hoặc đầu lưỡi chấn động liên tục, gây nên một loạt tiếng rung.

Theo vị trí cấu âm:

- Các âm môi: khi vật cản là hai môi gọi là âm môi-môi, môi dưới và răng gọi là môi-răng

- Các âm đầu lưỡi: khi đầu lưỡi quặt ngược chạm vào răng cửa hàm trên hoặc lợi, ngạc
- Các âm mặt lưỡi: mặt lưỡi được nâng lên phía ngạc cứng, ví dụ /c/, /ɲ/ trong *cha, nhà*
- Các âm cuối lưỡi, gốc lưỡi: phần cuối lưỡi được nâng lên tiếp xúc với ngạc mềm, ví dụ /g/, /k/, /ŋ/
- Các âm thanh hầu: không khí bị cản trở trong thanh hầu, ví dụ /h/ trong *hối hả.*

Các phụ âm được IPA mô tả trong một bảng tương quan cả hai tiêu chuẩn trên (Hình 3.2), các cột mô tả vị trí cấu âm, các hàng mô tả phương thức cấu âm.

Ngoài ra các phụ âm có thể được phân loại là vô thanh và hữu thanh, còn gọi là các âm vang và các âm ồn tuỳ theo thành phần cấu âm của chúng, thành phần tiếng thanh hay tiếng ồn là chính. Trong bảng trên, mỗi cột có hai âm thì âm bên trái là vô thanh, bên phải là hữu thanh.

Tuy nhiên, ngoài các tiêu chuẩn chính như trên, xu hướng phát âm cũng có tác dụng làm tạo ra các sắc thái mới cho âm vị. Chẳng hạn, một âm gốc lưỡi, khi phát âm nhích về phía trước gọi là ngạc hoá, ngược lại là mạc hoá; phát âm tròn môi gọi là môi hoá. Các vấn đề này nếu được mô hình rõ ràng sẽ rất thuận lợi cho quá trình điều chỉnh bộ tổng hợp để có thể tạo ra tín hiệu giống tiếng nói con người hơn.

3.2. Thông tin chung về ngữ âm tiếng Việt

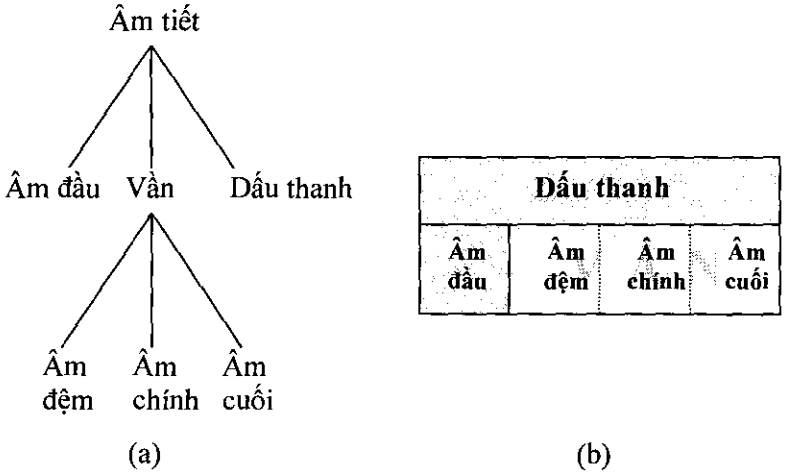
Tiếng Việt (Vietnamese) thuộc Ngữ hệ Phương nam, dòng Nam Á (Austroasiatique), ngành Môn-Khơ me [Chử00, Thuật99], là loại ngôn ngữ thanh điệu, sử dụng các ký hiệu La tinh để ghi chữ viết và các ký hiệu phụ để ghi dấu thanh. Tiếng Việt là ngôn ngữ đơn âm tiết, ranh giới âm tiết trùng với ranh giới hình vị, các thanh điệu là yếu tố ngữ điệu siêu đoạn trong phạm vi âm tiết và được các nhà ngôn ngữ học tiếng Việt coi như các âm vị, có chức năng khu biệt âm tiết. Có một số nhập nhằng trong cách ghi và đọc của các âm vị tiếng Việt, chẳng hạn: một

âm vị có thể có một số cách ghi và một cách ghi có thể biểu diễn nhiều hơn một âm vị.

Câu (sentence) tiếng Việt được chia thanh bốn loại, đó là: câu trần thuật (tường thuật); câu hỏi; câu cảm thán; và câu mệnh lệnh. Trên chữ viết, các loại câu được phân biệt bằng ký hiệu hết câu: dấu chấm cho câu trần thuật, dấu hỏi cho câu hỏi và dấu chấm than cho câu cảm thán và câu mệnh lệnh, còn trong tiếng nói, các loại câu được phân biệt bằng thay đổi ngữ điệu trong câu.

Từ (word) tiếng Việt gồm từ đơn và từ ghép, từ đơn chỉ gồm một âm tiết, từ ghép được cấu tạo từ 1 đến 4 âm tiết (phổ biến là 2). Trên chữ viết, các âm tiết được viết rời nhau (ngăn bằng dấu cách) và khi phát âm cũng có sự phân tách rõ rệt. Nói chung, tiếng Việt không có hiện tượng nối âm, luyến âm, nuốt âm khi phát âm.

Âm tiết (syllable) tiếng Việt được cấu tạo từ các âm vị. Âm tiết có cấu tạo thống nhất, mỗi âm tiết gồm 3 thành phần luôn có mặt và có thể dễ dàng bị phân tách là phụ âm đầu, vần và dấu thanh. Phần vần gồm có: âm đệm, âm chính và âm cuối, trong đó âm chính là nguyên âm, bắt buộc phải có mặt, âm đệm và/hoặc âm cuối có thể vắng mặt. Phụ âm đầu và vần là các thành phần đoạn tính theo thứ tự âm đầu-vần; dấu thanh là thành phần siêu đoạn (hình 3.1 a,b).



Hình 3.1. Cấu trúc một âm tiết tiếng Việt

3.3. Âm vị tiếng Việt

Tiếng Việt gồm 39 âm vị và 6 thanh điệu, trong đó có 23 phụ âm và 16 nguyên âm, là các vị trí được đánh dấu trong bảng chữ cái ngữ âm Quốc tế hình 3.2.

Hình 3.2. Bảng chữ cái ngữ âm Quốc tế

Âm vị tiếng Việt được chia thành 4 hệ thống khác nhau theo vị trí vai trò của nó trong cấu tạo âm tiết. Các phụ âm tiếng Việt chỉ có thể đứng ở đầu và/hoặc cuối âm tiết

3.3.1. Âm đầu

Tiếng Việt có 23 phụ âm làm nhiệm vụ âm đầu, mô tả âm vị, chữ viết và ví dụ theo bảng 3.1 sau:

Bảng 3.1. Hệ thống phụ âm đầu tiếng Việt

Âm vị	Chữ viết	Ví dụ
p	p	sa **p**a, **p**ác bó
m	m	**m**iệt **m**ài
b	b	**b**uôn **b**án
v	v	**v**ui **v**ẻ
f	ph	**ph**ương **ph**áp
t	t	**t**ương **t**ự
t'	th	**th**a **th**iết
d	đ	**đ**ây **đ**ó
n	n	**n**ước **n**on
s	x	**x**inh **x**ắn
ʂ	s	**s**ắp **s**ửa
l	l	**l**ập **l**oè
c	ch	**ch**âu **ch**ấu
ʈ	tr	**tr**ai **tr**áng
ɲ	nh	**nh**ập **nh**ằng
χ	kh	**kh**ó **kh**ăn
h	h	**h**ồi **h**ộp

z̬	r	**r**ắc **r**ối
z	d, gi	**d**a, **gi**a
k	k, q, c	**k**èn, **qu**á, **c**on
ɣ	gh, g	**gh**en, **g**ặp
ŋ	ngh, ng	**ngh**iên, **ng**ủ
ʔ		ai ơi, anh em

Các âm vị như /z/, /k/, /ɣ/, /ŋ/ có vài cách viết khác nhau; âm vị /ʔ/ là âm vị duy nhất không ghi bằng chữ viết, chính là âm vị xuất hiện các âm tiết mà ta không thấy âm đầu ghi bằng chữ cái nào.

3.3.2. Âm đệm

Âm đệm có tác dụng tròn môi khi phát âm, chỉ có 1 âm đóng vai trò âm đệm, có 2 dạng thể hiện khi viết như sau:

Bảng 3.2. Âm đệm tiếng Việt

Âm vị	Chữ viết	Ví dụ
w̬	o, u	**ho**a, kh**u**ẩn

Âm đệm có thể có mặt hoặc không có mặt trong âm tiết. Cấu tạo âm đệm giống nguyên âm /u/ làm âm chính nhưng khác về vị trí và chức năng nó đảm nhiệm trong âm tiết. Âm chính bao giờ cũng nằm ở đỉnh âm tiết, còn âm đệm nằm ở sườn đường cong đi lên. Trong phân tích và tổng hợp tiếng nói còn nhận thấy, âm đệm có độ dài rất ngắn so với âm chính.

3.3.3. Âm chính

Có 16 nguyên âm làm âm chính, gồm 9 nguyên âm thường, 4 nguyên âm ngắn và 3 nguyên âm đôi. Phiên âm, chữ viết và ví dụ các nguyên âm tiếng Việt

theo bảng sau:

Bảng 3.3. Hệ thống âm chính tiếng Việt

Âm vị	Chữ viết	Ví dụ
i	i, y	**m**i**nh**, **l**ý
e	ê	t**ết**
ɛ	e	**em**
ɛ̆	a	n**h**a**nh**, s**ách**
ɯ	ư	s**ư** t**ử**
ɤ	ơ	th**ơm**
ɤ̆	â	t**ất** b**ật**
a	a	ch**a**, ng**a**ng
ă	a, ă	đ**au** t**ay**, b**ắt**
u	u	th**u**
o	ô	c**ông** n**ông**
ɔ	o, oo	t**o**, b**oo**ng
ɔ̆	o	b**óc**, c**ong**
i‿e	iê, yê, ia, ya	v**iệt**, tuy**ên**, m**ía**, khu**ya**
ɯ‿ɤ	ươ, ưa	ch**ương**, l**ưa** th**ưa**
u‿o	uô, ua	l**uống** cu**ống**, m**ua**

Các âm chính bao giờ cũng phải có mặt trong một âm tiết, đóng vai trò đỉnh âm tiết. Âm chính có độ dài ngắn khác nhau tuỳ theo dấu thanh và các âm đầu, âm cuối có mặt trong âm tiết. Có một số mối quan hệ về việc có thể hoặc không thể đi cùng của âm chính với âm đầu và âm cuối, âm đệm, dấu thanh. Các quan hệ này góp phần tạo nên chính tả của tiếng Việt.

3.3.4. Âm cuối

Ngoài âm cuối zero (không có), tiếng Việt có 8 âm vị có thể làm âm cuối, biểu diễn âm vị, chữ viết và ví dụ theo bảng sau:

Bảng 3.4. Hệ thống âm cuối tiếng Việt

Âm vị	Chữ viết	Ví dụ
m	m	ha**m** là**m**
n	n	la**n** ma**n**
p	p	tấ**p** nậ**p**
t	t	lá**t** cắ**t**
k	c, ch	Cá**c** bá**c**, sạ**ch** bá**ch**
ŋ	ng, nh	nha**nh** chó**ng**
i̯	i, y	la**i** ra**i**, mâ**y** ba**y**
u̯	u, o	kê**u** cứ**u**, le**o** ca**o**

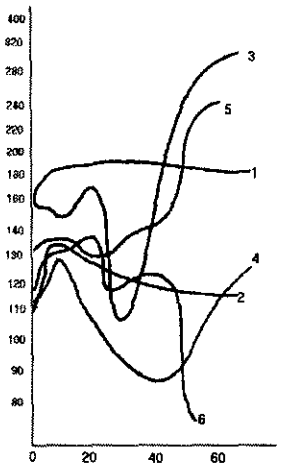
Căn cứ vào âm cuối, các âm tiết được chia thành 4 loại:

- Âm tiết nửa đóng: âm cuối là các phụ âm mũi /m, n, ŋ/
- Âm tiết đóng: âm cuối là các phụ âm vô thanh /p, t, k/
- Âm tiết nửa mở: âm cuối là các bán nguyên âm /u̯, i̯/
- Âm tiết mở: không có âm cuối, kết thúc bằng cách giữ nguyên âm sắc của nguyên âm chính.

3.3.5. Thanh điệu

Mỗi âm tiết tiếng Việt đều có một dấu thanh, tiếng Việt có 6 thanh điệu gồm không dấu, huyền, hỏi, ngã, sắc, nặng. Trong chữ viết là các ký hiệu (\?~/.) đặt trên hay dưới các âm chính, trong phiên âm là các số từ 1-5 viết nhỏ trên cao cuối âm tiết theo thứ tự trên (thanh không dấu được biểu diễn là không viết gì). Dấu thanh có

tác động siêu đoạn, ảnh hưởng đến toàn bộ âm tiết, mạnh nhất là phần vần. Sự đo và vẽ lại bằng máy kymograph về biến thiên tần số rung động của dây thanh (F0) ứng với các dấu thanh theo hình 2.3 (trục tung là tần số Hz, trục hoành là thời gian ms)



1- Không dấu
2- Huyền
3- Hỏi
4- Ngã
5- Sắc
6- Nặng

Hình 3.3. Biến thiên tần số rung động dây thanh với các thanh điệu khác nhau

Ta có thể hình dung một không gian 5 chiều, mỗi chiều là một thành phần của một âm tiết tiếng Việt, ý nghĩa thang chia chỉ các giá trị khác nhau của thành phần đó, nên thang chia này chỉ mô tả sự có mặt, không có ý nghĩa về giá trị lớn nhỏ. Các chiều của không gian này như sau:

- Chiều âm đầu: có 23 giá trị khác nhau, gồm các phụ âm và âm tắc thanh hầu (không có giá trị 0)
- Chiều âm đệm: có 2 giá trị khác nhau (không hoặc có)
- Chiều âm chính: có 16 giá trị khác nhau (không có giá trị 0)
- Chiều âm cuối: có 9 giá trị khác nhau
- Chiều thanh điệu: có 6 giá trị khác nhau (không có giá trị 0)

Như vậy, ta có thể xem mỗi âm tiết tiếng Việt như là một điểm trong không gian 5 chiều trên. Về lý thuyết tối đa sẽ có (23*2*16*9*6) = 39744 âm tiết có thể

trong tiếng Việt. Thực tế, các âm tiết tiếng Việt không lấp đầy không gian này, có rất nhiều tổ hợp không tồn tại (thống kê cho thấy, chỉ có khoảng 7000 âm tiết hay dùng trong tiếng Việt), ta có thể xây dựng bảng quan hệ các cặp giá trị với nhau để xác định có giá trị hợp lệ, chẳng hạn các âm cuối là /t/, /k/ thì thanh điệu chỉ có thể là "sắc", "nặng" ... Ta có thể ứng dụng sự mô tả này trong việc kiểm tra chính tả với đơn vị là âm tiết tiếng Việt mà không cần tra từ điển trong các ứng dụng như kiểm tra chính tả hoặc bộ điều khiển bàn phím đồng thời cho cả tiếng Việt và tiếng Anh, cũng có thể ứng dụng trong mã hoá và nén chữ Việt.

3.4. Kho ngữ liệu và công cụ nghiên cứu tiếng Việt

Dữ liệu ngôn ngữ, các công cụ phân tích tiếng nói và văn bản là không thể thiếu trong thực hiện đề tài, tuy nhiên, đối với tiếng Việt, các phương tiện này đều chưa sẵn sàng. Đề tài đã tự tiến hành xây dựng một kho ngữ liệu nhỏ để phục vụ cho các mục đích cụ thể, lựa chọn công cụ có sẵn hợp lý nhất và xây dựng công cụ phần mềm phục vụ cho nghiên cứu.

3.4.1. Kho ngữ liệu tiếng Việt

Kho ngữ liệu (Corpus) có vai trò đặc biệt quan trọng khi tiến hành nghiên cứu ngôn ngữ. Các hướng nghiên cứu khác nhau có các yêu cầu khác nhau đối với dữ liệu. Do vậy, yêu cầu đối với kho ngữ liệu là phải bao gồm tất cả các tình huống của ngôn ngữ, ngoài ra còn phải đủ nhỏ để khả thi khi xây dựng và dữ liệu tiếng nói phải được gán nhãn chi tiết, hệ thống phải cung cấp tính năng để có thể khai thác, tra cứu nhanh chóng, chính xác và đầy đủ theo các yêu cầu khác nhau. Chính vì điều này, xây dựng kho ngữ liệu đầy đủ là một công việc rất phức tạp.

