LỜI NÓI ĐẦU

Kể từ khi xuất hiện, máy tính càng ngày càng chứng tỏ rằng đó là một công cụ vô cùng hữu ích trợ giúp con người xử lý thông tin. Cùng với sự phát triển của xã hội, khối lượng thông tin mà máy tính cần xử lý tăng rất nhanh trong khi thời gian dành cho những công việc này lại giảm đi. Vì vậy, việc tăng tốc độ xử lý thông tin, trong đó có tốc độ trao đổi thông tin giữa con người và máy tính, trở thành một yêu cầu cấp thiết. Hiện tại, giao tiếp người-máy được thực hiện bằng các thiết bị như bàn phím, chuột, màn hình,... với tốc độ tương đối chậm nên cần có các phương pháp trao đổi thông tin mới giúp con người làm việc hiệu quả hơn với máy tính. Một trong những hướng nghiên cứu này là sử dụng tiếng nói trong trao đổi thông tin người-máy. Những nghiên cứu này liên quan trực tiếp tới các kết quả của chuyên ngành xử lý tiếng nói, trong đó có tổng hợp tiếng nói.

Tổng hợp tiếng nói là lĩnh vực đang được nghiên cứu khá rộng rãi trên thế giới và đã cho những kết quả khá tốt. Có ba phương pháp cơ bản dùng để tổng hợp tiếng nói là mô phỏng bộ máy phát âm, tổng hợp bằng formant và tổng hợp bằng cách ghép nối. Phương pháp mô phỏng bộ máy phát âm cho chất lượng tốt nhưng đòi hỏi nhiều tính toán vì việc mô phỏng chính xác bộ máy phát âm rất phức tạp. Phương pháp tổng hợp formant không đòi hỏi chi phí cao trong tính toán nhưng cho kết quả chưa tốt. Phương pháp tổng hợp ghép nối cho chất lượng tốt, chí phí tính toán không cao nhưng số lượng từ vựng phải rất lớn.

Ở các nước phát triển, những nghiên cứu xử lý tiếng nói, đã cho các kết quả khả quan, làm tiền đề cho việc giao tiếp người-máy bằng tiếng nói. Ở Việt Nam, các nghiên cứu trong lĩnh vực này tuy mới được phát triển trong những năm gần đây nhưng cũng đã có một số kết quả khả quan.

Với mục đích góp phần vào sự phát triển của tổng hợp tiếng Việt, đề tài này nghiên cứu về phương pháp tổng hợp tiếng Việt bằng phương pháp ghép nối dựa trên giải thuật TD-PSOLA.

TD-PSOLA là phiên bản trên miền thời gian của giải thuật PSOLA (Pitch Synchronous Overlap-Add). Với PSOLA, tín hiệu tổng hợp được tạo nên bằng cách cộng xếp chồng (Overlap-Add) các đoạn tín hiệu thành phần. Giải thuật này cho phép thao tác trực tiếp với tín hiệu tiếng nói trên miền thời gian, thay đổi tần số cơ bản và độ dài của tín hiệu. Để giảm số lượng từ vựng khi xây dựng ứng dụng, các từ tiếng Việt sẽ được tổng hợp từ các diphone.

Sau khi nghiên cứu về mặt lý thuyết, báo cáo này cũng trình bày việc áp dụng thuật toán để xây dựng một ứng dụng tổng hợp tiếng Việt từ văn bản.

Với nội dụng như vậy, báo cáo được chia làm 4 chương:

- **Chương I**: **Tiếng nói và xử lý tiếng nói**. Chương này đề cập tới những vấn đề cơ bản nhất về các đặc trưng của tín hiệu tiếng nói và các lĩnh vực của xử lý tiếng nói.
- **Chương II**: **Tổng hợp tiếng nói** sẽ trình bày các phương pháp khác nhau trong tổng hợp tiếng nói đồng thời đưa ra đánh giá về hiệu quả của các phương pháp này.
- **Chương III**: **Giải thuật TD-PSOLA**. Chương này trình bày chi tiết về giải thuật PSOLA và phiên bản trên miền thời gian TD-PSOLA, đồng thời cũng đề cập tới các vấn đề liên quan khi áp dụng cho tín hiệu tiếng nói.
- **Chương IV**: **Thiết kế chương trình tổng hợp tiếng Việt**. Dựa trên các nghiên cứu lý thuyết trong chương III, chương này sẽ trình bày cách áp dụng thuật toán TD-PSOLA để xây dựng chương trình tổng hợp tiếng Việt từ văn bản và các kết quả liên quan.

Các kết quả thu được khi áp dụng:

- Có thể biến đổi tần số cơ bản của tín hiệu tiếng nói để tạo các thanh điệu trong tiếng Việt.
- Có thể thay đổi thời gian, biên độ và ngữ điệu của từ, làm cơ sở cho việc tổng hợp câu trong tiếng Việt.
- Khắc phục được khó khăn về số lượng dữ liệu: Số lượng diphone không lớn (389 diphone).

Với những kết quả này, trong tương lai có thể phát triển tiếp đề tài theo những hướng nghiên cứu như mở rộng cơ sở dữ liệu, xử lý văn bản ở mức cao...

MỤC LỤC

CHƯƠNG 1

TIẾNG NÓI VÀ XỬ LÝ TIẾNG NÓI

1.1. MỞ ĐẦU

Tiếng nói là một phương tiện trao đổi thông tin của con người. Tiếng nói được tạo ra từ tư duy của con người: trung khu thần kinh điều khiển hệ thống phát âm làm việc tạo ra âm thanh.

Tiếng nói được phân biệt với các âm thanh khác bởi các đặc tính âm học có nguồn gốc từ cơ chế tạo tiếng nói.Về bản chất, tiếng nói là sự dao động của không khí có mang theo thông tin. Các dao động này tạo thành những áp lực đến tai và được tai phát hiện, phân tích và chuyển kết quả đến trung khu thần kinh. Lúc này tại trung khu thần kinh, thông tin được tái tạo lại dưới dạng tư duy logic mà con người có thể hiểu được.

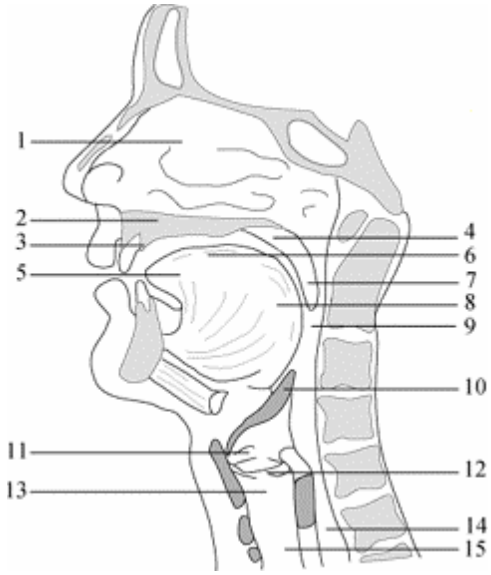
Tín hiệu tiếng nói được tạo thành bởi các chuỗi các âm vị liên tiếp. Sự sắp xếp của các âm vị được chi phối bởi các quy tắc của ngôn ngữ. Việc nghiên cứu một cách chi tiết về những quy tắc này cũng như những khía cạnh khác bên trong tiếng nói thuộc về chuyên ngành ngôn ngữ. Việc phân loại các âm vị của tiếng nói thuộc về chuyên ngành ngữ âm học. Khi nghiên cứu các mô hình toán học của cơ chế tạo tiếng nói, việc nghiên cứu về các âm vị là rất cần thiết.

1.2. BỘ MÁY PHÁT ÂM

1.2.1. Bộ máy phát âm

Bộ máy phát âm bao gồm các thành phần riêng rẽ như phổi, khí quản, thanh quản, và các đường dẫn miệng, mũi. Trong đó:

- Thanh quản chứa hai dây thanh có thể dao động tạo ra sự cộng hưởng cần thiết để tạo ra âm thanh.
- Tuyến âm là ống không đều bắt đầu từ môi, kết thúc bởi dây thanh hoặc thanh quản.
- Khoang mũi là ống không đều bắt đầu từ môi, kết thúc bởi vòm miệng, có độ dài cố định khoảng 12cm đối với người lớn.
- Vòm miệng là các nếp cơ chuyển động.



1. Hốc mũi
2. Vòm miệng trên
3. Ổ răng
4. Vòm miệng mềm
5. Đầu lưỡi
6. Thân lưỡi
7. Lưỡi gà
8. Cơ miệng
9. Yết hầu
10. Nắp đóng của thanh quản
11. Dây thanh giả
12. Dây thanh
13. Thanh quản
14. Thực quản

Hình 1.1. Bộ máy phát âm của con người

1.2.2. Cơ chế phát âm

Trong quá trình tạo âm thanh không phải là âm mũi, vòm miệng mở, khoang mũi đóng lại, dòng khí sẽ chỉ đi qua khoang mũi. Khi phát âm mũi, vòm miệng hạ thấp và dòng khí sẽ chỉ đi qua khoang mũi.

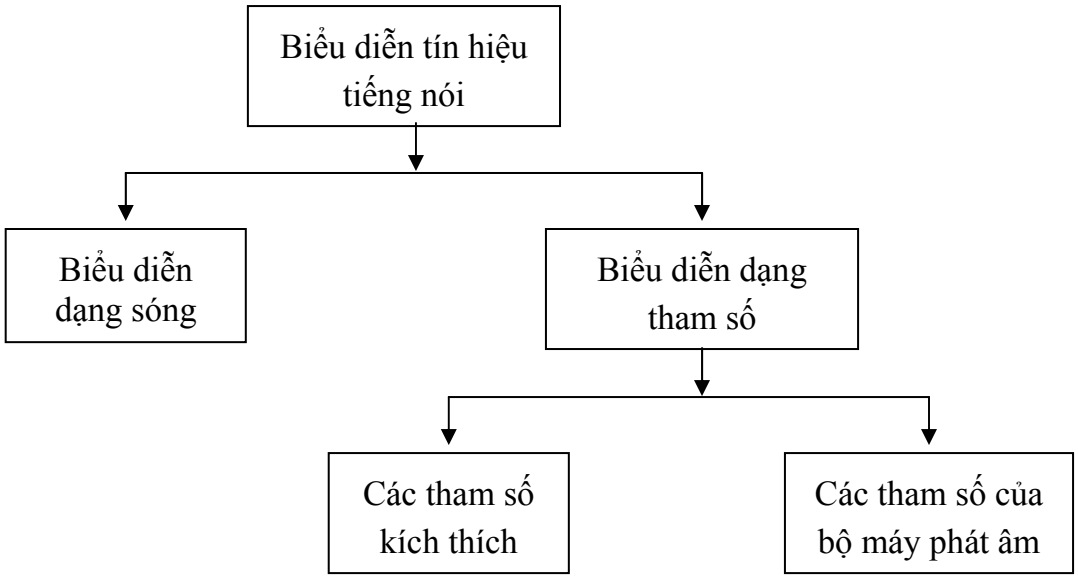
Tuyến âm sẽ được kích thích bởi nguồn năng lượng chính tại thanh môn. Tiếng nói được tạo ra do tín hiệu nguồn từ thanh môn phát ra, đẩy không khí có trong phổi lên tạo thành dòng khí, va chạm vào hai dây thanh trong tuyến âm. Hai dây thanh dao động sẽ tạo ra cộng hưởng, dao động âm sẽ được lan truyền theo tuyến âm (tính từ tuyến âm đến khoang miệng) và sau khi đi qua khoang mũi và môi, sẽ tạo ra tiếng nói.

1.3. BIỂU DIỄN TÍN HIỆU TIẾNG NÓI

Tín hiệu tiếng nói là tín hiệu tương tự. Do đó khi biểu diễn tín hiệu tiếng nói trong môi trường tính toán của tín hiệu số, việc biểu diễn và lưu trữ sao cho không bị mất mát thông tin là vấn đề hết sức quan trọng trong các hệ thống thông tin có sử dụng tín hiệu tiếng nói. Việc xem xét các vấn đề xử lý tín hiệu tiếng nói trong các hệ thống này dựa trên ba vấn đề chính:

- Biểu diễn tín hiệu tiếng nói dạng số.
- Cài đặt các kỹ thuật xử lý.
- Các lớp ứng dụng dựa trên kỹ thuật xử lý tín hiệu số.

Phần này trình bày vấn đề biểu diễn tiếng nói dưới dạng số. Mô hình tổng quát các phương pháp biểu diễn tín hiệu tiếng nói được trình bày trên hình 1.2.