Hiện chưa thấy tồn tại bất kỳ kho ngữ liệu tiếng Việt nào bao quát hết tất cả tình huống của ngôn ngữ như trên và trong phạm vi đề tài không thể đủ điều kiện xây dựng một kho ngữ liệu như vậy. Để đáp ứng nhu cầu cụ thể của đề tài, nghiên cứu được thực hiện trên một kho ngữ liệu nhỏ, bao gồm:

Khoảng 6000 âm tiết tiếng Việt được đọc rời rạc với tốc độ vừa phải trong điều kiện phòng làm việc, số hoá trực tiếp trên máy tính.

Khoảng 10000 ngữ đoạn có nghĩa, dài không quá 20 âm tiết, gồm các thể loại câu chính của tiếng Việt như câu trần thuật, câu hỏi, câu cảm thán và câu mệnh lệnh, được lọc từ các tác phẩm văn, thơ, tục ngữ phổ biến chứa khoảng 4000 âm tiết hay dùng nhất của tiếng Việt.

Công cụ phần mềm tự xây dựng để tập hợp các ngữ đoạn thỏa yêu cầu và lọc lấy các ngữ đoạn theo nhu cầu khảo sát từ tập hợp trên.

Dữ liệu này đã được sử dụng trong phân tích âm vị, liên cấu âm trong âm tiết, khảo sát các đặc tính ngữ điệu cũng như trong đánh giá chất lượng sản phẩm.

3.4.2. Công cụ phân tích tiếng nói

Công cụ được sử dụng để phân tích là phần mềm Wavesurfer [WaveSurfer] (có thể download tự do) và máy phân tích phổ tiếng nói CSL (Computerized Speech Lab) của hãng Kay Elemetrics.

Tiếng nói được thu và số hoá trực tiếp trên máy PC trong điều kiện phòng làm việc. Phương pháp phân tích bằng tổng hợp cũng được sử dụng thường xuyên để hiệu chỉnh các giá trị các tham số đặc trưng.

Ngoài ra, còn sử dụng một số chức năng phân tích và biểu diễn các thành phần đặc trưng của tiếng nói tự xây dựng kèm theo trong phần mềm Vnspeech.

3.5. Phân tích các tham số đặc trưng của âm vị tiếng Việt

Âm vị là các đơn vị trừu tượng để phân loại các thành phần nhỏ nhất tham gia vào cấu âm tiếng nói. Các âm vị tiếng Việt có thể được chia thành hai nhóm là nguyên âm và phụ âm căn cứ vào nguồn kích thích. Tuy nhiên, phương pháp phân tích sử dụng các đoạn tiếng nói tự nhiên chỉ là làm việc với các âm tố, là một thể hiện nào đó của âm vị. Các kết quả trình bày dưới đây là các giá trị trung bình của các âm tố của âm vị đó, được ghi nhận trong các ngữ cảnh khác nhau. Các tham số

này đóng vai trò giá trị khởi tạo cho bộ tổng hợp, và sau đó sẽ được hiệu chỉnh trong quá trình tổng hợp để xác định giá trị phù hợp.

3.5.1. Hệ thống nguyên âm tiếng Việt

Tiếng Việt có tổng cộng 16 nguyên âm, nhưng chỉ có 9 nguyên âm đơn là có các đặc trưng phân biệt, các nguyên âm khác có thể được tổng hợp từ đặc trưng của các nguyên âm đơn với sự thay đổi hay tổ hợp các tham số. Chẳng hạn, nguyên âm ngắn là sự tồn tại ngắn hơn khi cấu âm của nguyên âm đơn tương ứng, nguyên âm đôi là sự cấu âm đồng thời của hai nguyên âm đơn tương ứng. Do vậy, chỉ cần quan tâm nghiên cứu đặc trưng của 9 nguyên âm đơn.

Kết quả phân tích ngữ âm xác định các thành phần đặc trưng của tiếng nói bằng phần mềm và máy phân tích, sau đó sử dụng phương pháp phân tích bằng tổng hợp rút ra được bảng đặc trưng về tần số formant và độ rộng dải thông (bảng 2.1) của 9 nguyên âm đơn tiếng Việt, kết quả này hoàn toàn tương đương với [Bảng02, Loan, Loan99]. Mẫu giọng nói đưa phân tích là giọng nam trung (có F0 = 125 Hz),

Bảng 3.5. Các tham số đặc trưng của nguyên âm đơn tiếng Việt

Tham số / Âm vị	F1	B1	F2	B2	F3	B3
a	928	200	1491	150	2360	200
ɛ	614	200	1861	155	2545	214
e	426	150	1906	105	2560	184
ɔ	670	80	1050	105	2700	93
o	450	102	1000	104	2200	133
ɤ	467	112	1202	200	2614	120
u	350	121	722	201	2690	325
ɯ	328	96	1191	132	2536	114
i	310	57	2663	110	3073	191

3.5.2. Hệ thống phụ âm tiếng Việt

Nguyên lý của phương pháp tổng hợp formant là sử dụng các tần số formant để thiết lập các bộ lọc, các bộ lọc mô tả hình dáng của tuyến âm. Với các nguyên âm, do nguồn âm là tín hiệu tuần hoàn nên tín hiệu này được cộng hưởng khi đi qua tuyến âm, điều này được thể hiện rất rõ trên tiếng nói tạo ra. Trong khi đó, đối với các phụ âm, nhất là phụ âm vô thanh, điều này hoàn toàn ngược lại. Lý thuyết quỹ tích (locus theory) [Klatt87] được sử dụng để xác định sơ bộ các tần số đặc trưng của các phụ âm từ tiếng nói tự nhiên và sau đó sử dụng tổng hợp để hiệu chỉnh các giá trị.

Bảng 3.6. Bảng đặc trưng các phụ âm xát tiếng Việt

	F1	F2	F3	B1	B2	B3	AF	A4	A5	A6	AB
f	400	1420	2560	60	90	150	35	0	0	0	50
v	280	900	2250	60	90	150	40	33	15	15	50
s	320	1390	2530	200	80	200	40	0	0	52	0
z	280	1720	2560	60	90	150	55	36	15	15	0
ʂ	300	1840	2750	200	100	300	40	48	48	46	20
ʐ	490	1180	1600	60	90	150	15	0	0	0	0
χ	300	1200	3285	60	300	200	60	20	15	15	30
ɣ	200	1480	2620	60	90	150	40	0	0	0	0
h	490	1480	2500	60	90	150	40	22	0	0	40

Bảng 3.7. Bảng đặc trưng các phụ âm bật hơi tiếng Việt

	F1	F2	F3	B1	B2	B3	A4	A5	A6	ASP	AB
p	400	1100	2150	300	150	220	0	0	0	32	40
b	190	500	2500	60	90	150	15	15	15	40	0

t	190	1780	2680	60	90	150	0	0	0	30	0
t'	400	1780	2680	60	90	150	22	0	0	60	0
d	190	2200	2680	60	90	150	15	15	15	40	0
ʈ	190	1780	2680	60	90	150	30	30	30	30	0
c	850	2320	3000	60	90	150	30	30	30	30	0
k	300	1990	2850	250	160	330	43	45	45	20	0

Bảng 3.8. Bảng đặc trưng các phụ âm mũi tiếng Việt

	F1	F2	F3	B1	B2	B3	A1	A2	A3	FN	AN
m	200	900	2400	40	200	200	26	30	33	360	50
n	480	1500	2535	40	300	300	35	15	35	450	55
ɲ	250	2100	2200	42	290	100	50	20	35	450	55
ŋ	480	1780	2620	40	300	260	35	15	15	450	55

Bảng 3.9. Bảng các đặc trưng phụ âm vang bên tiếng Việt

	F1	F2	F3	B1	B2	B3	A1	A2	A3
l	400	1380	2700	60	90	150	36	26	26

Các giá trị đặc trưng của các âm vị tiếng Việt được sử dụng như các giá trị đích để tạo các bộ tham số điều khiển cho phần chuyển văn bản thành tham số điều khiển trong chương 4.

3.6. Liên cấu âm trong âm tiết tiếng Việt

Do đặc điểm cấu âm, trong các âm tiết tiếng Việt tồn tại 3 tình huống liên cấu âm sau:

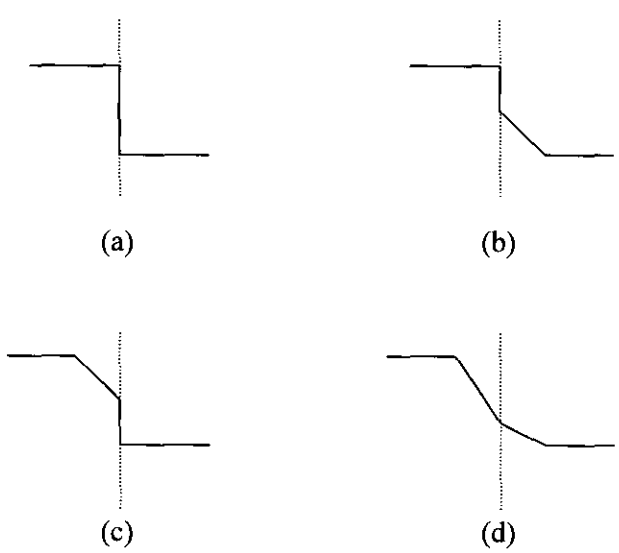
- Phụ âm – Nguyên âm (CV): ở các tình huống: phụ âm đầu - âm đệm, hay phụ âm đầu- âm chính
- Nguyên âm - Phụ âm (VC): xuất hiện ở tình huống âm chính – âm cuối
- Nguyên âm – Nguyên âm: (VV): ở tình huống âm đệm – âm chính, hay âm chính – âm cuối và cũng có thể coi như cả tại nguyên âm kép.

Mô tả liên cấu âm có nghĩa là mô tả sự ảnh hưởng lẫn nhau của các thông số đặc trưng của các âm vị kết nối, quan tâm nhiều nhất là các tần số formant, đặc biệt là các tần số F2, F3.

Khảo sát tiếng nói tự nhiên, biểu diễn hiệu ứng liên cấu âm các tham số đặc trưng là các đường cong, tuy nhiên, sự mô tả các tham số theo các đoạn được tuyến tính hoá không thấy xuất hiện sự cảm nhận đứt gãy tại thời điểm liên cấu âm trong tiếng nói tổng hợp. Dù sao, đây cũng có thể là một hướng cần quan tâm cho các bước nâng cao chất lượng.

Các phụ âm, nói chung không thể nào xác định tường minh các quỹ đạo formant của nó, sự thể hiện của các phụ âm khi cấu âm, phần quan trọng là sự thay đổi như thế nào các nguyên âm đi với nó (theo lý thuyết focus [Klatt87]).

Về nguyên tắc, sau khi tuyến tính hoá, ta có thể có 4 dạng thay đổi khi liên cấu âm như hình 3.4 sau:



Hình 3.4. Các dạng liên cấu âm

Dạng như hình 3.4 (a) là biến đổi đột ngột, trong khi (b), (c) là biến đổi bán đột ngột, (d) là sự biến đổi mềm mại. Hầu hết các tình huống cấu âm tiếng Việt đang được mô tả như tình huống hình 3.4 (d).

Giai đoạn chuyển dài hay ngắn và âm vị nào là âm vị gây ảnh hưởng phụ thuộc vào các cặp âm vị khi cấu âm cụ thể. Trong tiếng Việt, các phụ âm là âm vị gây ảnh hưởng đến nguyên âm, nhưng nhiều tình huống phụ âm đầu không ảnh hưởng nhiều đến âm chính.

Các thông tin để thể hiện liên cấu âm khi phát sinh các tham số điều khiển, các mô tả về độ ảnh hưởng lẫn nhau giữa các âm vị cũng như tình huống chuyển tiếp khi ghép nối đều được mô tả bằng các bảng dữ liệu xây dựng sẵn. Chi tiết về cấu trúc các bảng dữ liệu và phát sinh các tham số điều khiển được mô tả trong chương 4.

Chương 4

CHUYỂN VĂN BẢN THÀNH THAM SỐ ĐIỀU KHIỂN

Quá trình chuyển từ văn bản thành các tham số điều khiển bộ tổng hợp formant trong hệ Vnspeech gồm: phân tích và chuẩn hoá văn bản đầu vào, xác định các thông tin ngữ điệu và trên cơ sở đặc điểm ngữ âm tiếng Việt, các tham số đặc trưng của các âm vị cũng như cấu tạo của âm tiết phát sinh các tham số điều khiển bộ tổng hợp. Ngoại trừ bộ các tham số điều khiển được sinh để phù hợp với bộ tổng hợp formant, các nội dung như chuẩn hoá văn bản, xác định thông tin ngữ điệu là công việc chung cho mọi hệ TTS tiếng Việt.

4.1. Phân tích văn bản

Phân tích văn bản để chuyển các câu của văn bản cần đọc từ dạng biểu diễn thông thường theo các quy tắc viết chính tả tiếng Việt thành dạng biểu diễn duy nhất về ngữ âm, thuận tiện cho các bước xử lý tiếp theo.

4.1.1. Chuẩn hoá

Phần này trình bày phương pháp chuyển mọi dạng viết khác nhau của văn bản điện tử tiếng Việt thành một dạng duy nhất với các âm tiết thuần Việt, theo cách đọc tiếng Việt thông thường.

***Vấn đề gặp phải*:**

Văn bản điện tử tiếng Việt có một số vấn đề sau có thể dẫn đến nhầm lẫn khi xử lý, đọc tự động bằng máy:

- Văn bản được mã hoá bằng các bảng mã khác nhau: bảng mã 1 byte (phổ biến là theo bảng mã TCVN3, VNI); bảng mã 2 byte (phổ biến là Unicode dựng sẵn và Unicode tổ hợp). Các bảng mã khác nhau vì lý do kỹ thuật xử lý trên máy tính điện tử, còn đối với việc đọc của người, máy thì không có sự phân biệt.
- Trong văn bản tiếng Việt bình thường có một số dạng biểu diễn sau: Các âm tiết của hệ thống từ vựng tiếng Việt; các ký hiệu được phát âm như "+, -, %"; các ký hiệu không phát âm như các dấu ".,:?!" và các dấu ngoặc; các chữ số; các chữ viết tắt và các từ tiếng nước ngoài.

Vấn đề là cần phải có quy tắc thống nhất về đọc các tình huống trên.

Giải pháp của Vnspeech:

1. Vấn đề bảng mã khác nhau: tự động xác định bảng mã và chuyển sang biểu diễn bên trong theo bảng mã 1 byte TCVN3. Hiện tại Vnspeech có thể hiểu văn bản đầu vào theo bảng mã TCVN3 và Unicode dựng sẵn.
2. Xây dựng một bảng tham chiếu cách đọc theo kiểu tiếng Việt tất cả các ký tự (ASCII) có thể gặp.
3. Các chữ viết tắt: lập bảng tham chiếu các ký hiệu viết tắt phổ biến, nếu khi phân tích không gặp trong bảng thì đọc từng ký tự theo mục 2.
4. Các từ không có trong tiếng Việt: được chọn một hoặc một số giải pháp sau:
    - Sử dụng tham chiếu đến bộ tổng hợp tiếng nước ngoài (ví dụ tiếng Anh, hiện tại chưa thực hiện)
    - Đọc theo kiểu tiếng Việt: tách thành từng âm tiết theo kiểu tiếng Việt, ví dụ "love" đọc thành "lo ve" (hiện tại chưa thực hiện)
    - Đọc từng ký tự: hiện đang thực hiện theo giải pháp này
    - Bỏ qua không đọc
5. Các biểu thức số:
    - Đọc rời rạc các chữ số (cách đọc số điện thoại) như giải pháp 2, ví dụ: 8695484 – "tám sáu chín năm bốn tám bốn".

- Đọc theo dạng số đếm, đã giải quyết một số điểm lưu ý của cách đọc tiếng Việt:

  + Các số 0 hàng trăm trở lên đọc là "không trăm", hàng chục đọc là "linh", hàng đơn vị là "mươi", ví dụ: 2000033 – "hai triệu không trăm ba mươi ba"; 2134507- "hai triệu một trăm ba mươi tư ngàn năm trăm linh bảy"; 150- "một trăm năm mươi"

  + Số 1 đọc là "một" và "mốt", ví dụ: 11 - "mười một"; 31 - "ba mươi mốt"

  + Số 4 đọc là "bốn" và "tư", ví dụ: 14- mười bốn; 34- "ba mươi tư"

  + Số 5 đọc là "lăm","năm" và "nhăm", ví dụ: 5- "năm"; 15 – "mười lăm"; 25- "hai nhăm".

- Tự động xác định nếu là biểu thức chỉ ngày tháng, ví dụ: 25/12/2003 hay 25-12-2003 đọc là "ngày hai mươi nhăm tháng mười hai năm hai ngàn không trăm linh ba". Nếu không, đọc kiểu biểu thức số hoặc đọc từng chữ số.

4.1.2. Biểu diễn ngữ âm

Vấn đề: Mặc dù tiếng Việt sử dụng bảng ký hiệu La tinh để ghi âm, nhưng do quy định của quy tắc chính tả Tiếng Việt nên có hiện tượng sử dụng nhiều chữ cái khác nhau để cùng ghi một âm vị, sử dụng một ký tự để biểu diễn hơn một âm vị. Ngoài ra, cần phải phân biệt các ký tự nguyên âm như "o, u, i, y" là âm đệm, âm chính hay âm cuối, để xác định đúng âm vị nhiều trường hợp cần phải có ngữ cảnh. Do vậy, cần thiết phải chuyển thành dạng biểu diễn về ngữ âm duy nhất theo cấu trúc của âm tiết tiếng Việt, thuận tiện cho các bước xử lý tiếp theo. Ví dụ một số tình huống nhập nhằng:

Âm đầu:

'c', 'q', 'k': biểu diễn /k/

'gh', 'g': biểu diễn /ɣ/

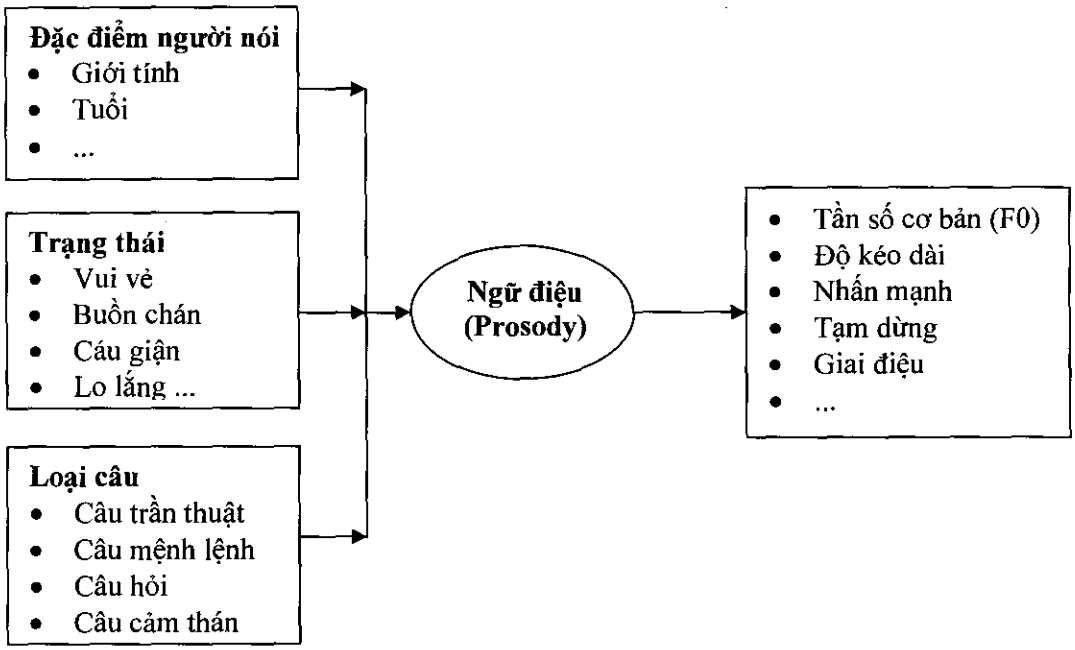
'ng', 'ngh'; biểu diễn /ŋ/

'd', 'gi', 'z': biểu diễn /z/, (âm 'i' có thể nhầm với âm chính, ví dụ: chữ gì)

Âm đệm: có thể nhầm lẫm với âm chính, ví dụ: mua – qua, quốc - thuốc; hò – hoà;
Âm chính: chữ 'a' có thể biểu diễn cho nguyên âm /a/ như 'mai', /ă/ như 'may', /ĕ/ như 'anh'; các nguyên âm đôi có nhiều cách viết khác nhau, như /i͡e/ được viết là ia (mía), yê (huyền), ya (khuya), iê (tiến)...
Âm cuối: có thể nhầm lẫn các bán nguyên âm với âm chính, ví dụ: 'đau - đu'
Giải pháp: Chuỗi ký tự thuần Việt được chuyển thành biểu diễn bằng phiên âm quốc tế theo quy ước của IPA, sử dụng thêm các con số "1, 2, 3, 4, 5" viết nhỏ cuối âm tiết để ký hiệu các dấu thanh 'huyền', 'hỏi', 'ngã', 'sắc', 'nặng' tương ứng (thanh không dấu không viết gì). Mỗi âm tiết tiếng Việt, trước hết được xác định dấu thanh (bỏ dấu khỏi âm tiết), sau đó theo bảng đối chiếu xây dựng sẵn gồm các thông tin: chữ viết, âm vị và ngữ cảnh (nếu cần thiết) , xác định các thành phần âm đầu, âm đệm, âm chính và âm cuối
Kết quả: Đoạn "biểu diễn phiên âm quốc tế văn bản tiếng Việt" sẽ là */bi͡ɐ̯u*^{2} *zi͡en*^{3} *fi͡en ʂm kʷ̣ok*^{4} *te*^{4} *văn ban*^{2} *ti͡eŋ*^{4} *vi͡et*^{5}*/.*

4.2. Phân tích xác định các thông tin ngữ điệu

Ngữ điệu (prosody) thể hiện cảm xúc của người khi đọc một đoạn văn bản, nó phụ thuộc vào nhiều yếu tố như: đặc điểm người nói, trạng thái, loại câu. Ngữ điệu được thể hiện ở tiếng nói đầu ra bằng các đặc điểm vật lý như: đường nét tần số cơ bản, độ kéo dài, độ nhấn mạnh... (hình 4.1)



Hình 4.1. Những thành phần ảnh hưởng và thể hiện của ngữ điệu

Một điều rõ ràng là tiếng nói nói chung (tự nhiên và tổng hợp) ngoài vấn đề phát âm rõ tiếng, ngữ điệu là yếu tố quan trọng đánh giá chất lượng tiếng nói. Nhiệm vụ của các hệ TTS là phải dự đoán được ngữ điệu từ văn bản, sau đó thể hiện trong tiếng nói tổng hợp. Do vậy, bộ tổng hợp (phần xử lý tín hiệu số) phải có khả năng thể hiện được các yếu tố ngôn điệu cần thiết, phần xử lý ngôn ngữ tự nhiên phải xác định được các thông tin ngữ điệu từ văn bản. Đây là một công việc hết sức khó khăn, là một thách thức lớn đối với tất các các hệ TTS trên thế giới [Keller02, Syndal, Taylor, Tuấn00a, VanSanten97]. Trên văn bản viết, có rất ít thông tin về ngữ điệu, chủ yếu do người đọc nhận thức thông qua hiểu nội dung hoặc theo thói quen. Công việc phân tích ngữ pháp để "hiểu" nội dung có thể là phương án tốt để xác định các thông tin ngôn điệu.

Ngữ điệu với các ngôn ngữ không thanh điệu thường được quan tâm khi tổng hợp ngữ đoạn (lớn hơn từ), với tiếng Việt, điều đó cũng là tương tự nhưng còn một vấn đề phải nhắc đến là một số các yếu tố thể hiện ngữ điệu trong phạm vi một âm

tiết là bắt buộc phải giải quyết vì đó chính là đặc điểm của thanh điệu của âm tiết. Do vậy, giải quyết ngữ điệu khi tổng hợp tiếng Việt phải quan tâm từ mức âm tiết.

Hiện tại, trong hệ Vnspeech, các thông tin ngữ điệu tại mức âm tiết đã được xác định và tổng hợp đạt kết quả tốt. Tại mức ngữ đoạn, đã tổng hợp được đặc tính trường độ cho câu trần thuật, để thể hiện được các đặc tính khác của ngữ điệu, cần phải có sự đầu tư nghiên cứu sâu rộng hơn.

4.2.1. Biến đổi cao độ trong âm tiết tiếng Việt

Mục tiêu của phân tích cao độ âm tiết là phục vụ cho nhiệm vụ tổng hợp dấu thanh. Dấu thanh tiếng Việt được xác định là sự biến đổi cao độ (tần số cơ bản F0) theo các quỹ đạo khác nhau. Các nghiên cứu chỉ ra rằng, đường nét của sự biến thiên là quan trọng chứ không phải giá trị tuyệt đối tại các điểm phân tích. Mỗi người nói có một tần số cơ bản đặc trưng, đó là giá trị cơ sở: giọng nam thường trong khoảng 100-150 Hz, nữ và trẻ em cao hơn, khoảng từ 150-250 Hz. Tổng hợp dấu thanh là tạo các giá trị mô tả đường nét đặc trưng của các thanh điệu dựa trên giá trị cơ sở. Tiếng Việt là ngôn ngữ thanh điệu (tone language) nên đường nét của sự biến đổi F0 (F0 contour) là yếu tố đóng vai trò quyết định trong việc thể hiện thanh điệu - một trong các thành phần cấu thành nội dung âm tiết tiếng Việt (các ngôn ngữ không thanh điệu, đường nét F0 diễn tả ngữ điệu chứ không quyết định nội dung). Xác định đường nét thể hiện F0 cần hai yếu tố: giá trị các điểm đánh dấu và độ kéo dài của đường.

Sự biến đổi cao độ trong các hệ TTS ở mức cơ bản cho các ngôn ngữ không thanh điệu thường không cần quá chi tiết, cao độ biến đổi một cách đơn giản để khỏi tạo cảm giác đều đều. Tiếng Việt là ngôn ngữ thanh điệu nên thông tin này rất quan trọng trong việc tổng hợp dấu thanh. Trên cơ sở cao độ đặc trưng cho từng giọng nói, biến đổi cao độ theo quỹ đạo đặc trưng và kết hợp với thông tin trường độ tạo các dấu thanh cho âm tiết tiếng Việt. Trong sáu dấu thanh của tiếng Việt, cao độ thanh không dấu là mức cơ bản đặc trưng cho mỗi giọng nói, tương tự như trong các ngôn ngữ không thanh điệu. Các dấu thanh còn lại đều có quỹ đạo biến đổi cao

độ đặc trưng, quỹ đạo của các dấu thanh này được vẽ trên cơ sở các thiết bị đo hay phân tích từ tiếng nói tự nhiên thường là các đường cong phức tạp, tuy nhiên trong mô tả đặc trưng, chúng thường được đơn giản hoá thành vài đoạn tuyến tính hoặc vài cách mô tả đơn giản khác. Hai cách mô tả phổ biến là bằng hình vẽ và bằng các con số [Giáp01]. Các biểu diễn bằng hình vẽ được trình bày chi tiết trong [Chử00, Giáp01, Minh02a, Quỳnh01, Thuật99]. Ví dụ của cách mô tả đơn giản hoá, tuyến tính hoá bằng con số: không dấu - 44; huyền - 32; hỏi - 323; ngã - 325; sắc - 45; nặng - 31.

Trong quá trình tổng hợp, chúng tôi tiến hành đơn giản và tuyến tính hoá quỹ đạo của biến đổi cao độ để mô tả dấu thanh như miêu tả bằng các con số như trên. Mỗi dấu thanh có một hàm để có thể dự đoán được cao độ tại một thời điểm với các thông tin đầu vào là cao độ cơ bản, độ kéo dài và vị trí cần dự đoán.

Phát sinh các giá trị diễn tả đường nét F0 ứng với một âm tiết:

Đường nét thanh điệu chỉ xác định trong đoạn gồm các âm vị hữu thanh của âm tiết (phần âm đầu hoặc cuối vô thanh không tồn tại tần số cơ bản). Sau khi phân tích để xác định các âm vị cấu thành âm tiết thì đồng thời xác định luôn điểm bắt đầu và kết thúc (và sẽ có giá trị trường độ) của đường nét thanh điệu. Nếu âm tiết có âm đầu và/hoặc âm cuối là hữu thanh thì đường nét thanh điệu thể hiện trên cả âm đầu và/hoặc âm cuối, ngược lại, nó chỉ thể hiện trên âm đệm (nếu có) và âm chính. Độ kéo dài của đường nét F0 liên quan đến trường độ các âm vị của âm tiết, xác định các giá trị này được trình bày trong phần sau.

Để thể hiện đường nét dấu thanh, mỗi âm tiết được xác định một giá trị F0 cơ sở (giá trị này phụ thuộc vào người nói và ngữ điệu của câu, có thể điều chỉnh được trong quá trình tổng hợp). Quá trình nghiên cứu ngữ âm tiếng Việt nhận thấy rằng, có thể tuyến tính hoá từng đoạn đường nét trên cơ sở giá trị cơ bản để diễn tả đường nét các thanh điệu. Các hàm tuyến tính (bậc nhất) này có dạng chung là:

$$y = a*(F_b/d)t + b*F_b$$

trong đó:

- F_b: tần số F0 cơ sở (Hz)

- d: miền xác định của hàm
- t: vị trí (thuộc miền xác định) cần xác định giá trị tuyệt đối của F0
- a, b: các hệ số, phụ thuộc vào các đoạn mô tả các thanh điệu; b chỉ vị trí xuất phát của đoạn so với F0 cơ sở; a - độ nghiêng của đường, a < 0: đường đi xuống và ngược lại.
- y: giá trị F0 tại thời điểm t

Các giá trị thực nghiệm của các biến số mô tả các dấu thanh của tiếng Việt được trình bày trong bảng 4.1 dưới.

Bảng 4.1. Các hệ số mô tả dấu thanh tiếng Việt

Stt	Thanh điệu	Số đoạn	Giá trị các hệ số a, b	Vị trí điểm gãy	Ghi chú
1	Không dấu	1	a = -0,1; b = 1	Không có điểm gãy	
2	Huyền	1	a = -0,1; b = 0,8	Không có điểm gãy	
3	Sắc	2	a1 = 0,1; b1 = 0,7 a2 = 0,55; b2 = 0,8	Tại ¼ đoạn hữu thanh	Âm tiết đóng, nguyên âm ngắn
			a1 = 0,2; b1 = 0,7 a2 = 0,4; b2 = 0,8	Tại 1/3 đoạn hữu thanh	Âm tiết đóng, nguyên âm thường
			a1 = 0,3; b1 = 0,7 a2 = 0,35; b2 = 0,8	Tại ½ đoạn hữu thanh	Âm tiết không đóng
4	Nặng	2	a1 = -0,3; b1 = 1 a2 = -0,35; b2 = 1	Tại vị trí giữa âm chính	
5	Hỏi	2	a1 = -0,35; b1 = 0,7 a2 = 0,45; b2 = 0,35	Tại vị trí giữa âm chính	

6	Ngã	2	a1 = -0,5; b1 = 0,8 a2 = 1,2; b2 = 0,3	Tại vị trí giữa âm chính	

4.2.2. Trường độ tự nhiên các âm vị

Trường độ các âm vị cấu thành âm tiết cũng đóng vai trò quan trọng trong thể hiện dấu thanh. Mỗi âm vị được giả thiết có một trường độ tự nhiên và giá trị này sẽ biến thiên trong các ngữ cảnh khác nhau [Minh03b]. Trường độ các âm vị cấu thành âm tiết được xác định bằng cách phân tích các âm tiết và thống kê bằng đơn vị mili giây (ms), chia làm 3 nhóm, đó là: giá trị ngắn nhất (min), giá trị dài nhất (max) và giá trị trung bình của tất cả các quan sát (avg). Giá trị trung bình được coi là trường độ tự nhiên của âm vị đó. Bảng 4.2 là các giá trị quan sát về trường độ của 23 âm đầu, 1 âm đệm và 8 âm cuối trong tình huống thanh điệu là "không dấu".