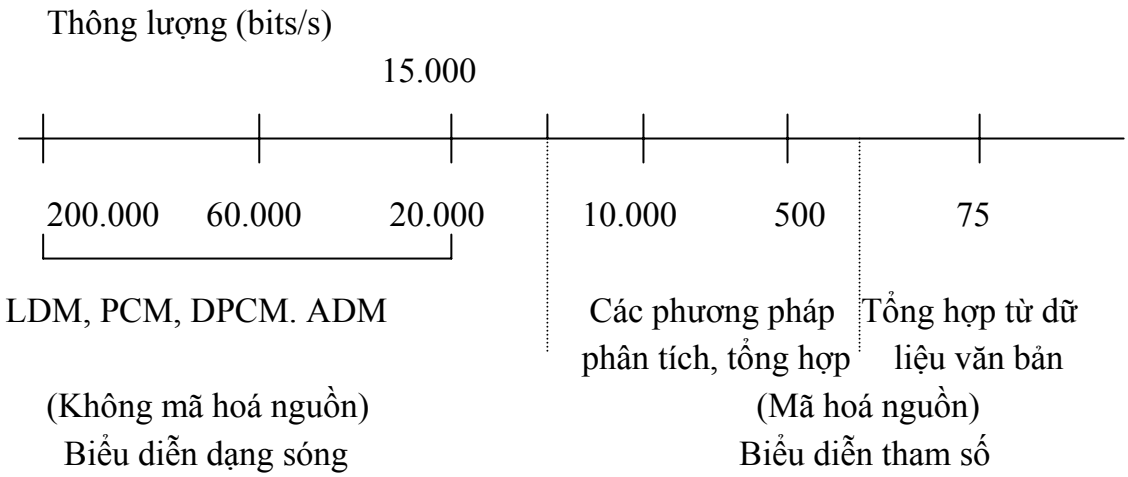
Hình 1.2. Biểu diễn tín hiệu tiếng nói

Biểu diễn tín hiệu tiếng nói theo dạng số chịu ảnh hưởng quan trọng của lý thuyết lấy mẫu, theo đó các trạng thái của tín hiệu có dải tần giới hạn có thể được biểu diễn dưới dạng các mẫu lấy tuần hoàn theo một chu kỳ cố định gọi là chu kỳ lấy mẫu. Việc lấy mẫu này sẽ cung cấp cho hệ thống những mẫu tín hiệu với tỷ lệ đủ lớn để xử lý. Tất cả các quá trình xử lý lấy mẫu được chỉ rõ trong các tài liệu về xử lý tín hiệu số. Có nhiều phương pháp biểu diễn rời rạc tín hiệu tiếng nói. Hình 1.2 chỉ ra những phương pháp biểu diễn này. Các khả năng biểu diễn như thế được phân thành hai nhóm chính: nhóm biểu diễn tín hiệu dạng sóng (waveform) và nhóm biểu diễn tín hiệu theo tham số (parametric).

Phương pháp biểu diễn tín hiệu theo dạng sóng như bản thân tên của nó, được xem xét đến với việc bảo mật thông tin theo cách thông thường là giữ nguyên hình dạng sóng của tín hiệu tương tự sau khi đã qua các bước lấy mẫu và lượng tử hoá tín hiệu.

Trên phương diện khác, phương pháp biểu diễn tín hiệu theo tham số được xem xét đến trên khía cạnh biểu diễn tín hiệu tiếng nói như đầu ra của hệ thống tạo tiếng nói. Để thu được các tham số biểu diễn, bước đầu tiên của phương pháp này lại thường là biểu diễn tín hiệu theo dạng sóng. Điều này có nghĩa là tín hiệu tiếng nói được lấy mẫu và lượng tử hoá giống như phương pháp biểu diễn tín hiệu tiếng nói dạng sóng, sau đó tiến hành xử lý để thu được các tham số của tín hiệu tiếng nói của mô hình tạo tiếng nói nêu trên. Các tham số của mô hình tạo tiếng nói này thường được phân loại thành các tham số tín hiệu nguồn (có quan hệ mật thiết với nguồn của tiếng nói) và các tham số của

bộ máy phát âm tương ứng (có quan hệ mật thiết với giọng nói của từng người). Hình 1.3 chỉ ra những sự khác nhau của một số dạng biểu diễn tín hiệu tiếng nói theo các yêu cầu của thông lượng (bits/s):



Hình 1.3. Thông lượng cho các phương pháp biểu diễn tiếng nói

Đường phân cách ở giữa (tương ứng với thông lượng 15.000 bits/s) chia khoảng dữ liệu thành hai phần riêng biệt: phần thông lượng cao dành cho dạng biểu diễn tín hiệu dạng sóng ở phía trái và phần thông lượng thấp ở bên phải dành cho biểu diễn tín hiệu dạng tham số. Hình vẽ trên chỉ ra sự thay đổi trong khoảng từ 75 bits/s (xấp xỉ thông lượng khi tổng hợp văn bản) cho tới thông lượng trên 200.000 bits/s cho các dạng biểu diễn sóng đơn giản. Điều này cho phép biểu diễn từ 1 đến 3.000 cách cho thông lượng tuỳ thuộc vào tín hiệu nói cần biểu diễn. Tất nhiên là thông lượng không chỉ phụ thuộc tín hiệu cần biểu diễn mà nó còn phụ thuộc vào các yếu tố khác như giá thành, sự mềm dẻo của phương pháp biểu diễn, chất lượng của tiếng nói.

Vì tiếng nói là tín hiệu liên tục nên để áp dụng các phương pháp xử lý tín hiệu thì tiếng nói phải được biểu diễn dưới dạng rời rạc. Quá trình rời rạc hoá tín hiệu tiếng nói bao gồm các bước sau:

- Lấy mẫu tín hiệu tiếng nói với tần số lấy mẫu f_0.
- Lượng tử hoá các mẫu với các bước lượng tử q.
- Mã hoá và nén tín hiệu.

Sau đây chúng ta xét qua các bước này.

1.3.1. Xác định tần số lấy mẫu

Khi lấy mẫu một tín hiệu tương tự với tần số lấy mẫu f_0 cần đảm bảo rằng việc khôi phục lại tín hiệu đó từ tín hiệu rời rạc tương ứng phải thực hiện được. Shanon đã đưa ra một định lý mà theo đó người ta có thể xác định tần số lấy

mẫu đảm bảo yêu cầu trên. Theo Shanon, điều kiện cần và đủ để khôi phục lại tín hiệu tương tự từ tín hiệu đã được rời rạc hoá với tần số f_0 là: $f_0 \geq f_{MAX}$ với f_{MAX} là tần số lớn nhất của tín hiệu tương tự.

Phổ của tín hiệu tiếng nói trải rộng trong khoảng 12 kHz, do đó theo định lý Shanon thì tần số lấy mẫu tối thiểu là 24 kHz. Với tần số lấy mẫu lớn như thế thì khối lượng bộ nhớ dành cho việc ghi âm sẽ rất lớn và làm tăng sự phức tạp trong tính toán. Nhưng chi phí cho việc xử lý tín hiệu số, bộ lọc, sự truyền và ghi âm có thể giảm đi nếu chấp nhận giới hạn phổ bằng cách cho tín hiệu qua một bộ lọc tần số thích hợp. Đối với tín hiệu tiếng nói cho điện thoại, người ta thấy rằng tín hiệu tiếng nói đạt chất lượng cần thiết để mức độ ngữ nghĩa của thông tin vẫn bảo đảm khi phổ được giới hạn ở 3400 Hz. Khi đó tần số lấy mẫu sẽ là 8000 Hz. Trong kỹ thuật phân tích, tổng hợp hay nhận dạng tiếng nói, tần số lấy mẫu có thể dao động trong khoảng 6.000 – 16.000 Hz. Đối với tín hiệu âm thanh (bao gồm cả tiếng nói và âm nhạc) tần số lấy mẫu cần thiết là 48 kHz.

1.3.2. Lượng tử hoá

Việc biểu diễn số tín hiệu đòi hỏi việc lượng tử hoá mỗi mẫu tín hiệu với một giá trị rời rạc hữu hạn. Mục tiêu của công việc này hoặc là để truyền tải hoặc là xử lý có hiệu quả. Trong trường hợp thứ nhất mỗi mẫu tín hiệu được lượng tử hoá, mã hoá rồi truyền đi. Bên thu nhận tín hiệu giải mã và thu được tín hiệu tương tự. Tính thống kê của tín hiệu được bảo toàn sẽ ảnh hưởng quan trọng đến thuật toán lượng tử hoá. Trong trường hợp xử lý tín hiệu, luật lượng tử hoá được quy định bởi hệ thống xử lý, nó có thể được biểu diễn bằng dấu phẩy tĩnh hay dấu phẩy động. Việc xử lý bằng dấu phẩy động cho phép thao tác với tín hiệu khá mềm dẻo mặc dù chi phí tính toán cao. Việc xử lý bằng dấu phẩy tĩnh đơn giản hơn nhiều nhưng đòi hỏi các điều kiện chặt chẽ đối với các thuật toán xử lý.

1.3.3. Nén tín hiệu tiếng nói

Lượng tử hoá tín hiệu gây ra các lỗi có thành phần giống nhiễu trắng, như vậy số bước lượng tử cần được phân bố theo tỷ lệ trên lỗi thích hợp. Nếu số bước lượng tử là cố định thì tỷ số này là hàm của biên độ tín hiệu, người ta sử dụng luật lượng tử logarithm và mỗi mẫu tín hiệu được biểu diễn bằng 8 bit. Đối với tín hiệu âm thanh kích thước mẫu thường là 16 bit.

Một đặc trưng cần thiết của phép biểu diễn tín hiệu số là tốc độ nhị phân tính bằng bit/s. Đó là giá trị quan trọng trong khi thực hiện truyền dữ liệu cũng như lưu trữ dữ liệu. Đường truyền điện thoại có tốc độ là

8(kHz)*8(bit)=64kb/s. Khi thực hiện truyền và ghi lại tín hiệu âm thanh, tốc độ cần thiết 768 kb/s.

Ta biết rằng tín hiệu tiếng nói có độ dư thừa rất lớn, do đó có thể giảm tốc độ tín hiệu tuỳ thuộc mục đích xử lý khi xem xét đến mức độ phức tạp của các thuật toán cũng như xem xét đến chất lượng của việc biểu diễn tín hiệu tiếng nói. Có nhiều kỹ thuật đưa ra để đạt được các mục đích trên. Sự lựa chọn một phương pháp biểu diễn số tín hiệu thoả mãn giữa các tiêu chuẩn về chất lượng của của phép biểu diễn, tốc độ lưu truyền hay lưu trữ và cuối cùng là các điều kiện môi trường (như nhiễu,...).

Thông thường số bit có nghĩa dùng để biểu diễn chuỗi lượng tử cần phải giảm bớt vì lý do kỹ thuật. Việc này có thể thực hiện được bằng cách bỏ đi các bit ít có nghĩa nhất, nếu phép lượng tử là tuyến tính, lỗi lượng tử tăng cùng với khoảng giá trị của chuỗi. Nhưng đối với một vài ứng dụng, mức lượng tử ở vùng tần số cao có yêu cầu thấp hơn so với mức lượng tử ở vùng tần số thấp hay ngược lại, trong trường hợp đó cần sử dụng toán tử tuyến tính để biến đổi tín hiệu.

Kỹ thuật truyền tin trong điện thoại thường sử dụng luật nén tín hiệu theo đường cong logarithm. Có hai luật nén được sử dụng phổ biến hiện nay là luật μ và luật A.

1.3.4. Mã hoá tín hiệu tiếng nói.

a. Mã hoá trực tiếp tín hiệu

Phương pháp mã hoá trực tiếp hay phổ tín hiệu cho phép biểu diễn một cách trung thực nhất tín hiệu. Mã hoá trực tiếp thực chất là biểu diễn mỗi mẫu tín hiệu hay phổ tín hiệu độc lập khác với các mẫu khác. Một hệ thống mã hoá tín hiệu khá phổ biến hiện nay theo phương pháp này thực hiện trong miền thời gian là mã hoá xung PCM (Pulse Code Mudulation).

Để bảo đảm biểu diễn tín hiệu đạt chất lượng cao phải bảm bảo được thông lượng cần thiết. Do tần số lấy mẫu đã được cố định, muốn giảm được thông lượng này phải giảm số bit dùng biểu diễn một mẫu. Muốn vậy phải áp dụng luật lượng tử phù hợp với thống kê bậc một của tín hiệu, nghĩa là phù hợp với mật độ phân bố và sự thay đổi của tín hiệu. Hệ thống PCM có thể giảm thông lượng xuống còn 64 kb/s.

Cũng theo hướng này người ta dùng hàm tự hồi quy để thực hiện nén tín hiệu. Khi đó mỗi mẫu mới của tín hiệu tiếng nói lại không chứa các đặc điểm hoàn toàn mới, nó chắc chắn có liên quan đến các mẫu trước đó.

Như vậy mỗi mẫu tín hiệu tiếng nói, bằng nhiều phương pháp có thể tiên đoán nhờ một số mẫu trước đó, khi đó chỉ cần tính toán sai số dự đoán và biến

đổi. Tại nơi nhận tín hiệu, một phép biến đổi ngược lại được thực hiện và người ta thấy rằng hệ số khuếch đại của hệ thống đối với thông lượng là hàm chất lượng của phép tiên đoán. Các hệ thống hoạt động theo nguyên tắc này có:

- DPCM (Differential PCM): Hệ thống PCM dùng phép tiên đoán cố định. Thay vì truyền mẫu tín hiệu, phương pháp này truyền đi các hệ số tiên đoán và sai số dự đoán.
- ADPCM (Adaptive DPCM): Hệ thống PCM dùng phép tiên đoán thích nghi. Hệ thống này là hệ thống cải tiến của hệ thống DPCM, người ta sẽ dùng hàm tự hồi quy trong thời gian ngắn để tính toán các hệ số tiên đoán với một đoạn mẫu tín hiệu khoảng 20 ms. Những tính toán này thực hiện trong thời gian thực.

Biểu diễn số của tín hiệu có thể thực hiện trong cả miền tần số bằng cách mã hoá biến đổi Fourier của tín hiệu. Trong miền tần số, phép mã hoá trực tiếp ít được áp dụng. Các kỹ thuật giảm bớt thông lượng được thực hiện bằng cách giảm độ dư thừa tự nhiên của tín hiệu tiếng nói trên phổ tín hiệu. Theo phương pháp này người ta dùng cách mã hoá băng thấp hay mã hoá thích nghi theo biến đổi ATC.

b. Mã hoá tham số tín hiệu

Để giảm hơn nữa thông lượng của tiếng nói tới khoảng giá trị 2000 – 3000 b/s, cần phải dùng các kết quả nghiên cứu về phương thức tạo ra tiếng nói con người. Có nhiều phương pháp cho phép đánh giá các tham số của mô hình tạo tiếng nói bao gồm hàm đặc trưng của tuyến âm và các đặc trưng của nguồn âm.

Tín hiệu tiếng nói được coi gần như dừng trong khoảng thời gian là 20 ms; như vậy các tham số được tính toán lại sau 20 ms và được thực hiện trong thời gian thực. Người ta thấy rằng việc truyền tham số này cho phép thông lượng giảm xuống còn khoảng 2500b/s. Phương pháp mã hoá này gọi là phương pháp mã hoá nguồn tham số tín hiệu.

Một tập hợp các tham số khi truyền hay lưu trữ đặc trưng cho phổ thời gian ngắn, có nghĩa là nó chỉ được chấp nhận trong một thời gian hạn chế. Tai người rất nhạy cảm với các phổ thời gian này, do đó có thể cho rằng tai người có thể phân biệt được một số hữu hạn các phổ thời gian ngắn. Giả sử M = 2B. Như vậy với mỗi phổ thời gian ngắn, ta gán cho nó một giá trị biểu diễn bằng một từ B bit và từ này sẽ được truyền đi hay lưu trữ. Bằng cách này thông lượng có thể giảm xuống còn 1000 b/s.

Tín hiệu tổng hợp bằng mã hoá theo tham số các tín hiệu tiếng nói thường không bảo đảm chất lượng trong hệ thống điện thoại thông thường. Giọng nói sẽ rất khó nhận ra trong trường hợp dùng phương pháp này. Do đó kỹ thuật mã hoá này chỉ ứng dụng trong điện thoại di động và quân sự...

1.4. ĐẶC TÍNH ÂM HỌC CỦA TIẾNG NÓI

1.4.1. Âm hữu thanh và âm vô thanh

a. Âm hữu thanh

Âm hữu thanh được tạo ra từ các dây thanh bị căng đồng thời và chúng rung động ở chế độ dãn khi không khí tăng lên làm thanh môn mở ra và sau đó thanh môn xẹp xuống do không khí chạy qua.

Do sự cộng hưởng của dây thanh, sóng âm tạo ra có dạng tuần hoàn hoặc gần như tuần hoàn. Phổ của âm hữu thanh có nhiều thành phần hài tại giá trị bội số của tần số cộng hưởng, còn gọi là tần số cơ bản (pitch).

b. Âm vô thanh

Khi tạo ra âm vô thanh dây thanh không cộng hưởng. Âm vô thanh có hai loại cơ bản là âm xát và âm tắc.

Âm xát (ví dụ như âm s) được tạo ra khi có sự co thắt tại vài điểm trong tuyến âm. Không khí khi đi qua điểm co thắt sẽ chuyển thành chuyển động hỗn loạn tạo nên kích thích giống như nhiễu ngẫu nhiên. Thông thường điểm co thắt xảy ra gần miệng nên sự cộng hưởng của tuyến âm ảnh hưởng rất ít đến đặc tính của âm xát được tạo ra.

Âm tắc (ví dụ như âm p) được tạo ra khi tuyến âm đóng tại một số điểm làm cho áp suất không khí tăng lên và sau đó được giải phóng đột ngột. Sự giải phóng đột ngột này tạo ra kích thích nhất thời của tuyến âm. Sự kích thích này có thể xảy ra với sự cộng hưởng hoặc không cộng hưởng của dây thanh tương ứng với âm tắc hữu thanh hoặc vô thanh.

1.4.2. Âm vị

Tín hiệu tiếng nói là tín hiệu tương tự biểu diễn cho thông tin về mặt ngôn ngữ và được mô tả bởi các âm vị khác nhau. Như vậy, âm vị là đơn vị nhỏ nhất của ngôn ngữ. Tuỳ theo từng ngôn ngữ cụ thể mà số lượng các âm vị nhiều hay ít (thông thường số lượng các âm vị vào khoảng 20 – 30). Các âm vị được chia thành hai loại: nguyên âm và phụ âm.

a. Nguyên âm

Nguyên âm là âm hữu thanh được tạo ra bằng sự cộng hưởng của dây thanh khi dòng khí được thanh môn đẩy lên. Khoang miệng được tạo lập thành nhiều hình dạng nhất định tạo thành các nguyên âm khác nhau. Số lượng các nguyên âm phụ thuộc vào từng ngôn ngữ nhất định.

b. Phụ âm

Phụ âm được tạo ra bởi các dòng khí hỗn loạn được phát ra gần những điểm co thắt của đường dẫn âm thanh do cách phát âm tạo thành. Phụ âm có đặc tính hữu thanh hay vô thanh tuỳ thuộc vào việc dây thanh có dao động để tạo nên cộng hưởng không. Dòng không khí tại chỗ đóng của vòm miệng tạo ra phụ âm tắc. Phụ âm xát được phát ra từ chỗ co thắt lớn nhất.

1.4.3. Các đặc tính khác

a. Tỷ suất thời gian

Trong khi nói chuyện, khoảng thời gian nói và khoảng thời gian nghỉ xen kẽ nhau. Tỷ lệ % thời gian nói trên tổng số thời gian nói và nghỉ được gọi là tỷ suất thời gian. Giá trị này biến đổi tuỳ thuộc vào tốc độ nói và từ đó ta có thể phân loại thành nói nhanh, nói chậm hay nói bình thường.

b. Hàm năng lượng thời gian ngắn

Hàm năng lượng thời gian ngắn của tiếng nói được tính bằng cách chia tín hiệu tiếng nói thành nhiều khung, mỗi khung chứa N mẫu. Các khung này được đưa qua một cửa sổ có dạng hàm như sau:

$$W(n) = \begin{cases} W(n) & \text{Với } 0 \le n \le N \\ 0 & \text{Với } n \ge N \end{cases}$$

Hàm năng lượng ngắn tại mẫu thứ m được tính theo công thức sau:

$$E_m = \sum_{n=0}^{N-1} \{x(n+m) * W(n)\}^2$$

Thông thường có ba dạng cửa sổ được sử dụng đó là cửa sổ Hamming, cửa sổ Hanning và cửa sổ chữ nhật. Hàm năng lượng thời gian ngắn của âm hữu thanh thường lớn hơn so với âm vô thanh.

c. Tần số cơ bản

Dạng sóng của tiếng nói gồm hai phần: Phần gần giống nhiễu (trong đó biên độ biến đổi ngẫu nhiên) và phần có tính chu kỳ (trong đó tín hiệu lặp lại gần như tuần hoàn). Phần tín hiệu có tính chu kỳ chứa các thành phần tần số có dạng điều hòa. Tần số thấp nhất chính là tần số cơ bản và cũng chính là tần số dao động của dây thanh.

Đối với những người nói khác nhau, tần số cơ bản cũng khác nhau. Dưới đây là một số giá trị tần số cơ bản tương ứng với giới tính và tuổi:

Giá trị tần số cơ bản	Người nói
80 – 200 Hz	Nam giới
150 – 450 Hz	Phụ nữ
200 – 600 Hz	Trẻ em

d. Formant

Với phổ của tín hiệu tiếng nói, mỗi đỉnh có biên độ lớn nhất xét trong một khoảng nào đó (cực đại khu vực) tương ứng với một formant. Ngoài tần số, formant còn được xác định bởi biên độ và dải thông. Về mặt vật lý các formant tương ứng với các tần số cộng hưởng của tuyến âm. Trong xử lý tiếng nói và nhất là trong tổng hợp tiếng nói, để mô phỏng lại tuyến âm người ta phải xác định được các tham số formant đối với từng loại âm vị, do đó việc đánh giá, ước lượng các formant có ý nghĩa rất quan trọng.

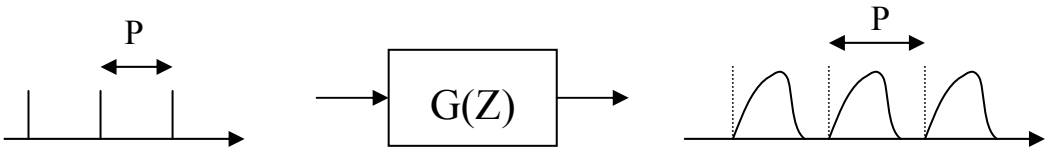
Tần số formant biến đổi trong một khoảng rộng phụ thuộc vào giới tính của người nói và phụ thuộc vào các dạng âm vị tương ứng với formant đó. Đồng thời, formant còn phụ thuộc các âm vị trước và sau đó. Về cấu trúc tự nhiên, tần số formant có liên hệ chặt chẽ với hình dạng và kích thước tuyến âm. Thông thường phổ của tín hiệu tiếng nói có khoảng 5 formant nhưng chỉ có 3 formant đầu tiên ảnh hưởng quan trọng đến các đặc tính của các âm vị, các formant còn lại cũng có ảnh hưởng song rất ít.

Tần số formant đặc trưng cho các nguyên âm biến đổi tuỳ thuộc vào người nói trong điều kiện phát âm nhất định. Mặc dù phạm vi của các tần số formant tương ứng với mỗi nguyên âm có thể trùm lên nhau nhưng vị trí giữa các formant là không đổi vì sự xê dịch của các formant là song song.

1.5. MÔ HÌNH TẠO TIẾNG NÓI

Nhằm đơn giản hoá việc phân tích và nghiên cứu bộ máy phát âm, người ta chia bộ máy phát âm ra làm hai phần cơ bản: nguồn âm và hệ thống đáp ứng.

- Hệ thống đáp ứng bao gồm thanh môn, tuyến âm, môi và mũi. Việc mô hình hoá này sử dụng hàm truyền đạt trong biến đổi Z.
- Đối với các âm hữu thanh, nguồn âm là một dạng sóng tuần hoàn đặc biệt. Dạng sóng này được mô phỏng bởi đáp ứng của bộ lọc thông thấp có hai điểm cực thực và tần số cắt vào khoảng 100 Hz.



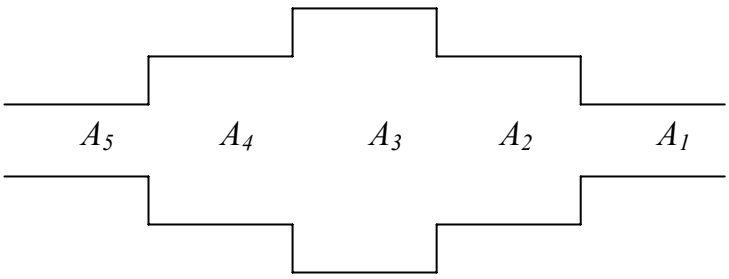
Hình 1.4. Mô Hình hoá nguồn âm đối với âm hữu thanh

$$G(Z) = \frac{A}{\left(1+\alpha z^{-1}\right)\left(1+\beta z^{-1}\right)}$$

Trong đó α,β là các hằng số đặc trưng cho nguồn âm với $\alpha<1$, $\beta<1$.

Đối với âm vô thanh nguồn âm là một nhiễu trắng với biên độ biến đổi gần như ngẫu nhiên.

Để tạo tiếng nói, người ta dùng các mô hình khác nhau để mô phỏng bộ máy phát âm. Theo quan điểm giải phẫu học, ta có thể giả thiết rằng tuyến âm được biểu diễn bằng một chuỗi *M* đoạn ống âm học lý tưởng, là những đoạn ống có độ dài bằng nhau, và từng đoạn riêng biệt có thiết diện mặt cắt là A_m (gọi tắt là thiết diện) khác nhau theo chiều dài đoạn ống. Tổ hợp thiết diện $\{A_m\}$ của các đoạn ống được chọn sao cho chúng xấp xỉ với hàm thiết diện *A(x)* của tuyến âm.



Hình 1.5. Chuỗi 5 đoạn ống âm học lý tưởng

Các đoạn ống được coi là lý tưởng khi:

- Độ dài mỗi đoạn đủ nhỏ so với bước sóng âm truyền qua nó được coi là sóng phẳng.
- Các đoạn đủ cứng sao cho sự hao tổn bên trong do dao động thành ống, tính dính và dẫn nhiệt không đáng kể.

Ngoài ra ta giả thiết thêm mô hình tuyến âm lúc này là tuyến tính và không nối với thanh môn, hiệu ứng của tuyến mũi được bỏ qua, ta sẽ có mô hình tạo tiếng nói lý tưởng và việc phân tích mô hình ống âm học trở nên phức tạp hơn. Tiếp theo chúng ta có thể thấy rằng mô hình này có nhiều tính chất chung với mạch lọc số nên nó có thể được biểu diễn bằng cấu trúc mạch lọc số với các tham số thay đổi phù hợp với sự thay đổi tham số của ống âm học.