Bảng 4.2. Giá trị trường độ các âm vị (không kể âm chính) trong các âm tiết không dấu

Stt	Âm vị		min, max, avg
1	Âm đầu	p	20,25,24
2		b	72,166,100
3		t'	55,92,72
4		t	8,17,12
5		d	42,111,67
6		ʈ	39,100,56
7		c	38,65,54
8		k	18,36,24
9		ʔ	19,51,30
10		m	56,144,105
11		n	68,90,81
12		ɲ	42,83,65
13		ŋ	47,83,69
14		f	116,202,148
15		v	31,82,60
16		s	106,245,187
17		z	114,170,136
18		ʂ	105,206,146
19		ʐ	68,133,100
20		χ	81,142,103
21		ɣ	31,102,68
22		h	52,98,78
23		l	38,91,76
1	Âm đệm	w̥	30,46,33
1	Âm cuối	p	20,25,24
2		t	8,17,12
3		k	18,36,24
4		m	56,144,105

5		n	68,90,81
6		ŋ	47,83,69
7		i̯	159,211,180
8		u̯	115,164,146

Các số liệu quan sát được sử dụng để dự đoán tham số trường độ cho tổng hợp theo một trong các công thức sau:

DUR_s = rand(min, avg);

DUR_l = rand(avg, max);

DUR_n = (rand(min, max) + avg)/2;

trong đó:

DUR_s,DUR_l,DUR_n: các giá trị trường độ ở tình huống cần ngắn, dài và bình thường tương ứng

rand(x,y) là hàm lấy một giá trị ngẫu nhiên trong phạm vi [x,y]

Các công thức trên được sử dụng để tạo ra sự biến thiên của trường độ các âm vị cũng như âm tiết trong phạm vi hợp lý, tương tự như người đọc.

Trường độ âm chính

Âm chính (luôn là nguyên âm) về nguyên tắc có thể có trường độ không hạn chế (chẳng hạn như tình huống phát âm đối với âm tiết khi viết chỉ có riêng âm chính hoặc như ngân dài khi hát) nên việc xác định trường độ các loại âm chính khác nhau như đối với các âm đầu hoặc cuối rất khó thực hiện (và cũng không có ý nghĩa). Các thử nghiệm khi tổng hợp cho thấy rằng trong âm tiết, âm chính phải có trường độ hợp lý và đặc biệt còn biến thiên tùy thuộc vào dấu thanh. Do vậy, các âm chính được dự đoán ban đầu với giá trị trường độ tự nhiên, sau đó được thay đổi với các giá trị tăng hoặc giảm theo các quy tắc trong bảng 4.3.

Bảng 4.3. Các quy tắc thay đổi trường độ âm chính

Đặc điểm âm tiết		Trường độ âm chính
Loại dấu thanh	Không dấu	Giá trị tự nhiên
	Dấu huyền	Tăng 10%
	Dấu hỏi	Giảm 20%
	Dấu ngã	Tăng 10%
	Dấu sắc	Giảm 10%
	Dấu nặng	Giảm 20%
Loại âm cuối	Đóng	Giảm 5%
	Nửa đóng	Giảm 10%
	Nửa mở	Giảm 15%
	Mở	Tăng 30%

- Giá trị phần trăm tăng lên hay giảm đi tính theo giá trị tự nhiên.
- Trường độ nguyên âm ngắn được lấy bằng 2/3 của nguyên âm thường tương ứng
- Các giá trị kết quả đều được làm tròn thành số nguyên, giá trị nhỏ nhất không bé hơn 50 ms, lớn nhất không lớn hơn 400 ms
- Giá trị tự nhiên được gán bằng 227 ms (là giá trị trường độ trung bình của các âm chính của 1348 âm tiết không dấu tiếng Việt)
- Trường hợp âm chính dài ra thì âm cuối hữu thanh được chọn theo dạng ngắn.

4.2.3. Yếu tố thay đổi trường độ âm tiết

Khảo sát sự thay đổi trường độ âm tiết của tiếng nói tự nhiên nhận thấy sự thay đổi chủ yếu diễn ra ở âm chính và một phần ở âm cuối (nếu âm cuối là hữu thanh). Do vậy, trong tổng hợp, khi cần thay đổi trường độ thì chúng tôi chọn 80% giá trị cần thay đổi áp dụng cho âm chính và 20% cho âm cuối hữu thanh, còn âm cuối là vô thanh hoặc vắng mặt thì áp dụng cả 100% cho âm chính. Đánh giá bằng nghe cảm nhận, nhận thấy sự thiết lập này không tạo ra sự bất bình thường khi thay đổi trường độ âm tiết tổng hợp.

4.2.4. Trường độ các âm tiết trong ngữ đoạn

Khảo sát tiếng nói tự nhiên nhận thấy, cùng một âm tiết, trong các ngữ khác nhau, thậm chí cùng một ngữ đoạn tuỳ các tình huống sẽ có trường độ khác nhau. Phần này đưa ra một số quy tắc về trường độ của âm tiết trong ngữ đoạn trên cơ sở nghiên cứu tiếng nói tự nhiên sau khi kiểm nghiệm, hiệu chỉnh bằng hệ TTS Vnspeech.

4.2.4.1. Thay đổi trường độ do vị trí

Vị trí của âm tiết trong ngữ đoạn là một trong các yếu tố ảnh hưởng đến trường độ, giá trị trường độ các âm tiết ở các vị trí trong ngữ đoạn coi như là trường độ tự nhiên. Các công bố về luật này đối với tiếng Anh [Klatt87] cũng như tiếng Việt [Tuấn00c] là âm tiết cuối ngữ đoạn dài ra. Việc kiểm nghiệm lại trong một số tình huống cho thấy những kết quả cụ thể hơn: âm tiết khi ở cuối ngữ đoạn có trường độ tăng khoảng từ 20-50% và khi ở vị trí đầu ngữ đoạn tăng khoảng 5-10% so với trường độ tự nhiên.

4.2.4.2. Thay đổi trường độ do tốc độ đọc

Tốc độ đọc trung bình cho bộ tổng hợp được thiết đặt trước là 150 âm tiết/phút (được chủ ý làm chậm hơn người với mục đích tăng khả năng nghe rõ, tham số này có thể điều chỉnh được). Khi tăng tốc độ đọc thì trường độ các quãng nghỉ và trường độ các âm tiết đều được rút ngắn đi và ngược lại. Quy tắc đơn giản sau được sử dụng để điều khiển trường độ ứng với thay đổi tốc độ: trường độ các âm tiết được dự đoán cho tốc độ đọc bình thường và vị trí không đặc biệt, khi tăng tốc, giả sử 200 âm tiết/ phút, nghĩa là tốc độ đọc tăng thành 200/150 = 1,33, như vậy mỗi âm tiết sẽ phải giảm trường độ 33% và âm chính phải điều chỉnh trường độ là chủ yếu theo như nguyên tắc thay đổi trường độ âm tiết đã trình bày trong phần trên.

4.2.5. Trường độ các phần nghỉ

Khi đọc thành tiếng một đoạn văn bản tiếng Việt với một tốc độ đọc nào đó, ngoài phân bố thời gian cho phát âm thành tiếng còn được dành cho các quãng nghỉ. Mỗi chúng ta khi đọc thành tiếng đều nhớ và vận dụng quy tắc: "*dấu chấm nghỉ dài, dấu phảy nghỉ ngắn*". Ngoài yếu tố cần thay đổi trường độ các quãng nghỉ phù hợp với tốc độ đọc, trường độ quãng nghỉ còn thể hiện ngữ điệu và có thể dẫn đến hiểu một mệnh đề theo các nghĩa khác nhau. Độ kéo dài quãng nghỉ cho các dấu cách rất dễ gây nên sự hiểu sai nghĩa của cả mệnh đề, đó là tình huống người nghe nhầm lẫn giữa nghỉ dấu cách với nghỉ dấu phảy. Ví dụ, ngữ đoạn: "*mẹ con đi chợ chiều mới về*", tuỳ theo sự nghỉ dài hay ngắn sau âm tiết "*mẹ*" hoặc "*chợ*" sẽ tạo cho đoạn trên có các nội dung khác hẳn nhau.

Nghiên cứu các quãng nghỉ chỉ cần thiết khi làm việc với ngữ đoạn, có 3 tình huống dẫn đến nghỉ (không có tín hiệu) khi đọc một đoạn tiếng Việt, đó là:

1) gặp các dấu chỉ hết đoạn, hết câu như "chấm, phảy, hỏi chấm, chấm than...";

2) gặp các từ nối các đoạn như "và, là, có nghĩa là..." hoặc do người đọc cố tình nghỉ để nhấn mạnh, tạo sự chú ý;

3) do chuyển từ âm tiết này sang âm tiết khác, được biểu hiện trên chữ viết là dấu cách.

4.2.5.1. Nghỉ ứng với các dấu ngắt đoạn

Không có yêu cầu chính xác và đặc biệt nào cho các giá trị này và sự thay đổi giá trị của nó không ảnh hưởng đến nghĩa của câu do vậy áp dụng trong tổng hợp là trường độ cho dấu chấm tương đương với trường độ trung bình của âm tiết, trường độ dấu phảy bằng 1/2 dấu chấm. Khi tốc độ đọc bình thường (150 âm tiết/phút) thì dấu ".?!:;" nghỉ khoảng 400 ms (bằng 1 âm tiết), dấu "," khoảng 200 ms.

Trường độ các dấu nghỉ này sẽ được thay đổi theo tốc độ đọc, khi đọc chậm lại hoặc nhanh lên, trường độ các dấu nghỉ này sẽ tăng lên hay giảm xuống theo tỷ lệ tương ứng.

4.2.5.2. Nghỉ do chủ ý người đọc

Tình huống này cần phải phân tích ngữ pháp và phụ thuộc vào cảm nhận và ý muốn của người đọc. Trường hợp này đòi hỏi các khảo sát sâu hơn về phân tích ngữ pháp, ngữ nghĩa tiếng Việt. Đây là nội dung cần sự tiếp tục nghiên cứu trong tương lai.

4.2.5.3. Nghỉ ứng với các dấu cách

Đây chính là đặc điểm đơn âm tiết của tiếng Việt, giữa các âm tiết không phân biệt thuộc hai từ hay cùng một từ đều có một dấu cách. Về cấu âm, nó là thời gian cần cho sự chuyển đổi vị trí của các bộ phận trong bộ máy phát âm và nguồn âm. Sự thiết lập đơn giản một giá trị như nhau cho các dấu cách là nguyên nhân làm cho tiếng nói tổng hợp có cảm giác đều đều vô cảm.

Phân tích các ngữ đoạn của tiếng nói tự nhiên cho thấy rằng, trường độ đoạn nghỉ ứng với các dấu cách có giá trị khác nhau, mặc dù trong chữ viết được biểu diễn giống nhau (tình huống nhiều dấu cách liên tiếp có thể chỉ việc phải nghỉ lâu hơn nhưng cũng có thể chỉ là cách trình bày, ở đây tạm quan niệm theo ý thứ hai). Tuy nhiên, khi ta đọc thành tiếng nhiều lần cùng một đoạn văn bản với tốc độ và ngữ điệu như nhau (tương đối) thì các giá trị độ kéo dài kể trên vẫn luôn khác nhau. Do vậy, Vnspeech sử dụng khái niệm trường độ "tự nhiên" của quãng nghỉ ứng với khoảng trống giữa các âm tiết. Trường độ này được xác định bằng công thức sau:

Dur_Pause = (Dur_Total – Dur_Signal)/ Syl_Count

trong đó:

Dur_Pause: Trường độ tự nhiên của khoảng trống

Dur_Total: Tổng trường độ đoạn tiếng nói

Dur_Signal: Tổng trường độ các đoạn tín hiệu

Syl_Count: Số âm tiết trong đoạn

Các khảo sát sơ bộ từ tiếng nói tự nhiên cho phép kết luận rằng: trường độ của âm tiết liền kề hay vị trí quãng nghỉ trong ngữ đoạn (đầu, giữa hay cuối) không phải nguyên nhân tạo sự thay đổi trường độ tự nhiên. Độ kéo dài quãng nghỉ giữa các âm tiết phụ thuộc chủ yếu là cặp 2 âm vị kết thúc của âm tiết trước và bắt đầu của âm tiết sau và cặp dấu thanh của âm tiết trước-âm tiết sau.

Để nghiên cứu ảnh hưởng các âm tiết liền kề đến trường độ khoảng trống, các âm tiết tiếng Việt được chia thành 3 nhóm theo các yếu tố được chọn làm đặc điểm phân biệt, đó là: theo dấu thanh (bảng 4.4); theo âm vị kết thúc (bảng 4.5) và theo âm vị bắt đầu (bảng 4.6).

Bảng 4.4. Phân loại âm tiết tiếng Việt theo dấu thanh

Stt	Loại âm tiết	Đặc điểm	Ví dụ
1	Không dấu	Thanh điệu là không dấu	lanh
2	Có dấu huyền	Thanh điệu là dấu huyền	lành
3	Có dấu hỏi	Thanh điệu là dấu hỏi	lảnh
4	Có dấu ngã	Thanh điệu là dấu ngã	lãnh
5	Có dấu sắc	Thanh điệu là dấu sắc	lánh
6	Có dấu nặng	Thanh điệu là dấu nặng	lạnh

Bảng 4.5. Phân loại âm tiết tiếng Việt theo âm vị kết thúc

Stt	Loại âm tiết	Đặc điểm	Ví dụ
1	Âm tiết mở	Vắng mặt âm cuối	la
2	Âm tiết nửa mở	Âm cuối là các bán nguyên âm như: /u̯, i̯/	lao, lai
3	Âm tiết nửa đóng	Âm cuối là các âm mũi như: /m, n, ŋ/	lan, lam, lanh, lang
4	Âm tiết đóng	Âm cuối là các âm tắc vô thanh như:	lát, lác, láp

		/p, t, k/	

Bảng 4.6. Phân loại âm tiết tiếng Việt theo âm vị bắt đầu

Stt	Loại âm tiết	Đặc điểm	Ví dụ
1	Không có phụ âm đầu	Âm tiết có âm vị bắt đầu là âm tắc vô thanh cửa hầu /ʔ/	anh, em, ê, ô
2	Bắt đầu là âm tắc vô thanh	Âm tiết có các âm đầu là âm tắc vô thanh như /p, t, t', ʈ, c, k/	pa, tác, thanh, trúc, chương, các, kỷ, quân
3	Bắt đầu là âm vô thanh khác	Âm tiết bắt đầu là các âm vô thanh còn lại như /f, s, ʂ, x, h/	phương, xa, sức, khoẻ, hoà
4	Bắt đầu là các âm hữu thanh	Âm tiết bắt đầu là các âm tắc hữu thanh, xát hữu thanh, âm mũi và âm vang như /b, d, m, n, ɲ, ŋ, v, z, ʐ, ɣ, l/	bình, đà, muôn, năm, nhanh, ngày, vui, gieo, dưa, rung, ghế gỗ, lá

Để khảo sát quãng nghỉ ứng với khoảng trống, đối tượng được chọn là cặp các âm tiết. Bảng 4.7 liệt kê các luật thay đổi trường độ khoảng thời gian ứng với khoảng trống được rút ra từ việc phân tích tiếng nói tự nhiên.

Bảng 4.7. Luật thay đổi trường độ khoảng thời gian tự nhiên ứng với khoảng trống giữa các âm tiết của tiếng Việt trong ngữ đoạn

Stt	Ngữ cảnh	Khoảng nghỉ
1	Bên trái là âm tiết đóng	Dài ra
2	Bên phải là âm tiết có bắt đầu là âm tắc vô thanh	Dài ra
3	Bên trái là âm tiết có dấu nặng và bên phải là âm tiết có dấu sắc	Dài ra
4	Bên trái là âm tiết mở	Ngắn lại

5	Bên phải là âm tiết không có phụ âm đầu	Ngắn lại
6	Bên trái là âm tiết nửa mở, nửa đóng	Ngắn lại
7	Bên phải là âm tiết bắt đầu bằng âm hữu thanh	Ngắn lại
8	Âm tiết bên trái và bên phải thuộc cùng một từ	Ngắn lại

Trường độ tối thiểu của khoảng nghỉ là 0 ms, tối đa là giá trị bằng 1/2 giá trị thiết lập cho dấu "phảy". Thực tế, trong tiếng nói tự nhiên có hiện tượng cấu âm chồng lên nhau (chẳng hạn: bên trái là âm nửa đóng, bên phải là âm bắt đầu bằng âm mũi với cặp dấu thanh phù hợp) nhưng hiện nay chưa thể hiện hiệu ứng này trong tổng hợp tiếng nói.

Luật số 8 hiện chưa áp dụng trong dự đoán trường độ khoảng trống được vì cần phải phân tích được ngữ pháp của đoạn. Đây là một hướng cần nghiên cứu để tiếp tục nâng cao chất lượng.