Sự chuyển động của không khí trong một đoạn ống âm học có thể được mô tả bằng áp suất âm thanh và thông lượng, đó là những hàm phụ thuộc độ dài ống *(x)* và thời gian *(t)*. Trong những đoạn riêng biệt đó, các giá trị của hai hàm này được coi là tổ hợp tuyến tính các giá trị của chúng đối với sóng thuận và sóng ngược (được ký hiệu lần lượt bằng dấu cộng '+' và dấu trừ '-'). Sóng thuận là sóng truyền từ thanh môn đến môi, trong khi sóng ngược lại truyền lừ môi đến thanh môn. Nếu đoạn thứ *m* chúng ta xét có thiết diện A_m thì hàm thông lượng và hàm áp suất của đoạn này là:

$$u_m(x,t) = u_m^+\left(t - \frac{x}{c}\right) - u_m^-\left(t + \frac{x}{c}\right)$$

$$p_m(x,t) = \frac{\rho.c}{A_m}\left[u_m^+\left(t - \frac{x}{c}\right) + u_m^-\left(t + \frac{x}{c}\right)\right]$$

ở đây u_m^+, u_m^- là sóng thuận và sóng ngược

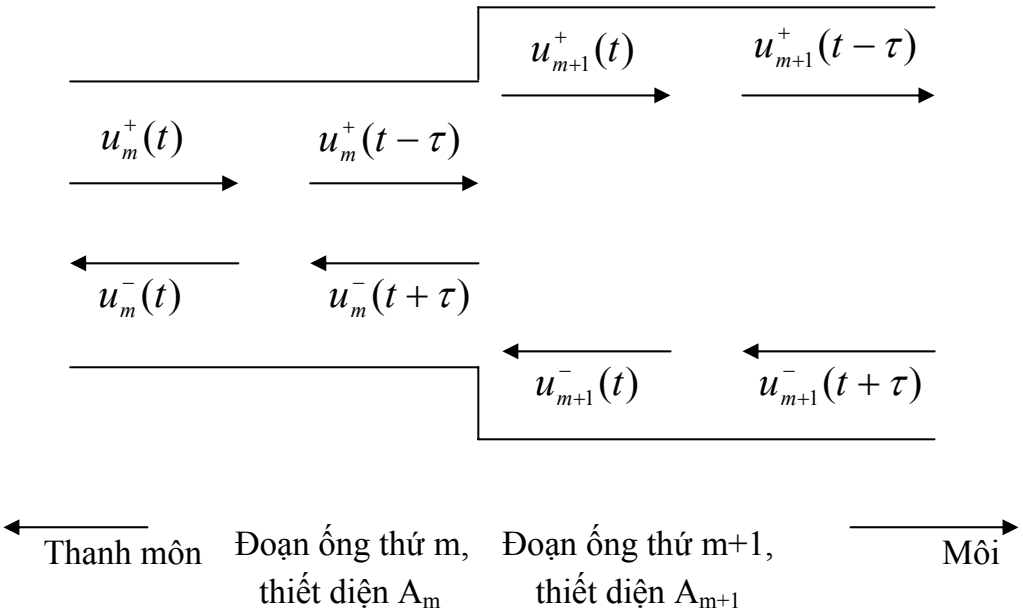
c là tốc độ âm thanh

ρ là mật độ không khí trong đoạn

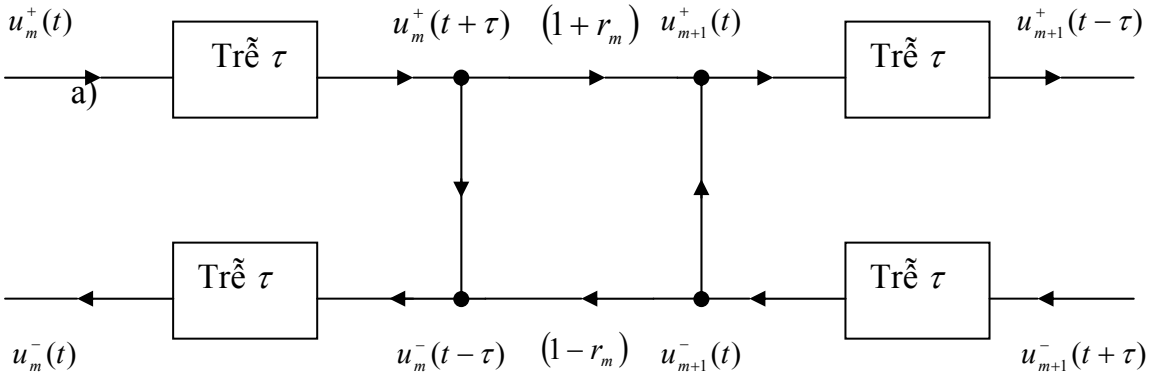
$x=0$ vị trí trung tâm của đoạn

Mối quan hệ giữa sóng thuận và sóng ngược trong những đoạn kế tiếp phải đảm bảo áp suất và thông lượng liên tục cả về thời gian và không gian tại mọi điểm trong hệ thống. Trong hình 1.6.a ta thấy khi sóng thuận trong một đoạn gặp phần thay đổi về thiết diện (mối nối giữa hai đoạn kế tiếp), một phần của nó truyền sang đoạn kế tiếp, một phần kia lại phản xạ dưới dạng sóng ngược. Hoàn toàn tương tự, khi sóng ngược gặp mối nối, một phần được chuyển tiếp sang đoạn trước đó, còn phần kia lại phản xạ lại dưới dạng sóng thuận.

.



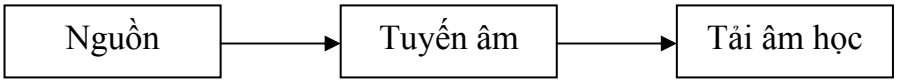
(a)



(b)

Hình 1.6 Cách biểu diễn lý học và toán học

a. Mô hình lý học giữa đoạn ống m và m+1
b. Mô hình toán học của đoạn ống thứ m



Hình 1.7. Mô hình số của hệ thống phát âm

Tuyến âm được coi như một chuỗi liên tiếp các ống âm học và được mô hình hoá bởi một chuỗi gồm *K* bộ cộng hưởng. Khi đó hàm truyền đạt của tuyến âm có dạng:

$$V(z) = \frac{B}{\prod_{i=1}^{K}\left(1 + b_{1i}z^{-1} + b_{2i}z^{-2}\right)}$$

Mỗi bộ cộng hưởng sẽ tạo ra một formant được đặc trưng bởi tần số trung tâm, tính theo công thức:

$$F_K = \frac{1}{2\pi} f_e \cos^{-1} \frac{-b_{1i}}{2\sqrt{b_{2i}}}$$

Với f_e là tần số lấy mẫu của tín hiệu lấy mẫu

Cuối cùng âm thanh được phát ra ở môi, nơi được coi như một tải âm học. Sự tán xạ của môi được biểu diễn bởi hàm truyền đạt:

$$R(z) = C\left(1 - z^{-1}\right)$$

Hàm truyền đạt của hệ thống có dạng:

$$T(z) = G(z).V(z).R(z)$$

Nếu giả thiết một trong hai điểm cực của thanh môn gần bằng 1(β = -1) ta có:

$$T(z) = \frac{C}{A(z)}$$

$$\text{Với } A(z) = \left(1 + \alpha z^{-1}\right)\prod_{i=1}^{K}\left(1 + b_{1i}z^{-1} + b_{2i}z^{-2}\right)$$

$$\text{Hay } A(z) = 1 + \sum_{i=1}^{2K+1}\alpha_i z^{-1}$$

là hàm truyền đạt của bộ lọc đảo. T(z) là hàm truyền đạt của mô hình toàn điểm cực. Các hệ số a_i của bộ lọc đảo sẽ là các tham số quan trọng trong phương pháp dự đoán tuyến tính để xác định các formant của tuyến âm.

Hạn chế của mô hình này là không thể tạo ra các âm xát hữu thanh và các âm mũi. Đối với các âm mũi mô hình trên được cải tiến bằng cách thêm vào phần đặc trưng cho mũi đặt song song với mô hình. Lúc đó hàm truyền đạt của hệ thống mới là:

$$\frac{\sigma_1}{A_1(z)} + \frac{\sigma_2}{A_2(z)} = \frac{\sigma_1 A_2(z) + \sigma_2 A_1(z)}{A_1(z)A_2(z)}$$

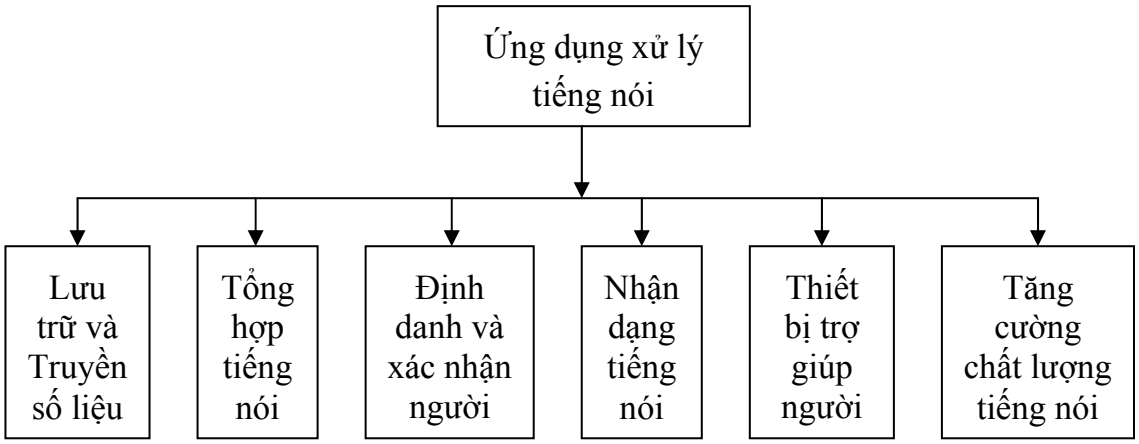
Hệ thống trên không còn là hệ thống toàn điểm cực mà nó còn xuất hiện các điểm không trong mặt phẳng Z. Việc xuất hiện các điểm không này sẽ gây khó khăn cho phương pháp tiên đoán tuyến tính là phương pháp áp dụng cho các hệ thống toàn điểm cực. Song người ta đã khắc phục được khó khăn trên bằng cách thay một điểm không bằng hai điểm cực theo phương pháp giảm bậc gần đúng, công thức giảm bậc như sau:

$$1-\alpha z^{-1} \approx \frac{1}{1+\alpha z^{-1}+\alpha^2 z^{-2}+...}$$

Tín hiệu âm thanh không phải là tín hiệu dừng, do đó mô hình phải được xây dựng một cách liên tục, nghĩa là các tham số của mô hình phải biến thiên theo thời gian. Sự biến thiên này rất chậm nên các tham số có thể coi như không đổi trong khoảng thời gian mà tín hiệu được coi là dừng: 20 ms.

1.6. XỬ LÝ TIẾNG NÓI

Dựa trên cơ sở lựa chọn các cách biểu diễn tín hiệu và phương pháp xử lý, đã có rất nhiều các ứng dụng quan trọng đã được triển khai. Hình vẽ dưới đây sẽ chỉ ra một số ứng dụng trong lĩnh vực xử lý tiếng nói.



Hình 1.8. Một vài ứng dụng xử lý tiếng nói

1.6.1. Tổng hợp tiếng nói

Tổng hợp tiếng nói là quá trình tạo ra tín hiệu âm thanh bằng cách điều khiển một mô hình mẫu với một tập các tham số. Nếu mô hình mẫu này và các tham số được xây dựng một cách hoàn hảo thì tiếng nói tổng hợp có thể giống với tiếng nói tự nhiên. Hiện có hai phương pháp tổng hợp tiếng nói:

a. Tổng hợp tiếng nói theo cách phát âm

Đây là cách tiếp cận trực tiếp để mô hình hoá hệ thống một cách chi tiết. Trong phương pháp này hệ thống tổng hợp được mô phỏng giống như quá trình tạo ra âm thanh và lan truyền âm thanh trong hệ thống phát âm của con người. Hướng nghiên cứu này vẫn đang tiếp tục và cho một số kết quả nhất định. Phương pháp này có thể tạo ra hầu hết các tiếng nói tự nhiên.

b. Tổng hợp đầu cuối tự nhiên

Theo hướng mô hình hoá này, người ta dựa trên các đặc tính đáp ứng tần số của dây thanh và tuyến âm để mô phỏng lại cơ chế tạo tiếng nói. Mô hình này gọi là mô hình nguồn-lọc. Bộ tổng hợp tiếng nói theo hướng này được thực hiện bằng cách sử dụng hệ thống tương tự với cơ chế tạo tiếng nói tại những điểm quan sát.

Cơ quan phát âm được mô hình hoá thành một hệ thống bao gồm một nguồn âm biểu diễn cho thanh môn và một bộ lọc biểu diên cho tuyến âm. Quá trình tổng hợp sẽ bao gồm hai phần cơ bản:

- Tổng hợp tín hiệu nguồn dựa vào tần số cơ bản và tính chất tuần hoàn của nguồn.
- Xây dựng lại hàm truyền đạt của tuyến âm (bao gồm cả mũi và miệng) dựa vào các tham số đặc trưng cho tuyến âm.

Hiện nay người ta thường sử dụng hai bộ tham số đặc trưng cho tuyến âm:

- Bộ tham số formant
- Bộ tham số của bộ lọc đảo

Các bộ tham số này có thể được tổng kết từ các quá trình phân tích tiếng nói.

1.6.2. Nhận dạng tiếng nói

Nhận dạng tiếng nói là lĩnh vực nghiên cứu với mục đích tạo ra được một thiết bị, máy móc hoặc phần mềm có khả năng nhận biết một cách chính xác tiếng nói của con người từ bất kỳ một nguồn phát âm nào. Nhận dạng tiếng nói có hai ứng dụng chính là nhận dạng tiếng nói và nhận dạng người nói.

a. Nhận dạng ngữ nghĩa

Thông thường để điều khiển các thiết bị máy móc người ta thường sử dụng cách giao tiếp thông qua sự vào ra cơ khí. Khi áp dụng tiếng nói vào giao tiếp, lợi ích của nó có thể dễ dàng nhận thấy: đó là tính tiện lợi, dễ sử dụng, tốc độ giao tiếp cao... Để có thể sử dụng tiếng nói như một công cụ giao tiếp thì hệ thống cần có khả năng tiếng nói về ngữ nghĩa. Nhận dạng ngữ nghĩa bao gồm nhận dạng từ và nhận dạng câu.

b. Nhân dạng người nói

Trong thế giới ngày nay tồn tại nhiều hệ thống yêu cầu độ an toàn bảo mật cao. Từ đó nảy sinh ra yêu cầu phải nhận dạng được người nói bằng những đặc điểm riêng biệt mà không ai có thể sao chép được. Bên cạnh các cách thức nhận dạng qua chữ ký, ảnh chân dung, chữ viết..., ngày nay người ta còn dùng

tiếng nói để nhận dạng bởi vì tiếng nói có những đặc tính riêng biệt với từng người. Tại một số công ty đã xuất hiện những hệ thống kiểm tra người qua cửa bằng nhận dạng tiếng nói hoặc nhận dạng mỗi người qua thẻ nhận dạng mà những thông tin lưu trữ trên thẻ chính là đặc điểm về tiếng nói của người đó.

Nguyên tắc của nhận dạng người nói là sử dụng những từ khoá đã được xác định từ trước mà những từ khoá này đặc trưng cho từng người một. Có hai yếu tố để khẳng định sự khác nhau trong tiếng nói của mỗi người:

- Các đặc tính cơ quan phát âm khác nhau như: độ dài của tuyến âm, tần số cộng hưởng của dây thanh, các tần số formant, dải thông, sự biến đổi của đường bao phổ... Đó là tập hợp những đặc tính có liên quan đến tính độc lập của nội dung âm vị của từ ngữ.
- Sự khác nhau trong cách phát âm của từng người: tốc độ và chiều dài từ luôn luôn khác nhau.

Trong tất cả các đặc tính trên đường bao phổ và tần số cơ bản là hai đặc tính quan trọng nhất. Đường bao phổ được miêu tả bằng những giá trị trung bình của các bộ lọc thông dải, của các tần số formant, của các hệ số tiên đoán tuyến tính, của hệ số cepstre và các tham số khác.

CHƯƠNG 2

TỔNG HỢP TIẾNG NÓI

2.1. CÁC PHƯƠNG PHÁP TỔNG HỢP TIẾNG NÓI

Tổng hợp tiếng nói là phát sinh tiếng nói từ sóng tiếng nói. Trong vài thập niên gần đây, các bộ tổng hợp tiếng nói có chất lượng ngày càng cao.Tuy nhiên chất lượng của các phương pháp hiện nay mới chỉ đạt đến mức phù hợp cho một vài ứng dụng, chẳng hạn như đa phương tiện và truyền thông.

Hiện nay có ba phương pháp tổng hợp tiếng nói. Phương pháp đơn giản nhất để phát sinh tiếng nói tổng hợp là phát các mẫu tiếng nói đã thu từ tiếng nói tự nhiên (như các từ hoặc câu). Phương pháp này cho chất lượng tương đối tốt nhưng gặp phải hạn chế là số lượng từ vựng trong cơ sở dữ liệu rất lớn. Bên cạnh đó tiếng nói cũng có thể tạo ra bằng cách mô phỏng hệ thống phát âm. Phương pháp này cho chất lượng rất tốt nhưng thực hiện khá phức tạp. Một phương pháp nữa cũng được dùng để tổng hợp tiếng nói là tổng hợp formant. Các phương pháp tổng hợp tiếng nói cùng với những đặc điểm cơ bản nhất sẽ được giới thiệu trong phần tiếp theo.

2.1.1. Phương pháp mô phỏng hệ thống phát âm

Phương pháp mô phỏng hệ thống phát âm (articulatory synthesis) cố gắng mô phỏng hệ thống phát âm của con người một cách hoàn hảo nhất, do đó có thể đạt tới chất lượng cao trong tổng hợp tiếng nói. Nhưng cũng chính vì vậy mà phương pháp này khó có thể thực hiện được, vì việc mô phỏng hệ thống phát âm của con người rất khó thực hiện.

Sau khi phương pháp tổng hợp Formant ra đời thì phương pháp mô phỏng hệ thống phát âm ít khi được sử dụng trong các hệ thống. Nhưng từ khi có sự xuất hiện của máy tính thì nó lại được phát triển.

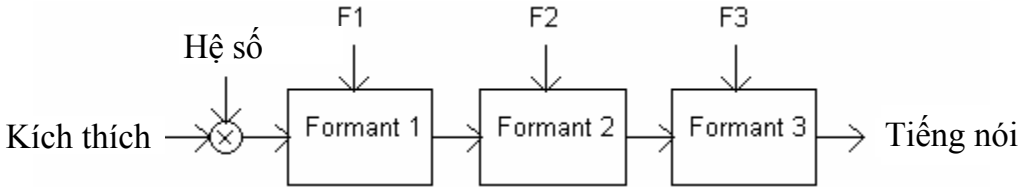
2.1.2. Phương pháp tổng hợp Formant

Phương pháp tổng hợp formant (formant synthesis) yêu cầu phải tổng hợp được tối thiểu 3 formant để hiểu được tiếng nói, và để có được tiếng nói chất lượng cao thì cần tới 5 formant. Tiếng nói được tạo ra từ các bộ tổng hợp

formant với thành phần chính là các bộ cộng hưởng. Tuỳ theo cách bố trí các bộ cộng hưởng mà ta có bộ tổng hợp formant là nối tiếp hay song song.

a. Bộ tổng hợp formant nối tiếp

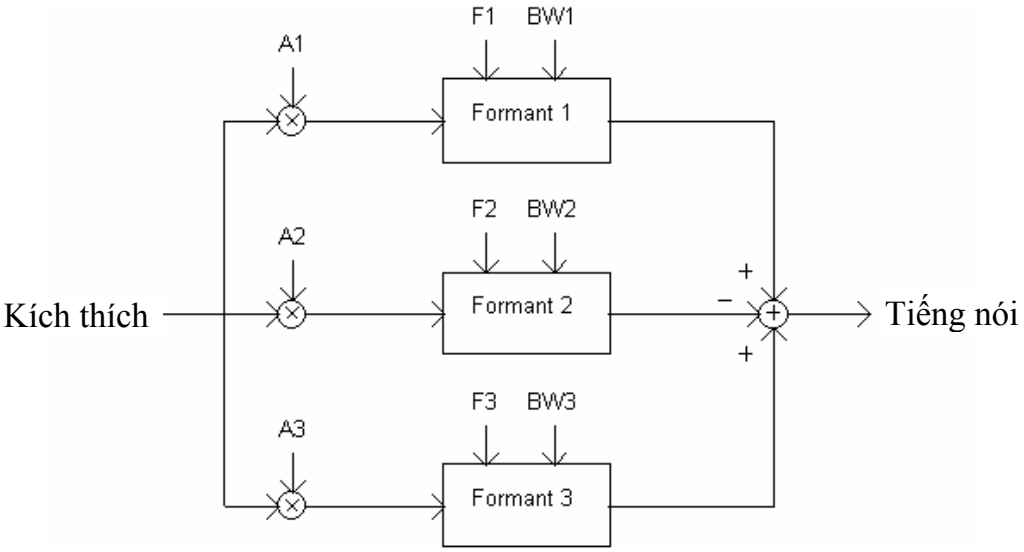
Bộ tổng hợp formant nối tiếp là một bộ tổng hợp formant có các tầng nối tiếp, đầu ra của bộ cộng hưởng này là đầu vào của bộ cộng hưởng kia.



Hình 2.1. Cấu trúc cơ bản của một bộ tổng hợp formant nối tiếp

b. Bộ tổng hợp formant song song

Bộ tổng hợp formant song song bao gồm các bộ cộng hưởng mắc song song. Đầu ra là kết hợp của tín hiệu nguồn và tất cả các formant. Cấu trúc song song cần nhiều thông tin để điều khiển hơn.



Hình 2.2. Cấu trúc cơ bản của một bộ tổng hợp formant song song

Tổng hợp formant là một phương pháp tổng hợp cho chất lượng chấp nhận được nhưng nếu yêu cầu chất lượng cao thì phương pháp này chưa đáp ứng được.

2.1.3. Phương pháp ghép nối

Tổng hợp bằng cách ghép nối các âm được tổng hợp từ các lời nói tự nhiên đã được thu từ trước có lẽ là cách dễ nhất để sản sinh lời nói. Phương pháp tổng hợp ghép nối cho chất lượng cao và tương đối tự nhiên. Phương pháp này rất phù hợp với các hệ thống phát thanh và các hệ thống thông tin. Tuy nhiên phương pháp này thường chỉ áp dụng cho một giọng và phải sử dụng nhiều bộ nhớ hơn các phương pháp khác do số lượng từ vựng rất lớn. Để khắc phục nhược điểm này người ta xây dựng các phương pháp tổng hợp ghép nối từ những đơn vị nhỏ như âm vị, âm tiết, diphone (âm vị kép)... Ngoài các diphone, chúng ta còn sử dụng triphone, tetraphone hay syllable, demisyllable, nhưng chủ yếu vẫn là các diphone, được thu từ tiếng nói tự nhiên. Các diphone được cắt ra từ tín hiệu rồi sau đó được tổng hợp lại theo yêu cầu dựa trên một thuật toán ghép nối.

Phương pháp này có một số khác biệt so với các phương pháp khác:

- Xuất hiện sự biến dạng của tiếng nói tổng hợp do tính không liên tục của việc ghép nối các diphone với nhau. Vì vậy phải sử dụng biện pháp làm trơn tín hiệu.
- Bộ nhớ yêu cầu cao, nhất là khi các đơn vị kết nối dài như là các âm vị hay các từ.
- Sưu tầm và gắn nhãn dữ liệu tiếng nói cần nhiều thời gian và công sức. Về lý thuyết tất cả các mẫu cần phải được lưu trữ. Số lượng và chất lượng các mẫu lưu trữ là một vấn đề cần giải quyết khi tiến hành lưu trữ.

Hiện nay phương pháp này đang được sử dụng rộng rãi trên thế giới và ngày càng cho chất lượng tốt hơn nhờ sự trợ giúp của máy tính.

Phần tiếp theo sẽ giới thiệu về một phương pháp tổng hợp ghép nối được áp dụng phổ biến cho tín hiệu tiếng nói, phương pháp ghép nối dựa trên giải thuật PSOLA.

a. Phương pháp tổng hợp PSOLA

PSOLA (Pitch Synchronous Overlap Add) là phương pháp tổng hợp dựa trên sự phân tích một tín hiệu thành một chuỗi các tín hiệu thành phần. Khi cộng xếp chồng (overlap-add) các tín hiệu thành phần ta có thể khôi phục lại tín hiệu ban đầu.

PSOLA thao tác trực tiếp với tín hiệu dạng sóng, không dùng bất cứ loại mô hình nào nên không làm mất thông tin của tín hiệu. PSOLA cho phép điều khiển độc lập tần số cơ bản, chu kỳ cơ bản và các formant của tín hiệu. Ưu điểm chính của phương pháp PSOLA là giữ nguyên đường bao phổ khi thay

đổi tần số cơ bản (pitch shifting). Phương pháp này cho phép biến đổi tín hiệu ngay trên miền thời gian nên chi phí tính toán rất thấp. PSOLA đã được dùng rất phổ biến với tín hiệu tiếng nói.

b. Các phiên bản của PSOLA

Dựa trên PSOLA, người ta đã đưa ra nhiều phiên bản khác nhau, dưới đây là các phiên bản chính:

- **TD-PSOLA**

Phương pháp TD-PSOLA (Time Domain- Pitch Synchronous Overlap Add) là phiên bản miền thời gian của PSOLA (TD-PSOLA). Phương pháp này thao tác với tín hiệu trên miền thời gian nên được sử dụng nhiều vì hiệu quả trong tính toán của nó. Phương pháp này sẽ được trình bày chi tiết trong chương tiếp theo.

- **FD-PSOLA**

Phương pháp tổng hợp FD-PSOLA (Frequency Domain- Pitch Synchronous Overlap Add) là phương pháp bao gồm các bước giống như TD-PSOLA nhưng thao tác trên miền tần số. Phương pháp này có chi phí tính toán cao hơn TD-PSOLA. Đối với mỗi trường hợp riêng biệt thì mỗi phương pháp sẽ cho hiệu quả khác nhau, nên phải dựa vào từng hoàn cảnh để chọn phương pháp thích hợp.

- **LP-PSOLA**

Ngoài các phương pháp trên miền thời gian, miền tần số, còn có một phương pháp gọi là phương pháp dự đoán tuyến tính (Linear Prediction - Pitch Synchronous Overlap Add). Phương pháp dự đoán tuyến tính được thiết kế để mã hoá tiếng nói nhưng phương pháp này cũng có thể dùng cho tổng hợp.