4.3. Phân tích xác định các thông số đặc trưng

Phần này trình bày phương pháp mô tả hiệu ứng liên cấu âm giữa các âm vị trong âm tiết, tạo các tham số điều khiển cho một âm tiết, phù hợp với bộ tổng hợp formant tạo tín hiệu tiếng nói, từ bảng các thông số đặc trưng của các âm vị (các giá trị đích), cấu trúc của các âm tiết và các thông tin về cao độ và trường độ của âm vị, âm tiết tiếng Việt.

4.3.1. Mô tả các âm vị tiếng Việt

Mục tiêu của phần này là chuẩn bị dữ liệu về ngữ âm tiếng Việt để phục vụ cho việc sinh các tham số điều khiển bộ tổng hợp formant cho một âm tiết tiếng Việt.

Để mô tả được các âm tố và hiệu ứng liên cấu âm thể hiện âm tiết khi tổng hợp, các thông số đặc trưng mô tả các âm vị (các giá trị đích) được tổ chức thành bảng, mỗi phần tử mô tả toàn bộ hoặc một phần của mỗi âm vị tuỳ theo đặc điểm của nó, chẳng hạn, các nguyên âm đơn cần một bộ giá trị, các âm vị ghép, phụ âm bật cần vài bộ giá trị.

Một phần tử trong bảng giá trị được mô tả bằng các thành phần:

- Tên gợi nhớ
- Bậc của phần tử (Rk): để mô tả ảnh hưởng của phần tử này đến phần tử kề nó trong cấu âm.
- Nhóm của phần tử
- Mảng giá trị các thông số đặc trưng

Các thông số đặc trưng bao gồm:

- fn: tần số phản formant (mô tả đặc tính mũi)
- an: biên độ đặc tính mũi
- fi: các tần số formant
- bi: độ rộng dải thông các formant tương ứng
- ai: biên độ các formant tương ứng
- ab: biên độ phần được chuyển thẳng
- av: biên độ (mô tả đặc tính hữu thanh)
- asp: biên độ âm nổ
- avc: biên độ (mô tả mức độ nhẹ nhàng của giọng nói)
- af: biên độ đặc tính xát

Mỗi thông số được định nghĩa 5 giá trị:

- Giá trị đích (Steady)
- Phần góp cố định (Fixd)
- Phần góp phần trăm (Prop)
- Độ kéo dài ngoài (Ed)
- Độ kéo dài trong (Id)

4.3.2. Phát sinh các tham số điều khiển

Để tổng hợp một âm tiết, cần phải phát sinh được các bộ tham số thích hợp mô tả âm tiết đó từ bảng các giá trị đích của các âm vị cấu thành và căn cứ vào cấu tạo ngữ âm của âm tiết. Chúng tôi sử dụng giải thuật của Holmes-Mattingley-Shearme (HMS) [Cawley96, Mattingly74, Klatt87] để tạo các

tham số điều khiển cho bộ tổng hợp formant theo bảng các giá trị đích các âm vị. Các thông tin cần thiết khác (như F0, năng lượng...) được tính toán cho từng frame dựa vào quy luật biến đổi hoặc các giá trị ngầm định xác định bằng thống kê và hiệu chỉnh dần.

Để mô tả sự chuyển tiếp (liên cấu âm - coarticulartory) khi ghép 2 phần tử, giá trị biên được tính theo công thức sau:

Giá trị biên = Fixd âm trội + (Prop âm trội Steady âm kém / 100).*

Các giá trị trung gian giữa giá trị biên và giá trị đích được xác định bằng nội suy tuyến tính.

Trường độ phần chuyển tiếp được xác định theo nguyên lý sau: Độ kéo dài ngoài (Ed) xác định mức độ ảnh hưởng trong cấu âm của một phần tử đến phần tử kề nó. Độ kéo dài trong (Id) của một phần tử chỉ ra mức độ ảnh hưởng trong cấu âm của phần tử lân cận đến phần tử này. Ed và Id của phần tử trội (có Rk cao hơn) xác định trường độ phần chuyển tiếp giữa hai phần tử (phần kết thúc của phần tử trước và phần bắt đầu của phần tử sau). Có hai tình huống khi xác định Ed và Id cho 2 phần tử:

- Nếu phần tử sau có Rk cao hơn phần tử hiện tại thì độ kéo dài phần kết thúc của phần tử hiện tại là Ed của phần tử sau và độ kéo dài phần bắt đầu của phần tử sau là Id của nó.
- Nếu phần tử hiện tại có Rk cao hơn phần tử sau thì độ kéo dài phần kết thúc của phần tử hiện tại là Id của nó và độ kéo dài phần bắt đầu của phần tử sau là Ed của phần tử hiện tại.

Ví dụ: Xác định đoạn chuyển tiếp giữa hai âm vị A và B {Kết thúc (A) và Bắt đầu(B)}

- Nếu Rk(A) > Rk(B) thì Kết thúc (A)= Id (A) và Bắt đầu (B) = Ed (A)

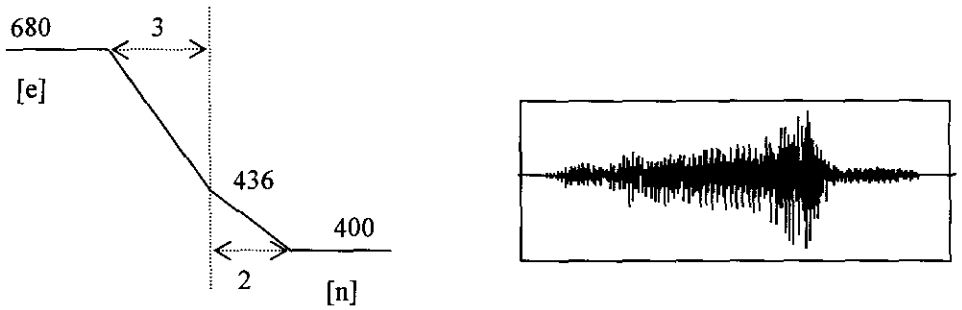
- Nếu Rk(A) < Rk(B) thì Kết thúc (A)= Ed (B) và Bắt đầu (B) = Id (B)

Chẳng hạn, ta có mô tả tham số F1 của:

Âm vị E: {680, 340, 50, 4, 4}, Rk(E) = 2 *Âm vị N:* {400, 300, 20, 3,2 }, Rk(N) = 10

cụ thể, trong âm tiết "***nguyễn***" [ŋwi̯en^3], thông số F1 tại chỗ ghép /e-n/ được tính như sau:

Giá trị biên(F1) = 300 + (20* 680/ 100) = 436



(a) Biến đổi F1 chỗ ghép /e-n/

(b) Phổ của âm tiết "***nguyễn***" được tạo bằng tổng hợp formant

Hình 4.2. Tạo các tham số điều khiển

Để sinh các tham số mô tả hiệu ứng liên cấu âm giữa các âm vị, các giá trị chuyển tiếp được tính như sau: tại giai đoạn chuyển bắt đầu (để nối với âm vị trước nó), các giá trị là nội suy tuyến tính giữa giá trị biên (được tính theo công thức trên) và giá trị đích; tại giai đoạn chuyển kết thúc (để nối với âm vị sau nó), các giá trị là nội suy tuyến tính giữa giá trị đích và giá trị biên sau (hình 4.2a). Nếu trường độ chuyển bắt đầu và chuyển kết thúc của một âm vị vượt quá trường độ của âm vị thì phần chuyển bắt đầu và kết thúc sẽ giao nhau, dẫn đến giá trị đích không được sử dụng, trong trường hợp này các giá trị được tính bằng nội suy tuyến tính giữa giá trị biên bắt đầu và giá trị biên kết thúc.

Hình 4.2b biểu diễn phổ tín hiệu tổng hợp được của âm tiết "*nguyễn*"' trong miền thời gian, sử dụng giải pháp sinh các tham số mô tả hiệu ứng liên cấu âm như trình bày trên, dấu ngã được tổng hợp bằng thay đổi đường nét F0 và trường độ âm tiết. Đánh giá bằng nghe không phát hiện điểm gãy chỗ ghép của các âm vị [ŋ,w̩,i̯e,n].

Phương pháp mô tả hiệu ứng liên cấu âm này còn được sử dụng để mô tả các âm vị ghép và các âm nổ (như [p,b,t,k,g]) tương đối hiệu quả, mỗi âm nổ

được chia thành 3 pha: pha giữ, pha thoát và pha sau thoát. Mỗi pha được mô tả như một phần tử trong bảng dữ liệu. Sự chuyển giữa các pha cũng áp dụng thuật toán trên. Tại điểm cần chuyển giá trị nhanh, ta gán giá trị Rk cao hơn để đảm bảo nó vượt qua các âm lân cận và đồng thời Id và Ed rất ngắn. Ví dụ pha thoát của các âm nổ được gán Rk cao hơn bất kỳ âm nào và Ed = Id = 0. Kết quả là sự chuyển vào và ra của âm nổ là không liên tục, điều này giống như đặc điểm tự nhiên của nó.

Thêm nữa, khi coi Rk, Id và Ed là các giá trị thay đổi, phương pháp này có thể mô tả được nhiều tình huống ghép nối các âm vị khác nhau. Các trường hợp thay đổi đột ngột, bán đột ngột hay dần dần; sự ảnh hưởng nhiều hay ít của âm vị này đến âm vị kia đều có thể diễn tả được. Điều này sẽ giúp tạo được tiếng nói tổng hợp đạt độ tự nhiên ngày càng cao theo đặc điểm tự nhiên của các tình huống cấu âm.

Chương 5

ĐÁNH GIÁ CHẤT LƯỢNG

Tất cả các thông số vật lý đặc trưng của tiếng nói đều cần phải quan tâm khi đánh giá chất lượng của tiếng nói tổng hợp, đó là các đánh giá sâu về kỹ thuật, còn mục tiêu cuối cùng vẫn là chất lượng của tín hiệu âm thanh tiếng nói được phát qua loa. Do vậy, thường có hai dạng đánh giá: so sánh các chỉ tiêu cần đánh giá của tiếng nói tổng hợp với tiếng nói tự nhiên và đánh giá bằng cảm nhận của người nghe theo các tiêu chí và mức độ khác nhau.

Tổng hợp tiếng nói formant là phương pháp trên cơ sở tín hiệu nên các thông số đặc trưng nhận được từ phân tích tín hiệu tiếng nói rất được quan tâm, tuy nhiên đây là các thông tin rất kỹ thuật nên thường chỉ dành cho các nhà nghiên cứu hay các chuyên gia tự đánh giá trong quá trình nghiên cứu phát triển bằng cách sử dụng các công cụ phân tích tiếng nói (thiết bị phần cứng hoặc phần mềm).

Đánh giá chất lượng tiếng nói dựa trên cảm nhận rất hay được nhắc đến vì đây là cách tự nhiên và không khó khăn khi lựa chọn người gia đánh giá. Tuy nhiên, hiện chưa có một công cụ hay phương pháp nào được đề xuất để đánh giá chất lượng tiếng Việt (tự nhiên và tổng hợp) bằng cảm nhận, mà thường chỉ là các nhận xét chung chung. Đề tài đã đề xuất và xây dựng bộ công cụ đánh giá chất lượng tiếng nói bằng cảm nhận, công cụ này có thể mở rộng để đánh giá chất lượng không chỉ riêng tiếng nói của hệ Vnspeech mà còn của cả các hệ TTS khác cũng như cho tiếng nói tự nhiên.

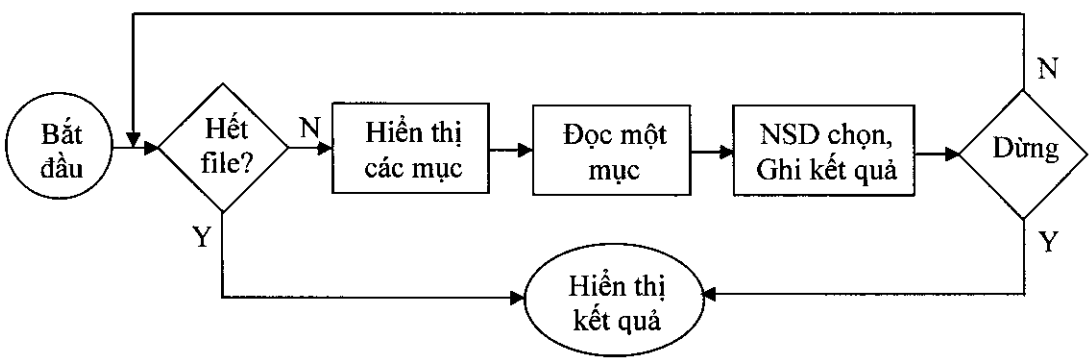
Bộ công cụ đánh giá chất lượng tiếng nói bằng cảm nhận được xây dựng thực hiện 4 dạng đánh giá: 1) đánh giá bằng lựa chọn; 2) đánh giá bằng nghe rõ dãy số và 3) đánh giá bằng nghe rõ câu bất kỳ; 4) đánh giá ngữ điệu.

5.1. Đánh giá sự phân biệt các thành phần bằng lựa chọn

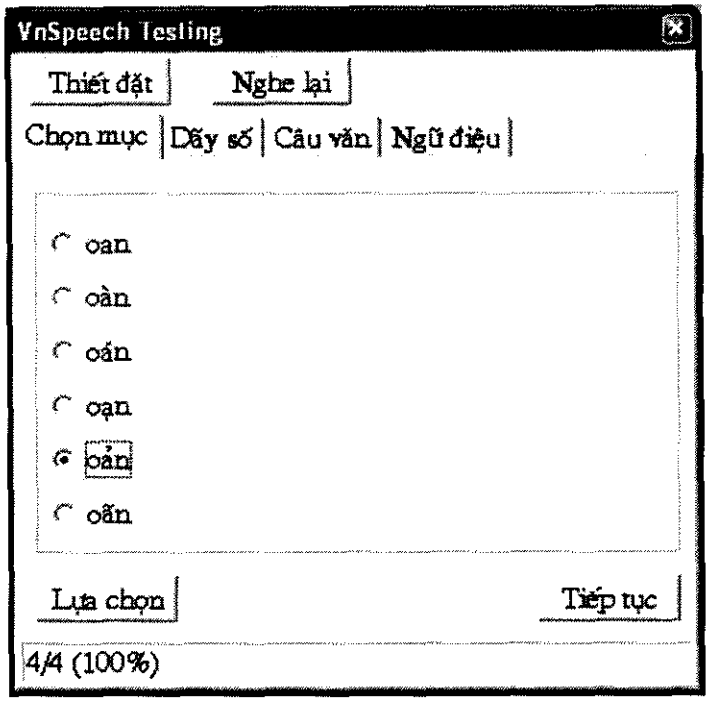
Mục tiêu của phương pháp này là đánh giá mức độ nghe rõ để phân biệt các âm vị cấu thành âm tiết, âm tiết, từ và đoạn. Các thành phần và đơn vị đưa ra để phân biệt gồm:

- Phân biệt các âm đầu
- Phân biệt âm đệm
- Phân biệt âm chính
- Phân biệt âm cuối
- Phân biệt dấu thanh
- Phân biệt âm tiết

Sơ đồ khối quy trình đánh giá bằng lựa chọn được trình bày trong hình 5.1(a).



(a) Sơ đồ khối quy trình



(b) Giao diện chương trình

Hình 5.1. Đánh giá kết quả bằng lựa chọn

Dữ liệu cho các chỉ tiêu cần đánh giá được chuẩn bị trước, ghi trong các file văn bản, mỗi chỉ tiêu đặt trong 1 file, được tổ chức như sau: các tình huống lựa chọn mỗi chỉ tiêu cần đánh giá được đặt trong 1 dòng, ngăn cách bằng dấu chấm phảy. Ví dụ, để phân biệt dấu thanh, file dữ liệu gồm các dòng dạng "a, à, á, ả, ã, ạ".

Quy trình đánh giá được thực hiện: lần lượt các dòng dữ liệu như trên được đọc vào bộ nhớ và trình bày lên màn hình thành dãy với mỗi mục có một nút chọn bên trái. Sau đó, engine Vnspeech sẽ đọc ngẫu nhiên một mục nào đó của dãy được hiển thị trên. Nhiệm vụ của người đánh giá là "chọn" mục nào giống nhất với âm thanh nghe được. Khi chọn nút "Tiếp tục", chương trình sẽ ghi nhận kết quả là đúng hay sai và lặp lại quá trình này cho đến hết toàn bộ dữ liệu. Khi hết toàn bộ dữ liệu hoặc người dùng chọn "kết thúc", chương trình sẽ thông báo kết quả đánh giá. Giao diện chương trình đánh giá được trình bày trong hình 5.1 (b).