Cơ sở của phương pháp dự đoán tuyến tính dựa trên các mẫu $y(n)$ có thể lấy xấp xỉ hoặc dự đoán từ p mẫu trước đó $y(n\text{-l})$ đến $y(n\text{-}p)$ với sai số nhỏ nhất. Như vậy:

$$y(n) = e(n) + \sum_{k=1}^{p} a(k) y(n-k)$$

và: $$e(n) = y(n) - \sum_{k=1}^{p} a(k) y(n-k) = y(n) - \tilde{y}(n)$$

Với $\tilde{y}(n)$ là giá trị dự đoán, p là thứ tự dự đoán tuyến tính, $a(k)$ là hệ số dự đoán tuyến tính được tìm bằng cách lấy min tổng bình phương của các khung lỗi.

Tín hiệu kích thích được lấy xấp xỉ bằng một dãy các tín hiệu tiếng nói và nhiễu ngẫu nhiên. Tín hiệu nguồn được cho qua bộ lọc số với hệ số a(k).

Phương pháp LP-PSOLA cho kết quả chưa tốt. Người ta đã cải biến phương pháp này để thu được chất lượng tốt hơn, mà đại diên là phương pháp

WLP (Warped Linear Prediction). Ý tưởng cơ bản là thay thế các đơn vị trễ trong bộ lọc số bởi các đoạn sau:

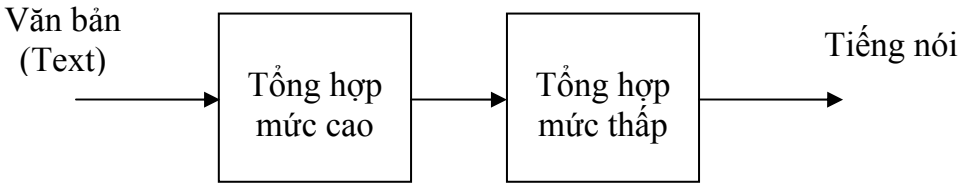
$$\tilde{z}^{-1} = D_1(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}$$

Với z là tham số cong nằm trong khoảng [-1,1] và $D_1(z)$ là nhân tử cong trễ với $\lambda = 0.63$ tại tần số lấy mẫu là 22 kHz. WLP đưa ra cách giải quyết tốt hơn cho tần số cao và tồi hơn cho tần số thấp.

2.2. MÔ HÌNH TỔNG HỢP TIẾNG NÓI TỪ VĂN BẢN

Một nhu cầu rất quan trọng trong lĩnh vực tổng hợp tiếng nói là tổng hợp tiếng nói từ văn bản (Text To Speech – TTS). Quá trình này được chia làm hai mức xử lý:

- High Level Synthesis: Tổng hợp mức cao
- Low Level Synthesis: Tổng hợp mức thấp



Hình 2.3. Mô hình tổng hợp tiếng nói

2.2.1. Tổng hợp mức cao

Tổng hợp mức cao là giai đoạn đầu của quá trình tổng hợp, giai đoạn chuyển đổi các văn bản text thành các đơn vị tiếng nói (ví dụ như diphone). Văn bản được nhập hoặc sao chép vào, sau đó qua tổng hợp mức thấp sẽ thành tiếng nói.

Tổng hợp mức cao gồm 3 bước:

- Xử lý trước văn bản với các chữ số, các ký tự đặc biệt, chữ viết tắt, và những từ viết tắt được ghép bằng các chữ đầu của các từ đầy đủ...
- Phân tích cách phát âm của từ, kể cả từ đồng âm khác nghĩa và các tên riêng.
- Phân tích ngữ điệu của tiếng nói.

Sau khi tổng hợp mức cao, thông tin được cung cấp cho hệ thống mức thấp để điều khiển. Chẳng hạn, với bộ tổng hợp formant thì cần các thông tin như tần số cơ bản, tần số formant, khoảng thời gian, và biên độ của mỗi đoạn âm thanh.

a. Xử lý văn bản

Nhiệm vụ đầu tiên của tất cả các hệ thống TTS là chuyển đổi dữ liệu (mẫu) về dạng thích hợp cho một bộ tổng hợp.Trong giai đoạn này tất cả các đặc tính như chữ cái, chữ số, chữ viết tắt... phải được chuyển đổi theo một khuôn dạng rõ ràng, đầy đủ. Để xử lý văn bản, người ta dùng những bảng đối chiếu một - một đơn giản. Trong một số trường hợp còn cần thêm thông tin bổ sung (ví dụ những từ gần nghĩa, những ký hiệu...). Điều này có thể dẫn đến một cơ sở dữ liệu khá lớn và tập luật phức tạp, đó sẽ là những vấn đề cần giải quyết khi thực hiện với các hệ thống thời gian thực.

Ví dụ:

- Văn bản đầu vào có thể chứa các từ viết tắt phải được hiểu như nhau trong tất cả các hoàn cảnh. Nhưng sự chuyển đổi từ viết tắt không phải lúc nào cũng dựa trên cách viết tắt mà phải dựa trên cả một cụm viết tắt (Ví dụ: tiếp đầu ngữ M trong ngữ cảnh nào đó được hiểu mega, nhưng viết MTV không thể chuyển thành megaTV).
- Tương tự như vậy, việc chuyển đổi chữ số cũng không đơn giản. Chữ số được sử dụng trong với nhiều vai trò như là số, là ngày tháng, giá trị đo đạc, và trong những biểu thức toán học. Những số nằm giữa 1100 và 2002 thông thường được chuyển đổi thành năm. 1/1/1111 chữ số trong mẫu trên thường được chuyển đổi thành ngày/tháng/năm. Nhưng 2/5 thì thật khó bởi vì nó có thể vừa là ngày/tháng vừa có thể là một phân số.

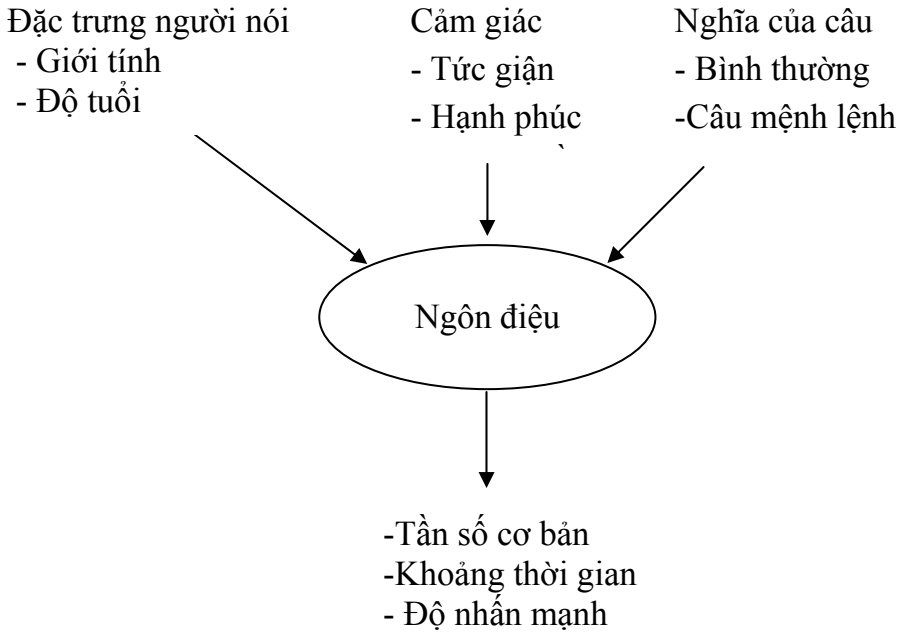
b. Phân tích cách phát âm

Với các ngôn ngữ trên thế giới mà việc phát âm không hoàn toàn tuân theo quy tắc (ví dụ như tiếng Anh) thì phát âm đúng các từ là một vấn đề khó trong tổng hợp tiếng nói. Đặc biệt với một ứng dụng điện thoại thì hầu hết các từ đều là tên hoặc là địa chỉ các đường phố và để đọc đúng những tên này là điều không dễ dàng. Một phương pháp giải quyết là có thể lưu vào một bảng phát âm đặc biệt, nhưng số lượng sẽ rất lớn. Vì vậy phương pháp trên không hiệu quả. Lúc này việc tạo ra các luật cơ bản để xây dựng nên một từ điển các từ với các luật chuyển từ sang âm vị (letter-to-phoneme) sẽ hợp lý hơn. Cách tiếp cận này cũng phù hợp với phát âm bình thường. Khi phân tích, một từ có thể được chia thành các phần độc lập bao gồm tiền tố, gốc từ, phụ tố.

c. Ngôn điệu

Xác định đúng được ngữ điệu, trọng âm và khoảng thời gian từ văn bản viết có lẽ là những vấn đề khó khăn nhất trong những năm tới. Các đặc tính này được gọi là ngôn điệu hoặc những đặc tính siêu đoạn và có thể được xem xét

như giai điệu, nhịp điệu và sự nhấn mạnh của tiếng nói ở mức cảm giác. Ngữ điệu có nghĩa là sự thay đổi của tần số cơ bản trong thời gian nói. Ngôn điệu của tiếng nói liên tục phụ thuộc vào nhiều yếu tố như nghĩa của các câu, đặc trưng và cảm xúc của người nói. Ngôn điệu phụ thuộc được mô tả ở hình 2.4.



Hình 2.4. Sự phụ thuộc của ngôn điệu vào các yếu tố

2.2.2. Tổng hợp mức thấp

Tổng hợp mức thấp là quá trình kết hợp các đoạn tín hiệu (ví dụ như diphone). Các đoạn tín hiệu này đã được phân tích, xử lý qua mức cao (xử lý văn bản, ngữ điệu).

Đối với phương pháp tổng hợp bằng cách mô phỏng hệ thống phát âm của con người thì sự chọn lựa dữ liệu và thực thi các luật là rất phức tạp. Hầu như không thể mô phỏng dưới dạng mô hình khối, sự chuyển động của lưỡi... một cách hoàn hảo. Lúc này, sự có mặt của máy tính đã trợ giúp một phần đáng kể.

Với tổng hợp formant thì tập luật để điều khiển tần số cơ bản, biên độ và đặc trưng của tín hiệu nguồn lại rất lớn. Vì vậy làm mất đi tính tự nhiên vốn có. Đặc biệt, âm mũi được xem là một vấn đề lớn đối với tổng hợp formant.

Còn với tổng hợp ghép nối thì việc thu thập các mẫu tín hiệu và gán nhãn mất rất nhiều thời gian, và có thể làm cho cơ sở dữ liệu rất lớn. Tuy nhiên số lượng dữ liệu có thể giảm xuống đáng kể nếu sử dụng những phương pháp nén dữ liệu thích hợp. Bên cạnh đó sự không đồng bộ các điểm ghép nối cũng có thể làm tín hiệu tổng hợp bị méo. Đối với những đơn vị ghép nối dài như từ

hoặc âm vị thì hiệu quả kết hợp là một vấn đề, ngoài ra bộ nhớ và hệ thống cũng là một khó khăn cần giải quyết.

2.3. SO SÁNH CÁC PHƯƠNG PHÁP TỔNG HỢP TIẾNG NÓI

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

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

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

Với mục đích nghiên cứu việc tổng hợp tiếng Việt và dựa trên những đặc điểm của các phương pháp tổng hợp, báo cáo này sẽ sử dụng phương pháp tổng hợp bằng ghép nối cho tiếng Việt. Trong số những phương pháp dùng để tổng hợp bằng ghép nối thì TD-PSOLA là phương pháp được sử dụng rộng rãi nhất với ưu điểm là chi phí tính toán thấp và giữ nguyên được nhiều thông tin trong

tiếng nói do thao tác trực tiếp với tín hiệu trên miền thời gian. Các chương tiếp theo sẽ trình bày chi tiết về phương pháp tổng hợp tiếng nói TD-PSOLA và áp dụng để xây dựng một chương trình tổng hợp tiếng Việt bằng diphone.

CHƯƠNG 3

GIẢI THUẬT TD-PSOLA

3.1. GIẢI THUẬT PSOLA

Như đã đề cập trong chương trước, người ta có thể tổng hợp tiếng nói theo nhiều phương pháp như mô phỏng hệ thống phát âm của con người, tổng hợp formant và tổng hợp ghép nối. Mỗi phương pháp đều có những ưu, nhược điểm riêng. Phương pháp mô phỏng hệ thống phát âm của con người cho chất lượng tốt, song rất khó mô phỏng một cách hoàn hảo bộ máy phát âm. Phương pháp tổng hợp bằng formant lại không cho chất lượng cao. Trong ba phương pháp này thì tổng hợp tiếng nói bằng ghép nối được sử dụng rộng rãi hơn cả.

PSOLA là giải thuật dùng cho phương pháp ghép nối. Trước hết tiếng nói được phân tích thành các tín hiệu thành phần, sau đó, khi cộng xếp chồng các thành phần này ta sẽ được tín hiệu tiếng nói tổng hợp. Phương pháp này thao tác trực tiếp với tín hiệu trên miền thời gian nên có chi phí tính toán thấp. Người ta kéo dãn thời gian trong tín hiệu tổng hợp bằng cách lặp lại các đoạn tín hiệu thành phần.

PSOLA có thể hiểu như sau:

- Tổng hợp tín hiệu từ các thành phần, trong đó mỗi thành phần có một tần số cơ bản.
- Tổng hợp dựa trên mô hình nguồn-lọc (source-filter).

Với phương pháp này tín hiệu phải điều hoà (harmonic) và phải thích hợp cho việc phân tích thành các tín hiệu thành phần khi sử dụng cửa sổ, điều này có nghĩa là năng lượng của tín hiệu phải tập trung xung quanh một khoảng thời gian nào đó trong mỗi chu kỳ.

3.1.1. Phân tích PSOLA

Phân tích PSOLA bao gồm việc phân tích một tín hiệu $s(t)$ thành các tín hiệu thành phần $s_i(t)$ bằng cách sử dụng cửa sổ $h(t)$:

$$s_i(t) = h(t - m_i)s(t)$$

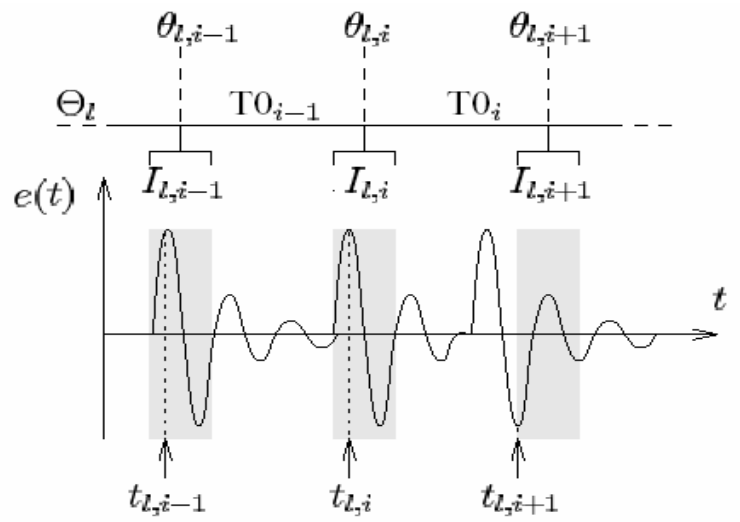
trong đó m_i được gọi là các điểm mốc (markers) phải thoả mãn các điều kiện sau:

- $m_i - m_{i-1}$ phải gần với chu kỳ cơ bản.
- Phải gần với điểm có biên độ cực đại (maxima energy). Điều kiện này được đưa ra để tránh làm hỏng tín hiệu khi lấy cửa sổ.

Sau khi tìm được chu kỳ cơ bản $T0(t)$ và hàm năng lượng $e(t)$, các điểm mốc m_i sẽ được xác định theo hai bước sau:

a. Bước 1: Tìm cực đại địa phương của hàm năng lượng.

Vì các điểm mốc phải gần các điểm có năng lượng cực đại nên bước đầu tiên là tìm các cực đại này. Xét vector $\theta_l = [\theta_{l,0}, \theta_{i,1}, ..., \theta_{l,i}, ...]$, trong đó $\theta_{l,i} - \theta_{l,i-1} = T0_{i-1}$. Xung quanh thời điểm $\theta_{l,i}$ xét khoảng thời gian $I_{l,i} = \left[\theta_{l,i} - \frac{T0_{i-1}}{\alpha}, \theta_{l,i} + \frac{T0_{i-1}}{\alpha}\right]$, ở đây α được gọi là độ mở rộng (extent). Trong mỗi khoảng $I_{l,i}$ gọi thời điểm có năng lượng lớn nhất là $t_{l,i}$. Với vector θ_L, tính tổng giá trị năng lượng tại các thời điểm $t_{l,i}$: $\sigma_l = \sum_i e(t_{l,i})$. Cuối cùng, chọn ra bộ $\tau_i = t_{l',i}$ mà tại đó σ_l đạt cực đại.



Hình 3.1. Xác định cực đại địa phương của hàm năng lượng

b. Bước 2: Tối ưu tính tuần hoàn và năng lượng cực đại.

Hai tiêu chuẩn này phải được tối ưu đồng thời vì các điểm mốc m_i vừa phải đồng bộ với tần số cơ bản vừa phải gần với các điểm có năng lượng cực đại. Có thể dùng giải thuật bình phương nhỏ nhất để tối ưu:

Gọi m_i là các điểm mốc phải tìm. τ_i là giá trị vừa tìm được trong bước 1, $T0_i$ là chu kỳ cơ bản ứng với τ_i. Dùng giải thuật bình phương nhỏ nhất để tìm

m_i sao cho $m_i - m_{i-1} \approx T0_{i-1}$ và $m_i \approx \tau_i$. Hàm phải tìm cực tiểu bây giờ sẽ là:

$$\varepsilon = \sum_i ((m_i - m_{i-1}) - T0_{i-1})^2 + \beta(m_i - \tau_i)^2$$

Gọi $\overline{m} = [m_0, m_1, ..., m_i, ...m_{N-1}, m_N]^T$, khi đó $\overline{m}$ được xác đinh như sau:

$$\overline{m} = M^{-1} \begin{pmatrix} 0 & -T0_0 & +\gamma\tau_0 \\ T0_0 & -T0_1 & +\beta\tau_1 \\ & & \\ T0_{N-2} & -T0_{N-1} & +\beta\tau_{N-1} \\ T0_{N-1} & 0 & +\gamma\tau_N \end{pmatrix},$$

trong đó M là một ma trận tam giác với đường chéo chính có dạng $[1+\gamma \quad 2+\beta \quad ... \quad 2+\beta \quad 1+\gamma]$, tam giác trên và dưới có dạng $[-1 \quad -1 \quad ... \quad -1 \quad -1]$.

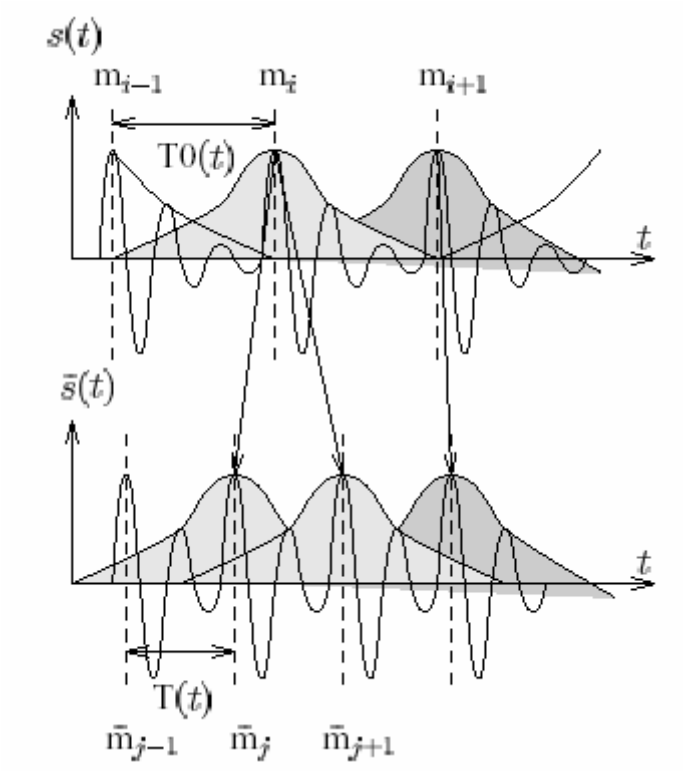
3.1.2. Tổng hợp PSOLA

Tổng hợp PSOLA được thực hiện bằng cách cộng xếp chồng các tín hiệu thành phần $s_i(t)$ được sắp xếp theo các thời điểm $\overline{m_j}$

$$\begin{cases} \overline{s_j}(t) = s_i(t + m_i) \\ \overline{s}(t) = \sum_j \overline{s_j}(t - \overline{m}_j) \end{cases}$$

ở đây m_i là các điểm mốc gần nhất với tín hiệu vào.

Chu kỳ cơ bản được điều chỉnh từ $T0(t)$ tới $T(t)$ bằng cách thay đổi khoảng cách giữa các đoạn tín hiệu liên tiếp $\overline{m}_j - \overline{m}_{j-1} = T(t)$. Với PSOLA việc co dãn trên miền thời gian được thực hiện bằng cách lặp lại các đoạn tín hiệu.



Hình 3.2. Cộng xếp chồng các đoạn tín hiệu

Tuy nhiên, khi thời gian được kéo giãn nhiều bằng cách lặp lại các tín hiệu thành phần có thể làm cho tín hiệu tổng hợp không liên tục.Giải thuật TD – PSOLA (Time Domain PSOLA) được trình bày ở phần tiếp theo sẽ khắc phục nhược điểm này. Hiện nay TD-PSOLA còn được mở rộng để sử dụng cho các phương pháp tổng hợp ghép nối khác, bởi vì nó là phương pháp tổng hợp chất lượng cao và chạy tốt ở cả những máy tính tốc độ thấp (tổng hợp thời gian thực có thể được thực hiện với bộ vi xử lý Intel 386).

3.2. GIẢI THUẬT TD-PSOLA

Giả sử rằng $s(n)$ là tín hiệu tuần hoàn , $\tilde{s}(n)$ là tín hiệu $s(n)$ sau khi đã thay đổi tần số bằng cách lấy tổng của các khung OLA của $s_i(n)$. $w(n)$ là cửa sổ, sự thay đổi chu kỳ tần số gốc T_0 tới chu kỳ tần số T tạo ra sự thay đổi của $s_i(n), \tilde{s}(n)$:

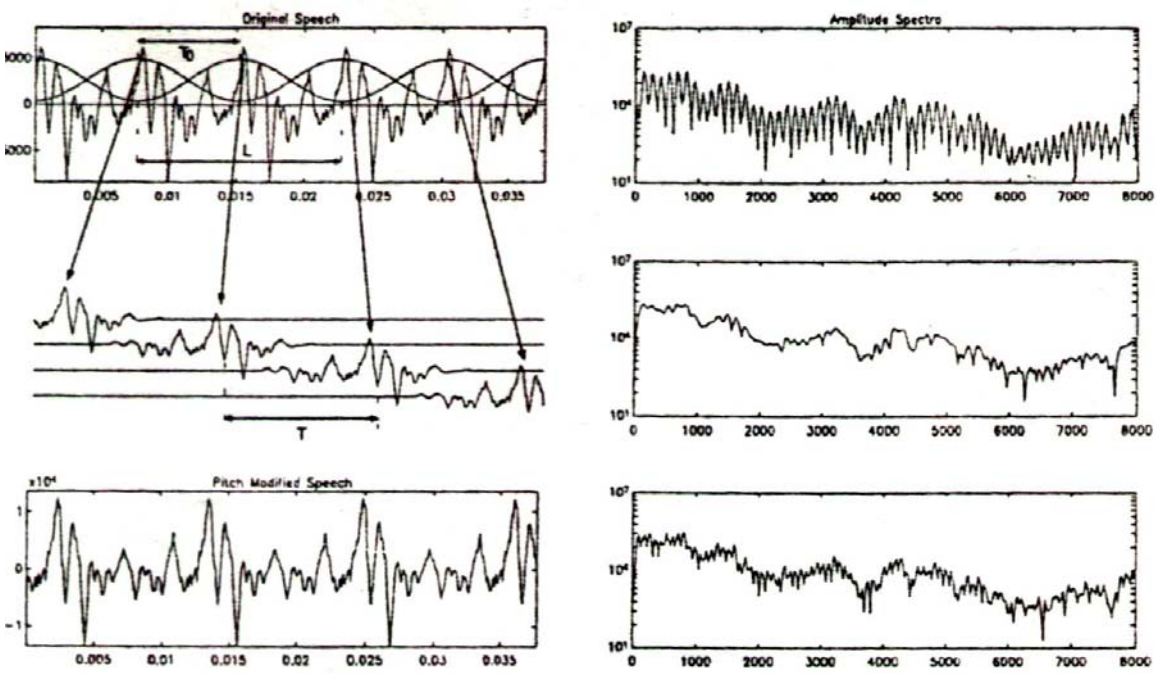
$$s_i(n) = s(n)w(n - iT_0)$$

$$\tilde{s}(n) = \sum_{i=-\infty}^{\infty} s_i(n - i(T - T_0))$$

Nếu $T \neq T_0$ thì ta phải làm hài hoà lại $s_i(n)$ với tần số cơ bản là $\frac{1}{T}$:

Nếu $s_i(n) \xleftrightarrow{\Im} S_i(\omega)$ thì $\tilde{s}(n) \xleftrightarrow{\Im} \frac{2\pi}{T}\sum_{i=-\infty}^{\infty} S_i\left(i\frac{2\pi}{T}\right)$

Công thức trên rất hiệu quả khi muốn thay đổi tần số của tín hiệu tuần hoàn.



Hình 3.3. Quá trình làm thay đổi tần số của tín hiệu

Nếu $T=T_0$ và cửa sổ phân tích đủ hẹp, tín hiệu tổng hợp gần như trùng với tín hiệu gốc

$$\tilde{s}(n) = \sum_{i=-\infty}^{\infty} s(n)w(n-iT) = s(n)\hat{w}(n) = Ks(n)$$

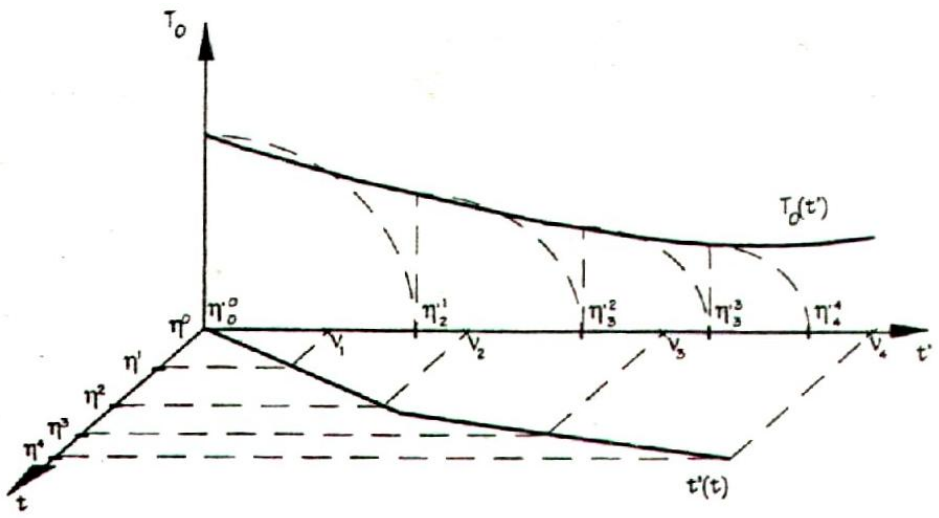
Trong trường hợp đặc biệt với cửa sổ tam giác thì kích thước của cửa sổ được chọn bằng 2 lần chu kỳ cơ bản, khi đó dấu gần đúng của biểu thức trên sẽ tịnh tiến tới dấu bằng với K=1.