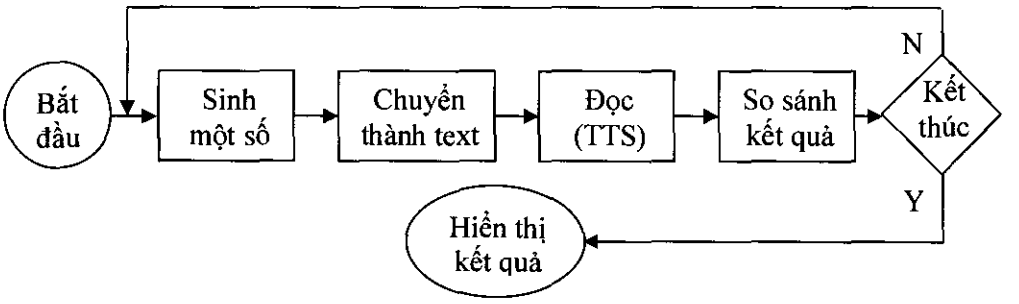
Người quản trị sẽ ghi nhận kết quả của từng người tham gia đánh giá và kết quả đánh giá sẽ là trung bình cộng kết quả của từng người. Số bộ test và số người tham gia phải đủ lớn để kết quả đánh giá được đặc trưng. File dữ liệu phục vụ test có thể thay đổi độc lập với chương trình để có thể nâng cao chất lượng quá trình test. Mỗi bộ test có thể gồm số mục chọn khác nhau (từ 2 – 10 mục chọn).

5.2. Đánh giá độ nghe rõ dãy số nguyên

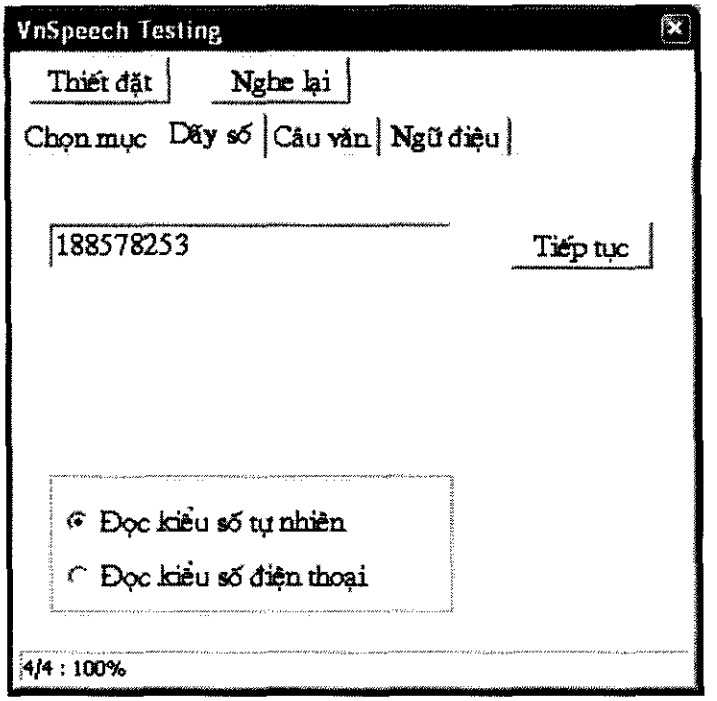
Mục tiêu của cách đánh giá này là kiểm tra sự nghe được chính xác một số không nhiều âm tiết biết trước với thứ tự không dự đoán được. Các âm tiết ở đây sẽ chỉ gồm các chữ số như "một", "hai", "ba",...và các bội số như "trăm","ngàn", "triệu", "tỷ".

Quy trình đánh giá được tiến hành như sau: chương trình sẽ sinh một số nguyên ngẫu nhiên nào đó và đọc theo cách đọc số tự nhiên tiếng Việt hoặc đọc kiểu số điện thoại (theo lựa chọn của người dùng), người kiểm tra sẽ nhập số nghe được để chương trình so sánh và chu trình kiểm tra sẽ được lặp lại đến khi người sử dụng quyết định dừng. Kết thúc chu trình, chương trình sẽ thông báo kết quả cuối cùng bằng tỷ lệ theo phần trăm số lần nhập vào đúng.

Sơ đồ khối quy trình đánh giá đo bằng độ nghe rõ một số ngẫu nhiên như hình 5.2 (a) và giao diện chương trình thực hiện trong hình 5.2 (b).



(a) Sơ đồ khối quy trình



(b) Giao diện chương trình

Hình 5.2. Đánh giá độ nghe rõ số nguyên ngẫu nhiên

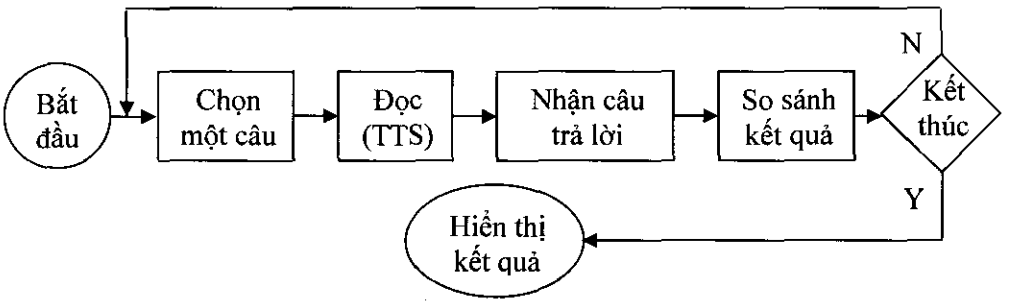
5.3. Đánh giá độ nghe rõ câu có nghĩa bất kỳ

Mục tiêu của các phương pháp này là đánh giá sự nghe rõ các âm tiết trong các câu có nghĩa (một số tình huống có thể đoán trước được âm tiết theo sau).

Dữ liệu được sử dụng để đánh giá là một file văn bản khoảng 10000 câu có nghĩa, có độ dài từ 5 đến 20 âm tiết được lọc từ các tác phẩm văn thơ, các câu thành ngữ của tiếng Việt như "Truyện Kiều", "Dế mèn phiêu lưu ký"..., mỗi câu đặt trên 1 dòng. Đây là mức độ đánh giá có độ khó và đặc trưng nhất vì có mặt gần như đầy đủ tất cả các âm tiết hay dùng của tiếng Việt (>5000) và rất nhiều các tình huống sử dụng và đủ các loại câu khác nhau.

Một chu trình đánh giá được tiến hành như sau: chương trình chọn và đọc một câu bất kỳ từ kho dữ liệu các câu văn trên, người dùng nhập câu nghe được vào ô nhập văn bản. Chương trình sẽ so sánh câu trả lời với câu được đọc theo mức độ chính xác từng âm tiết và ghi nhận kết quả. Sơ đồ khối của

quy trình đánh giá được trình bày trong hình 5.3 (a), giao diện chương trình trong hình 5.3 (b).



(a) Sơ đồ khối quy trình

(b) Giao diện chương trình

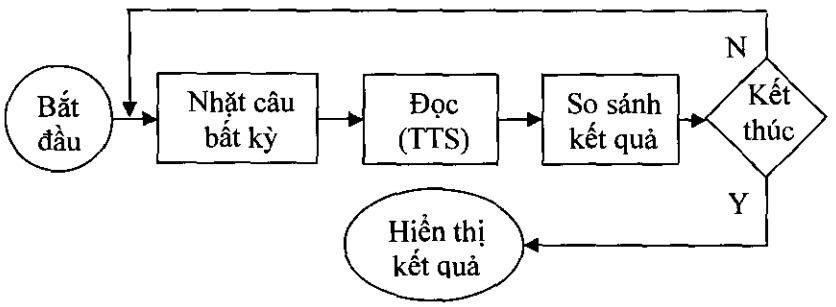
Hình 5.3. Đánh giá độ nghe rõ câu văn tiếng Việt

Chú ý: số lượng lần thử phải đủ lớn để kết quả được đặc trưng (khoảng 100 lần); nên nhập đủ số lượng âm tiết nghe được, kể cả âm tiết nghe không rõ.

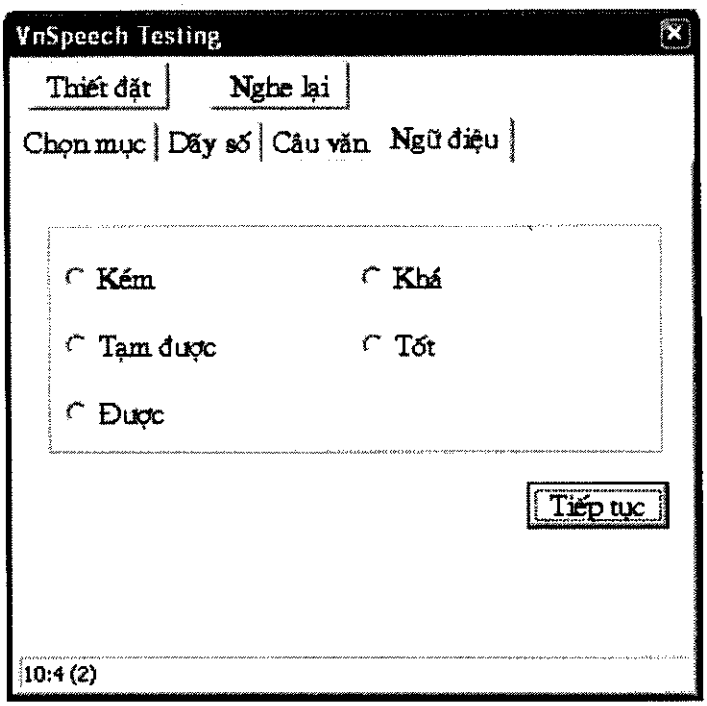
5.4. Đánh giá chất lượng ngữ điệu

Ngữ điệu ở đây được đánh giá dựa vào cảm giác của người nghe, người nghe cảm thấy câu vừa được đọc có độ ngân nga, lên xuống giọng, mạnh yếu có hợp lý hay không, sau đó quyết định đánh giá vào một trong 5 mức sau: "tốt", "khá", "được", "tạm được" và "kém", với các điểm 5, 4, 3, 2 và 1 tương ứng.

Quy trình đánh giá ngữ điệu được thực hiện như sau: chương trình đọc một câu tiếng Việt bất kỳ được chọn như phần 5.3; người dùng nghe và chọn vào mục chỉ ngữ điệu cảm thấy hợp lý nhất. Điểm số ngữ điệu được cộng dồn và lấy trung bình trên tổng số câu được đánh giá, kết quả lấy gần nhất với một trong 5 giá trị trên để đưa ra kết luận về ngữ điệu. Sơ đồ khối của quy trình đánh giá trong hình 5.4 (a) và giao diện chương trình thực hiện trong hình 5.4 (b).



(a) Sơ đồ khối quy trình



(b) Giao diện chương trình

Hình 5.4. Đánh giá ngữ điệu tiếng Việt tổng hợp

5.5. Kết luận

Cả bốn dạng đánh giá chất lượng trên đều có thể thay đổi một số các tham số đặc trưng của tiếng nói như tốc độ, tần số cơ bản, tần số lấy mẫu, âm lượng...sao cho thuận tiện nhất với khả năng nghe của mỗi người. Một điều hiển nhiên là sự quen với giọng nói, tốc độ đọc chậm, được phép nghe lại sẽ làm tăng sự đúng đắn của quá trình nhận diện. Bộ công cụ đánh giá này cho phép thực hiện các điều chỉnh và lựa chọn đó.

Để kết quả đánh giá được đặc trưng, số lần lặp lại của mỗi kiểu phải đủ lớn cũng như cần nhiều người khác nhau tham gia đánh giá và lấy kết quả trung bình. Mục tiêu thiết kế bộ công cụ đánh giá này là để không ai (kể cả người thiết kế) có thể biết trước đáp án đúng, sự đánh giá kết quả là do chương trình, tuy nhiên, chất lượng và sự tin cậy của kết quả đánh giá phụ thuộc vào dữ liệu. Dữ liệu được xây dựng để sử dụng đánh giá ở đây mới chỉ là bước đầu, cần phải tiếp tục hoàn thiện hơn.

Chương 6

SẢN PHẨM VÀ KẾT LUẬN

Đề tài được thực hiện với 3 mục tiêu chính: 1) Xây dựng hệ phần mềm có khả năng tự động đọc tất cả các từ tiếng Việt dễ nghe, gần với giọng người trên cơ sở tổng hợp tiếng nói formant từ các thành phần đặc trưng; 2) Đặt cơ sở cho các công việc xử lý tiếng nói khác; và 3) Nghiên cứu ngữ âm tiếng Việt. Sản phẩm cụ thể gồm bản báo cáo các kết quả nghiên cứu khoa học và phần mềm thực hiện tự động đọc văn bản tiếng Việt. Nội dung công việc được chia nhỏ thành 12 chuyên đề để thực hiện các mục tiêu và có kết quả trên. Có nhiều nội dung ngoài phạm vi của các chuyên đề đã được triển khai để thực hiện tốt mục tiêu, các nội dung nghiên cứu về lý thuyết đã được trình bày trong các chương trước của báo cáo này, các công việc về lập trình đã được tích hợp trong hệ phần mềm Vnspeech, với các tính năng sẽ được giới thiệu rõ hơn trong phần này.

6.1. Sản phẩm của đề tài

Đề tài đã xây dựng được hệ phần mềm "chuyển văn bản thành tiếng nói - VnSpeech" dựa trên tổng hợp tiếng nói formant. Phần mềm có thể được sử dụng như một engine tích hợp vào các ứng dụng khác hay phần mềm ứng dụng độc lập. Vnspeech có thể đọc thành tiếng được tất cả các âm tiết tiếng Việt khá dễ nghe, có thể nhận đầu vào là từng câu, phân tích để dự đoán đặc tính ngữ điệu và sau đó đọc ra loa hay ghi ra file nội dung tiếng nói ứng với câu đó.

Như một engine, Vnspeech là thư viện phần mềm thực hiện chuyển văn bản, nội dung tiếng Việt thành tiếng nói (TTS), gồm các hàm API:

- Phân tích và chuẩn hoá văn bản tiếng Việt
- Phân tích và biểu diễn các đặc tính ngôn điệu của câu tiếng Việt
- Chuyển thành biểu diễn ngữ âm của câu tiếng Việt

- Tổng hợp tín hiệu tiếng nói tiếng Việt bằng phương pháp ***tổng hợp formant dựa trên mô hình và nguyên lý tạo tiếng nói con người*** từ các thông tin ngữ âm
- Ghi ra file, phát ra loa tín hiệu tiếng nói tổng hợp

Đặc điểm:

- Các tham số đặc trưng của tiếng nói được điều khiển rất mềm dẻo, có thể biến đổi thành nhiều giọng khác nhau
- Kích thước toàn bộ chương trình rất nhỏ, thuận tiện cho tích hợp các thiết bị có tài nguyên hạn chế (cầm tay, di động...)
- Mã nguồn được viết bằng C++ nên tính khả chuyển cao

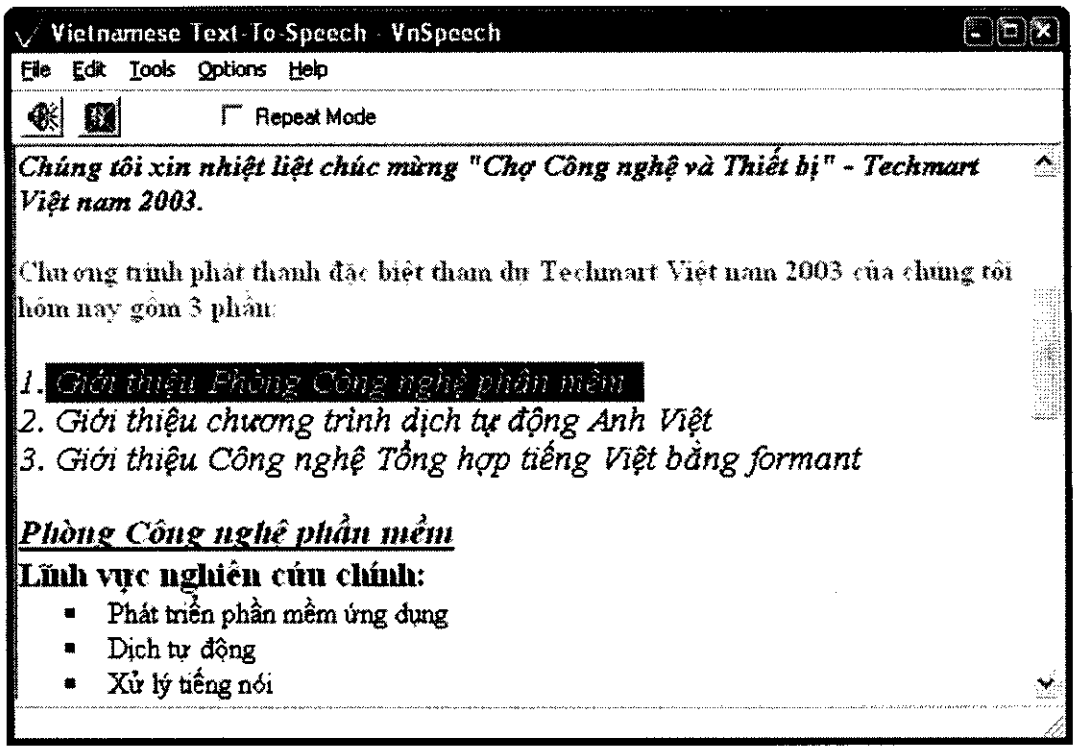
Thông số kỹ thuật:

- Tiếng Việt được tổng hợp bằng phương pháp ***tổng hợp formant***
- Xử lý văn bản tiếng Việt thuộc các bảng mã TCVN3, Unicode
- Tín hiệu tiếng nói tạo ra được mã hoá theo PCM 16 bit, mono (có thể thay đổi tần số lấy mẫu)
- Tổ chức thành các hàm API trong các thư viện DLL/COM, có thể sử dụng với mọi ngôn ngữ lập trình
- Chạy trên môi trường Windows 95/98/ME/NT/2000/XP/Server2003

6.1.1. Phần mềm ứng dụng

Phần mềm ứng dụng được xây dựng trên cơ sở engine Vnspeech bao gồm một số tính năng sau:

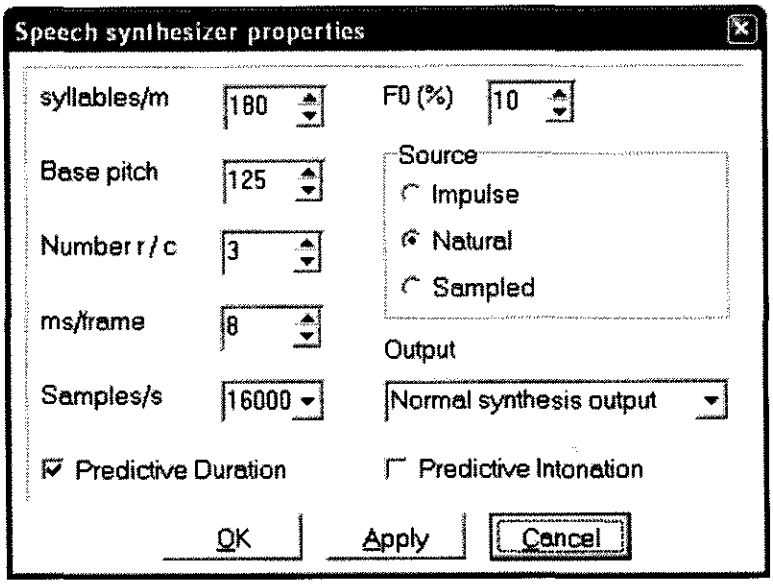
Môi trường soạn thảo văn bản tiếng Việt (hình 6.1) với một số tính năng chuẩn, có thể trực tiếp soạn thảo, mở hoặc lưu trữ các file dạng *.txt, *.rtf chứa tiếng Việt được mã hoá theo chuẩn TCVN3 hay Unicode dựng sẵn. Công cụ chuyển mã văn bản từ TCVN3 sang Unicode và ngược lại.



Hình 6.1. Giao diện chính của ứng dụng Vnspeech

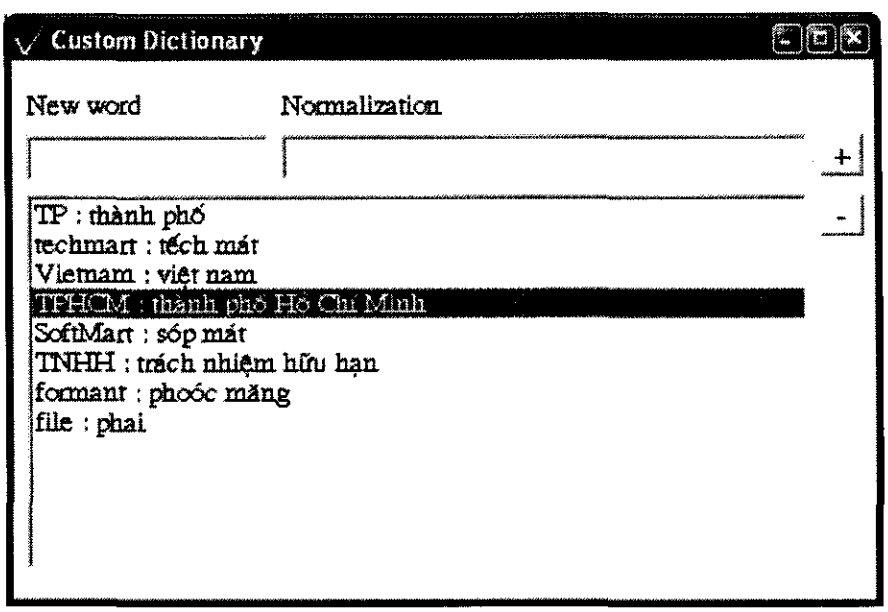
Để đọc, chỉ cần chọn hoặc đặt con trỏ về đầu đoạn văn bản cần đọc và chọn nút "đọc", tiến trình đọc có thể được ngừng bất kỳ lúc nào, câu đang xử lý được bôi đen để thuận tiện theo dõi, có thể chọn để quá trình đọc lặp lại nhiều lần...

Người dùng cuối có thể tuỳ chọn cho thay đổi các tham số đặc trưng của tiếng nói như cao độ cơ bản, tốc độ đọc, tần số lấy mẫu... (hình 6.2)



Hình 6.2. Bảng điều khiển các tham số đặc trưng

Từ điển cách đọc các từ viết tắt hay từ không có trong tiếng Việt (hình 6.3)

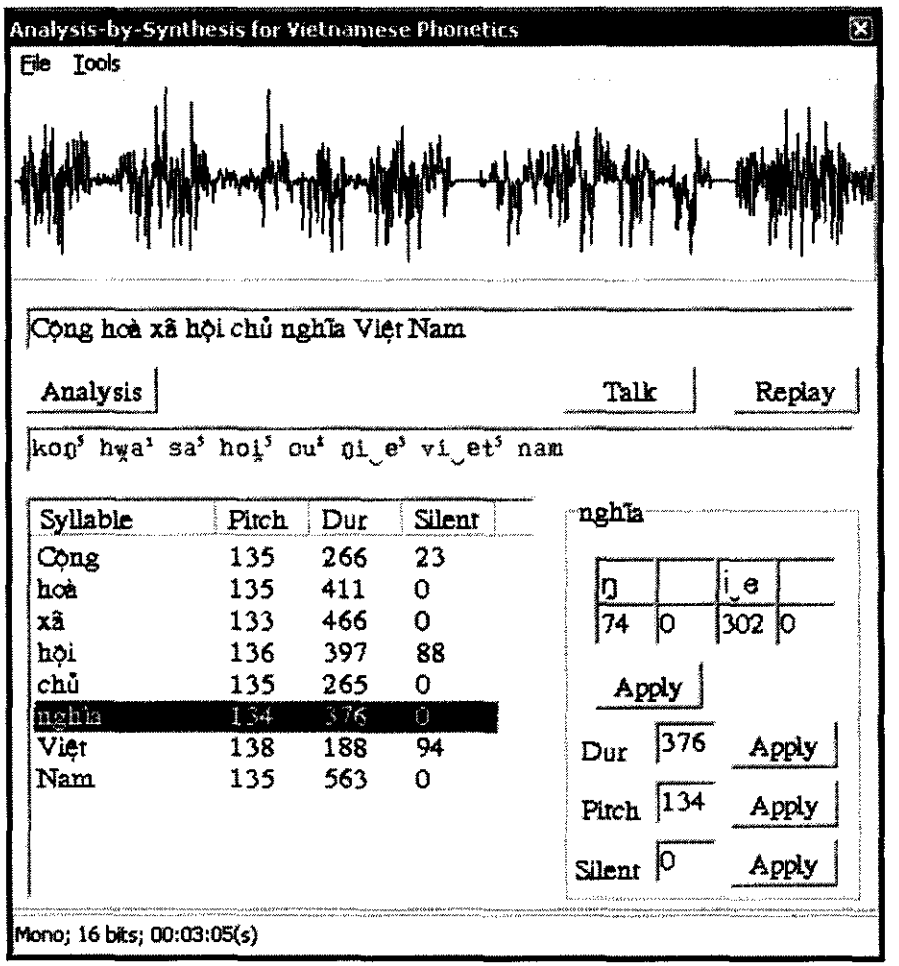


Hình 6.3. Từ điển cách đọc các từ lạ

6.1.2. Công cụ nghiên cứu ngữ âm tiếng Việt

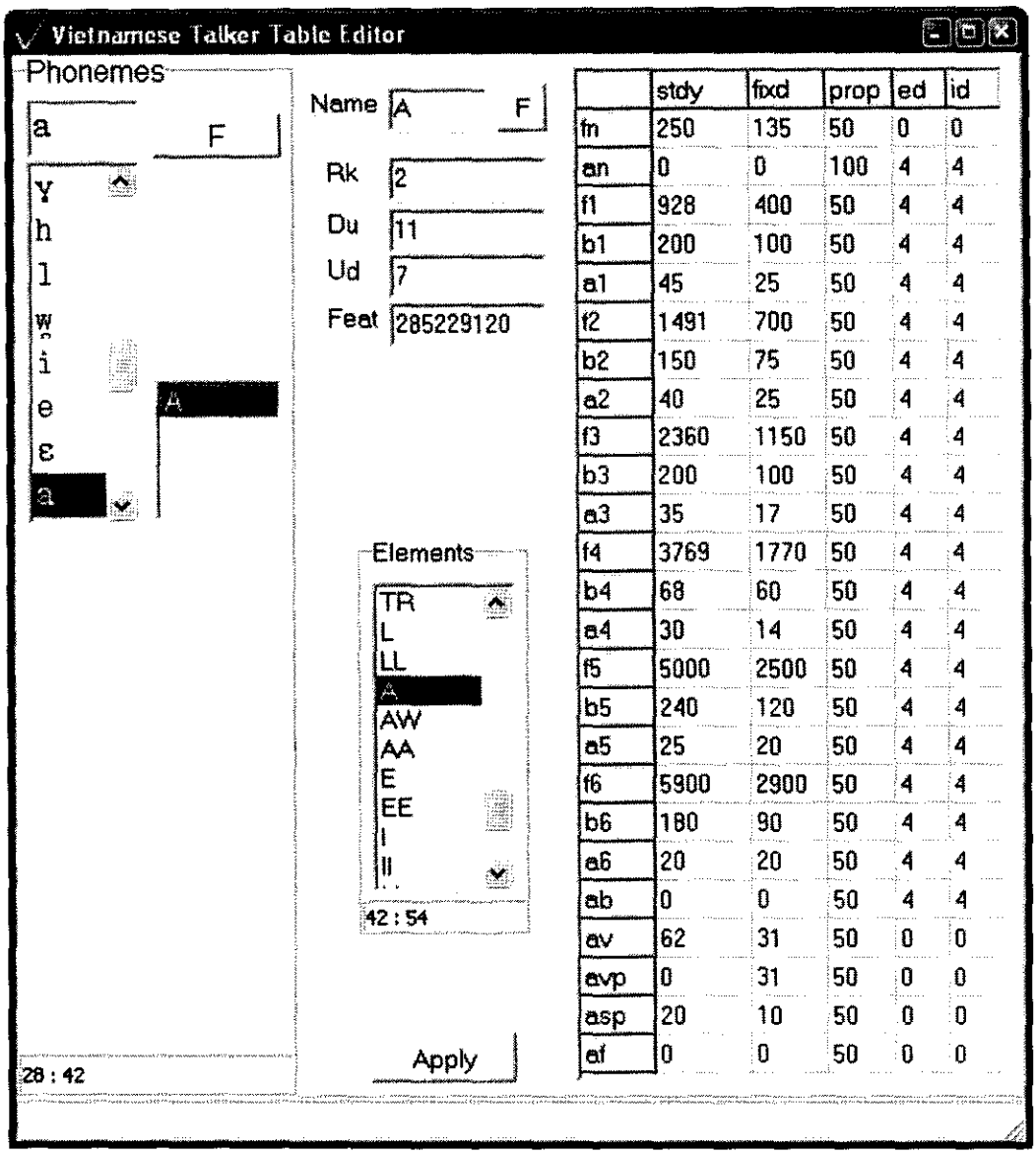
Một môi trường phần mềm để nghiên cứu trực quan ngữ âm tiếng Việt theo phương pháp "***Phân tích bằng Tổng hợp***" đã được xây dựng (hình 6.4).

Công cụ này cho phép khảo sát chuyển biểu diễn tiếng Việt từ dạng thông thường thành biểu diễn phiên âm theo bảng chữ cái ngữ âm quốc tế IPA. Người dùng có thể nghiên cứu ngữ điệu của ngữ đoạn tiếng Việt, phân tích ngữ đoạn thành các thành phần ngữ âm nhỏ hơn với các thông tin chi tiết về cao độ, trường độ. Đặc biệt, nó cho phép đánh giá một cách trực quan hiệu ứng của các sự điều chỉnh trường độ các âm vị cấu thành âm tiết, trường độ cả âm tiết, trường độ các khoảng nghỉ, biến thiên cao độ của các âm tiết khi được tổng hợp lại. Công cụ còn trình bày biểu diễn dạng sóng của tín hiệu tổng hợp, có thể lưu ra file dạng *.wav hay file các tham số điều khiển, đọc lại file các tham số điều khiển để tổng hợp.



Hình 6.4. Công cụ "Phân tích bằng Tổng hợp" ngữ âm tiếng Việt

Các thông số đặc trưng của các âm vị tiếng Việt, ảnh hưởng đến cấu thành âm tiết cũng có thể được khảo sát trực quan (hình 6.5).

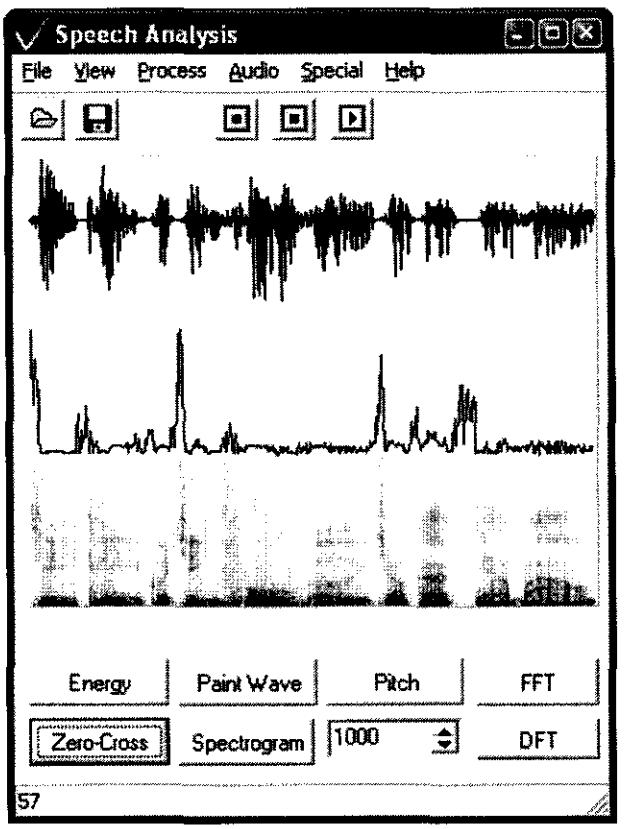


Hình 6.5. Editor khảo sát trực quan các đặc trưng của âm vị tiếng Việt

6.1.3. Công cụ phần mềm phân tích tín hiệu tiếng nói

Trong quá trình nghiên cứu, đề tài đã xây dựng được một số công cụ phần mềm để phân tích và biểu diễn với tín hiệu tiếng nói. Gồm có các chức năng ghi, phát lại, biểu diễn tín hiệu tiếng nói; tính toán và biểu diễn kết quả

phân tích FFT, năng lượng, tần số cắt 0, spectrogram, pitch. Hình 6.6 minh hoạ một số chức năng phân tích của chương trình.



Hình 6.6. Một số tính năng phân tích và biểu diễn tín hiệu tiếng nói

6.1.4. Chất lượng tiếng nói tổng hợp

Sử dụng bộ công cụ đánh giá chất lượng mô tả trong chương 5 với tiếng nói của hệ Vnspeech thu được kết quả như bảng 6.1. Các chỉ tiêu đánh giá độ nghe rõ của từng thành phần của âm tiết chỉ sử dụng trong quá trình phát triển, nên không trình bày ở đây, ngữ điệu chưa đặt ra trong phạm vi đề tài này nên cũng chưa thực hiện đánh giá. Các tiêu chí được đánh giá gồm: độ nghe rõ dãy số đọc kiểu số tự nhiên; đọc kiểu số điện thoại; và một ngữ đoạn có nghĩa bất kỳ. Người tham gia quá trình đánh giá gồm 3 nam và 2 nữ, thực hiện trong điều kiện phòng làm việc. Người đánh giá có thể thay đổi các tham số đặc trưng cho thích hợp nhất với khả năng nghe, được phép nghe trước để quen giọng và có thể nghe lại nếu chưa rõ, với mỗi chỉ tiêu cần đánh giá, người

đánh giá sẽ thực hiện chu trình lặp 100 lần trên phần mềm công cụ, kết quả chỉ ra trong bảng 6.1 dưới, dòng cuối là kết quả trung bình của cả 5 người.

Bảng 6.1. Chất lượng tiếng nói của Vnspeech

Stt	Người đánh giá	Kiểu số tự nhiên (%)	Kiểu số điện thoại (%)	Ngữ đoạn bất kỳ (%)
1	M1	100	100	94,08
2	M2	100	100	84,07
3	M3	100	100	78,25
4	F1	100	100	80,73
5	F2	100	100	77,41
Trung bình		100	100	82,90

Kết quả đánh giá cho thấy, mức độ nghe rõ phụ thuộc vào độ quen thuộc với giọng (tác giả - M1 có số mục nghe rõ lớn hơn cả). Với số lượng âm tiết hạn chế và biết trước (các chữ số và bội số), khả năng nghe rõ tốt hơn. Chất lượng tiếng nói của Vnspeech đã đạt được chỉ tiêu dự kiến, tuy nhiên, để có thể ứng dụng rộng rãi hơn nữa, vẫn tiếp tục cần tăng chất lượng nói chung và độ nghe rõ nói riêng, điều này hoàn toàn có thể thực hiện được.

6.2. Kết luận

Đề tài đã hoàn thành được mục tiêu đặt ra, đã xây dựng được phần mềm VnSpeech thực hiện chuyển văn bản thành tiếng nói bằng tổng hợp formant cho tiếng Việt. Hệ có thể được sử dụng như một ứng dụng độc lập hay thư viện phần mềm để tích hợp vào các ứng dụng khác. Chất lượng tiếng nói tổng hợp đã đạt mức chấp nhận được, có thể sử dụng trong nhiều ứng dụng. Hiện tại, có rất ít ngôn ngữ có thể tổng hợp được từ các thành phần đặc trưng nên đây là một thành công của đề tài đóng góp vào nghiên cứu ngôn ngữ tiếng Việt.

Các kết quả nghiên cứu ngữ âm tiếng Việt của đề tài không chỉ riêng áp dụng vào lĩnh vực tổng hợp mà còn có ích trong các công việc khác của xử lý tiếng nói tiếng Việt.

Kết quả của đề tài còn tạo thêm phương tiện mới để nghiên cứu phân tích ngữ âm tiếng Việt, tổng hợp là một phương pháp của nghiên cứu phân tích: ***phân tích bằng tổng hợp***.

6.3. Hướng nghiên cứu tương lai

Nghiên cứu chuyển văn bản thành tiếng nói tiếng Việt dựa trên tổng hợp formant liên quan đến nhiều ngành khoa học cơ bản khác nhau như ngôn ngữ, ngữ âm, xử lý tín hiệu số, khoa học máy tính, trong đó ngôn ngữ và ngữ âm tiếng Việt mang tính đặc thù của Việt nam. Để nhanh chóng có được các kết quả hữu ích, thành công nghệ/sản phẩm chung của xã hội, có thể áp dụng trong nhiều lĩnh vực, rất cần sự đầu tư của Nhà nước cũng như sự phối hợp nghiên cứu của các nhà khoa học thuộc các ngành liên quan.

Để tiếng nói tổng hợp cũng như hệ Vnspeech có thể ứng dụng hiệu quả trong các lĩnh vực cần phải nghiên cứu để tăng chất lượng hơn nữa. Các hướng chính để có thể tăng chất lượng tiếng nói tổng hợp cho hệ Vnspeech là:

- Tăng chất lượng tín hiệu: cần nghiên cứu sâu thêm về hệ thống âm vị, ngữ âm tiếng Việt, liên cấu âm giữa các âm vị trong âm tiết.
- Nghiên cứu và thể hiện các yếu tố ngữ điệu: cần phải sử dụng phân tích ngữ pháp, hiểu ngôn ngữ và thể hiện các đặc tính ngữ điệu khi tổng hợp, sự biến đổi các âm tiết khi nằm trong ngữ đoạn.
- Thêm nhiều giọng đọc khác nhau, đặc biệt phải có cả giọng nữ và trẻ em
- Engine tổng hợp phải nhỏ gọn và khả chuyển để thuận tiện cho tích hợp vào các loại ứng dụng.

TÀI LIỆU THAM KHẢO

[Acero] A.Acero: **Source-Filter Models for Time-Scale Pich-Scale Modification of Speech**, Microsoft Research. (from Internet).

[Bảng02] Vũ Kim Bảng: **Hệ formant của 9 nguyên âm đơn tiếng Hà Nội**, *Ngôn ngữ*, 15 (162) – 2002, tr. 56-63

[Bangayan97] Philbert Bangayan at al: **Analysis by synthesis of pathological voices using the Klatt synthesizer**, *Speech Communication 22* (1997) 343-368.

[Cawley96] G. Cawley: **The Application of Neural Networks to Phonetic Modelling**. *PhD. Thesis*, University of Essex, England, 1996. (from Internet)

[Cẩn97] Nguyễn Tài Cẩn: **Giáo trình Lịch sử ngữ âm tiếng Việt**, NXB Giáo dục - 1997.

[Chử00] Mai Ngọc Chử, Vũ Đức Nghiệu, Hoàng Trọng Phiến: **Cơ sở ngôn ngữ học và Tiếng Việt**, NXB Giáo dục, 2000.

[Donovan96] R. Donovan: **Trainable Speech Synthesis**, *PhD. Thesis*. Cambridge University Engineering Department, England, 1996. (from Internet)

[Festival] Festival homepage: http://www.cstr.ed.ac.uk/projects/festival.html

[Force] Force Computers - Products & Services - DECtalk® Text-To-Speech: http://www.forcecomputers.de/product/dectalk/dtalk.htm

[Freedom] Freedom Scientific Home Page: http://www.freedomscientific.com/

[Furui01] Sadaoki Furui: **Digital Speech: Processing, Synthesis, and Recognition**, (Second Edition, Revised and Expanded) - Marcel Dekker Inc. 2001.

[Giáp01] Nguyễn Thiện Giáp, Đoàn Thiện Thuật, Nguyễn Minh Thuyết: **Dẫn luận Ngôn ngữ học**, NXB Giáo dục - 2001.

[Hạo98] Cao Xuân Hạo: **Tiếng Việt- mấy vấn đề về ngữ âm, ngữ pháp, ngữ nghĩa**, NXB Giáo dục, 1998.

[Hùng03] Lê Xuân Hùng, Trịnh Văn Loan, Eric Castelli: **Automatic Synthesis of Vietnamese by concatenation of diphones**, *Kỷ yếu Hội thảo Khoa học Quốc gia lần thứ nhất* - Nghiên cứu Phát triển và Ứng dụng Công nghệ Thông tin và Truyền thông (ICT.rda) - Hà nội 22-23/2/2003. NXB Khoa học và Kỹ thuật, 7/2003, Tr 317-322.

[Jayant84] Jayant N.S., Peter Noll: **Digital coding of waveforms, Principles and Applications to Speech and Video**, Prentice-Hall. Inc. - 1984.

[Keller02] E.Keller, G.Bailly, A. Monaghan, J.Terken, M.Huckvale: **Improvements in Speech Synthesis,** *COST 258: The Naturaless of Synthetic Speech*, John Wiley & Sons Ltd. – 2002.

[Khánh00] Trịnh Đăng Khánh, Trịnh Văn Loan: **Lý thuyết thăng giáng và ứng dụng vào xác định hàm diện tích của các nguyên âm tiếng Việt**, *Tạp chí Bưu chính Viễn thông Chuyên san*: "Các công trình nghiên cứu và triển khai Công nghệ thông tin và Viễn thông" Tr. 70-76, Số 4 - Tháng 10/2000.

[Klatt87] Klatt D.H.: **Review of Text-To-Speech conversion for English,** *Journal of the Acoustical Society of America*, September 1987, pp v-793, 57 pp, 34 fig, 13 tab. (from Internet)

[Loan] Trịnh Văn Loan: **Xác định tham số đặc trưng cho nguyên âm không dấu tiếng Việt,** *Báo cáo khoa học* đề tài nghiên cứu KHCN 01-07.

[Loan99] Trịnh Văn Loan, Nguyễn Nam Hà, Phạm Việt Hà: **Xác định tham số đặc trưng của các nguyên âm không dấu của tiếng Việt**, *Tạp chí Bưu chính Viễn thông - Chuyên san*: "Các công trình nghiên cứu và triển khai Công nghệ thông tin và Viễn thông" Tr. 77-82, Số 2 - Tháng 12/1999.

[Mattingly74] I.G. Mattingly: "**Speech Synthesis for phonetic and phonological models**", *Current Trends in Linguistics*, Thomas A. Sebeok, Editor, Volume 12, Mouton, The Hague, pp. 2451-2487, - 1974. (from Internet).

[Minh01] Lê Hồng Minh: **Tổng hợp tiếng Việt,** *Luận văn Thạc sỹ* - Trường ĐHBK Hà nội, 2001.

[Minh02a] Lê Hồng Minh: **Nghiên cứu Tổng hợp tiếng Việt trên máy tính,** *Báo cáo Khoa học đề tài* - Viện Ứng dụng Công nghệ, 2002.

[Minh02b] Lê Hồng Minh: **Tổng hợp Formant Âm tiết tiếng Việt,** *Tạp chí Bưu chính Viễn thông*, kỳ 1 tháng 3/2002.

[Minh03a] Lê Hồng Minh: **Một số kết quả nghiên cứu và phát triển hệ phần mềm chuyển văn bản thành tiếng nói cho tiếng Việt bằng tổng hợp Formant,** *Kỷ yếu Hội thảo Khoa học Quốc gia lần thứ nhất* - Nghiên cứu Phát triển và Ứng dụng Công nghệ Thông tin và Truyền thông (ICT.rda) - Hà nội 22-23/2/2003. NXB Khoa học và Kỹ thuật, 7/2003, Tr 292-301.

[Minh03b] Lê Hồng Minh, Lê Khánh Hùng: **Phân tích và Tổng hợp đặc tính trường độ của tiếng Việt,** Trình bày tại *Hội thảo Quốc gia lần thứ 6* - Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông. Chủ đề: Xử lý ngôn ngữ và Đa phương tiện - Thái Nguyên 29-31/8/2003.

[Ngọc99] Quách Tuấn Ngọc: **Xử lý tín hiệu số**, NXB Giáo dục - 1999.

[Pelton93] G.E. Pelton: **Voice Processing**, McGraw-Hill, Inc., 1993.

[Potamianos99] A.Potamianos, P.Maragos: **Speech analysis and synthesis using an AM-FM modulation model**, *Speech Communication 28* (1999) 195-209.

[Pradit] Pradit Mittrapiyanuruk, at al.: **Issues in Thai Text-To-Speech Synthesis: The NECTEC Approach**, *NECTEC Technical Journal*, Vol.II, No.7. (from Internet)

[Quỳnh01] Nguyễn Hữu Quỳnh: **Ngữ pháp tiếng Việt**, Nhà xuất bản Từ điển Bách khoa, 2001

[Rabiner] L. Rabiner, Biing-Hwang Juang: **Fundamentals of Speech Recognition**, Prentice-Hall, Inc.

[Rabiner78] L.R. Rabiner, R.W. Schafer: **Digital Processing of Speech Signals**, Prentice-Hall, Inc., 1978.

[Richard96] Gael Richard, Christophe d'Alessandro: **Analysis/ synthesis and modification of speech aperiodic component**, *Speech Communication 19* (1996) 221-224.

[Rodman99] Robert D. Rodman: **Computer Speech Technoloy**, Artech House, 1999.

[Sairo85] Shuzo Sairo, Kazuo Nakara: **Fundamentals of Speech Signal Processing**, Academic Press - 1985.

[Sensimetrics] Sensimetrics Corporation homepage: http://www.sens.com

[SoftVoice] SoftVoice, Inc. Homepage: http://www.text2speech.com/;

[SpeakJets] SpeakJets Homepage http://www.speechchips.com/

[SpeechTech] Speech Technology Magazine, Homepage: http://speechtechmag.com/

[Sproat95] Richard W. Sproat, Joseph P.Ovive: **Text to Speech Synthesis**, *AT&T Technical Journal*, March/April 1995.

[Sproat99] R. Sproat, M. Ostendorf, A. Hunt: **The Need for Increased Speech Synthesis Reseach** (Report of the 1998 NSF Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis), March 1999 (from Internet)

[Styger94] Styger, T., Keller, E.: **Formant synthesis**. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges* (pp. 109-128). Chichester: John Wiley. (from Internet).

[Syndal] A.K. Syndal, Alan W. Black at al.: "**Three Methods of Intonation Modeling**" (from Internet).

[Taylor] Paul Taylor: **Analysis and Synthesis of Intonation using the Tilt Model**, Center for Speech Technology Research – University of Edinburgh. (from Internet).

[Thắng00] Phan Quốc Thắng, Trịnh Đăng Khánh: **Ứng dụng mô hình nguồn âm và bộ lọc của quá trình tạo tiếng nói để khảo sát nguồn âm có mang tính thanh điệu và dạng tuyến âm một số nguyên âm tiếng Việt**, *Tạp chí Bưu chính Viễn thông- Chuyên san*: "Các công trình nghiên cứu và triển khai Công nghệ thông tin và Viễn thông" Tr. 10-13, Số 3 - Tháng 6/2000.

[Thuật99] Đoàn Thiện Thuật: **Ngữ âm tiếng Việt**, NXB Đại học Quốc gia Hà nội - 1999.

[Trung99] Nguyễn Quốc Trung: **Xử lý tín hiệu và lọc số**, NXB KHKT - 1999.

[Tuấn00a] Trịnh Anh Tuấn, Đỗ Trung Tá: **Một số kết quả nghiên cứu về thanh điệu trong phát âm và sự biến thanh cho tổng hợp tiếng Việt**, *Tạp chí Bưu chính Viễn thông - Chuyên san*: "Các công trình nghiên cứu và triển khai Công nghệ thông tin và Viễn thông" Tr. 5-9, Số 3 - Tháng 6/2000.

[Tuấn00b] Trịnh Anh Tuấn: **Một số phương pháp nâng cao chất lượng hệ thống tổng hợp tiếng Việt V-TALK**, *Tạp chí Bưu chính Viễn thông- Chuyên san*: "Các công trình nghiên cứu và triển khai Công nghệ thông tin và Viễn thông" Tr. 19-23, Số 3 - Tháng 6/2000.

[Tuấn00c] Trịnh Anh Tuấn: **Nghiên cứu các đặc trưng để phân tích và tổng hợp tín hiệu âm tần**, *Luận án Tiến sỹ Kỹ thuật*, 2000.

[Tuý96] Hồ Anh Tuý: **Xử lý tín hiệu số**, NXB Khoa học và Kỹ thuật - 1996

[VanSanten97] Jan P.H. Van Santen, Richard W. Sproat, Joseph P.Olive, Julia Hirschberg: **Progress in Speech Synthesis**, Springer - 1997.

[WaveSurfer] WaveSurfer: http://www.speech.kth.se/wavesurfer/index.html