Rõ ràng, với giải thuật này,cơ sở dữ liệu phải được lưu trữ dưới dạng danh sách tham số (một danh sách cho mỗi đoạn). Trên thực tế, đối với TD-PSOLA, thì các danh sách này chứa chuỗi các điểm mốc (markers) là tâm các cửa sổ xếp chồng η^i. Vị trí các điểm mốc này được sắp xếp đồng bộ với tần số cơ bản của phần hữu thanh trong đoạn tín hiệu, nhờ vào một thuật toán xác định tần số cơ bản nào đó. Đối với phần vô thanh thì khoảng cách giữa các vị trí này là đều nhau.

Độ dài của cửa sổ $w(n)$ được lấy đồng bộ với chu kỳ hiện tại, vì thế các mẫu $s_i(n)$ chỉ khác 0 nếu như nó thuộc vào một của sổ nào đó, tức là nó phụ thuộc vào hệ số xếp chồng: $F_R = \frac{L}{T_0} - 1$

Nếu F_R quá lớn thì tần số vốn có trong các $s_i(n)$ sẽ tác động không tốt tới tần số của tín hiệu tổng hợp. Nếu F_R quá nhỏ thì tín hiệu tổng hợp sẽ khá thô. Hơn thế nữa, biểu thức xấp xỉ mà ta đưa ra ở phần trên sẽ không còn đúng nữa. Nếu chọn được giá trị thích hợp cho F_R thì có thể có được kết quả khá tốt: Nếu F_R=1 (và nếu như tín hiệu nguồn đủ phức tạp) thì phổ của các $s_i(n)$ sẽ xấp xỉ với đường bao phổ của $s(n)$. Khi đó việc tổng hợp sẽ không ảnh hưởng đến formant và các độ rộng của nó.

Những đoạn tín hiệu tiếng nói khác nhau sẽ có khoảng thời gian và tần số khác nhau. Do đó ta sẽ kết hợp mỗi điểm η^i với giá trị của tần số tuần hoàn địa phương T_0, tạo nên các một cặp (η^j, T_0^i) để phân tích các khung OLA của tín hiệu s_i(n). Cuối cùng, bộ ba tham số (η'^j, η^i, T_o^i) sẽ được dùng như một bộ tham số khi tổng hợp tín hiệu. Ở đây η'^j ứng với điểm tần số tổng hợp η^i thông qua hàm $t'(t)$, T_0^i là phân tích khung OLA của điểm tần số tổng hợp hiện tại. Những bộ ba này được minh hoạ ở hình 3.4.



Hình 3.4. Sự thay đổi tần số và thời gian với TD-PSOLA

3.3. TD-PSOLA VÀ TÍN HIỆU TIẾNG NÓI

Khi tổng hợp tiếng nói, kích cỡ của cửa sổ sẽ thay đổi theo từng khung tín hiệu:

$$w_j(n) = w_1\left(n\frac{T_e}{T_0^j}\frac{1}{F_R}\right)$$

$$s_j(n) = s(n)w_j\left(n-\eta^j\right)$$

$$\tilde{s}(n) = \sum_{j=-\infty}^{\infty} s_j\left(n-\eta'^j\right)$$

Trong đó $w_j(n)$ là kích cỡ của cửa sổ mà giá trị của nó phải nằm trong đoạn [0,1]. Các khung OLA đã được lấy từ các đoạn tín hiệu tại vị trí được xác định bởi điểm mốc η^j và gửi tới hệ thống cộng xếp chồng. Với giá trị chuẩn là F_R=1, thì tổng không xác định trên bị giới hạn bởi giá trị lớn nhất của bốn đoạn tín hiệu, đối với các *hệ số pitch* thì tỉ số của tần số tuần hoàn tổng hợp địa phương và tần số gốc được định nghĩa như sau $F_P = \frac{T}{T_0}$ và nằm trong đoạn [0.5,2].

Phải chú ý rằng, tính đúng đắn của công thức xấp xỉ nêu trên phụ thuộc nhiều vào giá trị của tần số tổng hợp. F_P>1 sẽ cho kết quả không tốt. Khi F_P<1, giá trị của K sẽ phụ thuộc nhiều vào các hệ số *pitch*. Để khắc phục, mỗi mẫu tổng hợp sẽ được nhân với hai nhân tố chuẩn hoá:

$$\tilde{s}(n) = \frac{\sum_i \alpha_i\left(n-\eta^i\right)}{\sum_i w_i\left(n-\eta^i\right)}$$

với giả thiết $\alpha_i = \frac{1}{F_P}$.

3.4. CÁC VẤN ĐỀ LIÊN QUAN

Trong giải thuật TD-PSOLA, một tham số tương đối quan trọng là tần số cơ bản (hay chu kỳ cơ bản) của các đoạn tín hiệu được ghép nối. Chính vì vậy trước khi dùng giải thuật này để tổng hợp tiếng nói ta phải tìm được tần số cơ bản của các đoạn tín hiệu. Ngoài ra còn có một vấn đề khác nảy sinh khi áp dụng thuật toán để tổng hợp tiếng nói là khi thực hiện thao tác cộng OLA thì các tín hiệu thành phần có thể không giống nhau dẫn đến sự không tương ứng

về biên độ, lúc đó ta phải làm trơn tín hiệu. Phần tiếp theo sẽ trình bày chi tiết hơn về các vấn đề này.

3.4.1 Xác định tần số cơ bản

Xác định tần số cơ bản là một trong những vấn đề rất quan trọng của xử lý tiếng nói. Nó được sử dụng trong các hệ thống nhận dạng, tổng hợp, thẩm định ghi âm hay phát âm tiếng nói. Do sự quan trọng của nó, có nhiều giải pháp được đưa ra. Phần này sẽ trình bày hai phương pháp đơn giản và dễ áp dụng là dựa vào hàm tự tương quan và hàm vi sai biên độ trung bình.

a. Dùng hàm tự tương quan

Trong xử lý tín hiệu số, hàm tự tương quan của tín hiệu *x(n)* được định nghĩa như sau:

$$R(k) = \sum_{m=-\infty}^{\infty} x(m).x(m+k)$$

Dễ thấy rằng nếu tín hiệu *x(n)* tuần hoàn với chu kỳ P thì hàm tự tương quan cũng tuần hoàn với chu kỳ *P*: $R(k) = R(k+\mathrm{P})$

Hơn nữa hàm tự tương quan còn có những tính chất quan trọng sau:

- Là hàm chẵn $R(k) = R(-k)$
- $R(k)$ đạt giá trị cực đại tại 0: $|R(k)| \le R(0)$ với mọi k
- Giá trị $R(0)$ chính bằng năng lượng của tín hiệu:

$$R(0) = \sum_{m=-\infty}^{\infty} x^2(m)$$

Dựa vào các tính chất trên ta có nhận xét: Hàm tự tương quan sẽ đạt giá trị cực đại tại các mẫu *0*, *±P, ±2P*, … và bằng giá trị năng lượng của tín hiệu, các điểm cực đại được gọi là các đỉnh (peak). Như vậy việc xác định chu kỳ cơ bản của tín hiệu tiếng nói sẽ đưa về việc xác định chu kỳ của hàm tự tương quan.

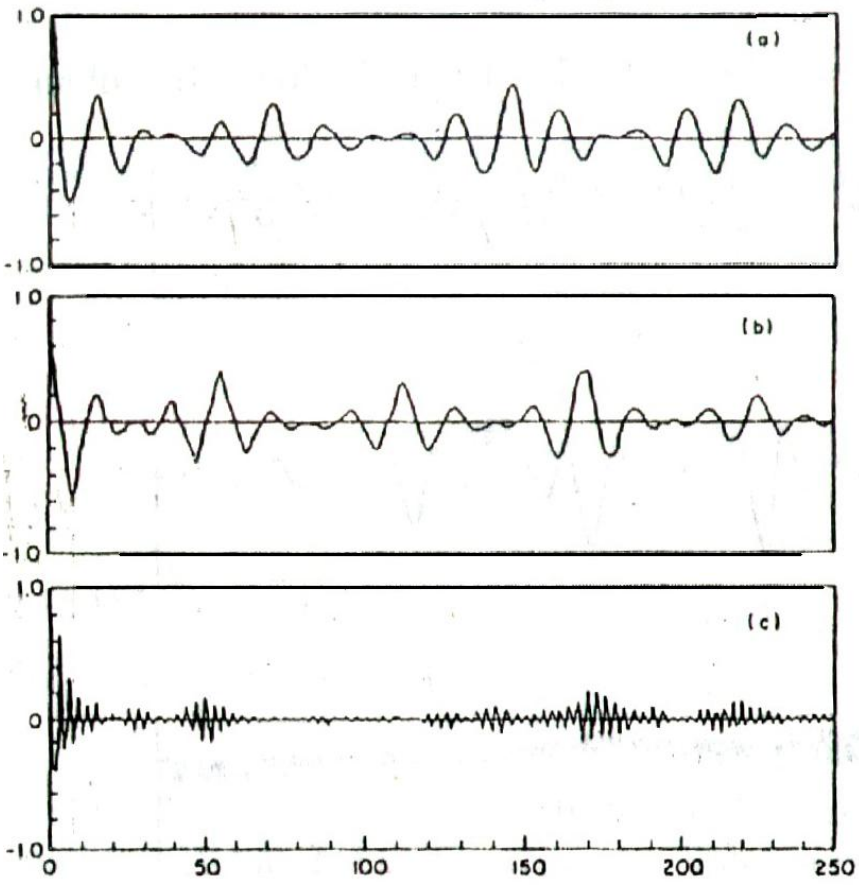
Để áp dụng cho một đoạn tín hiệu tiếng nói, ta phải xác định hàm tự tương quan thời gian ngắn.

Trước hết ta nhân tín hiệu với hàm cửa sổ thích hợp $w(n)$, khi đó hàm tự tương quan được biểu diễn bằng công thức:

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m).w(n-m).x(m+k).w(n-k-m)$$

Biểu thức trên có thể hiểu như sau: đầu tiên một đoạn của tín hiệu tiếng nói được lựa chọn bằng cách nhân với cửa sổ; sau đó việc xác định hàm tự

tương quan theo công thức định nghĩa được áp dụng cho đoạn tín hiệu đã qua cửa sổ.



Hình 3.5. Hàm tự tương quan đối với
âm hữu thanh (a) và (b); và vô thanh (c) dùng của sổ Hamming với *N*=401.

Dễ thấy:

$$R_n(-k) = R_n(k)$$

Và: $$R_n(-k) = R_n(k) = \sum_{m=-\infty}^{\infty} [x(m).w(m-k)][x(n-m).w(n+k-m)]$$

Nếu định nghĩa: $h_k(n) = w(n).w(n+k)$ thì ta có:

$$R_n(k) = \sum_{m=-\infty}^{\infty} [x(m).x(m-k)].h_k(n-m)$$

Tức là $R_n(k)$ đạt được bằng cách cho $x(m).x(m-k)$ qua bộ lọc có đáp ứng xung $h_k(n)$.

Việc tính toán hàm tự tương quan thời gian thực được tiến hành bằng việc sử dụng biểu thức định nghĩa được viết lại như sau:

$$R_n(k) = \sum_{m=-\infty}^{\infty} [x(n+m).w'(m)].[x(n+m+k).w'(k+m)]$$

với $w'(n) = w(-n)$.

Nếu w' là cửa sổ Hamming hoặc chữ nhật thì biểu thức trên có thể biểu diễn như sau:

$$R_n(k) = \sum_{m=0}^{N-l-k} [x(n+m).w'(m)].[x(m+n+k).w'(k+m)]$$

Khi tính toán hàm tự tương quan việc lựa chọn N là rất quan trọng. Do sự không ổn định của tín hiệu tiếng nói nên giá trị N càng nhỏ càng tốt. Mặt khác để hàm tự tương quan tuần hoàn thì cửa sổ phải có chiều dài ít nhất 2 nửa chu kỳ của sóng tín hiệu. Mặt khác khi tính toán hàm tự tương quan thường được chuẩn hoá về 1 đơn vị.

b. Dùng hàm vi sai biên độ trung bình

Xét chuỗi vi sai sau:

$$d(n) = x(n) - x(n-k)$$

Dễ thấy rằng d(n) tuần hoàn cùng chu kỳ P với tín hiệu gốc $x(n)$ và đạt giá trị bằng 0 tại các mẫu $0, \pm kP, \ldots$

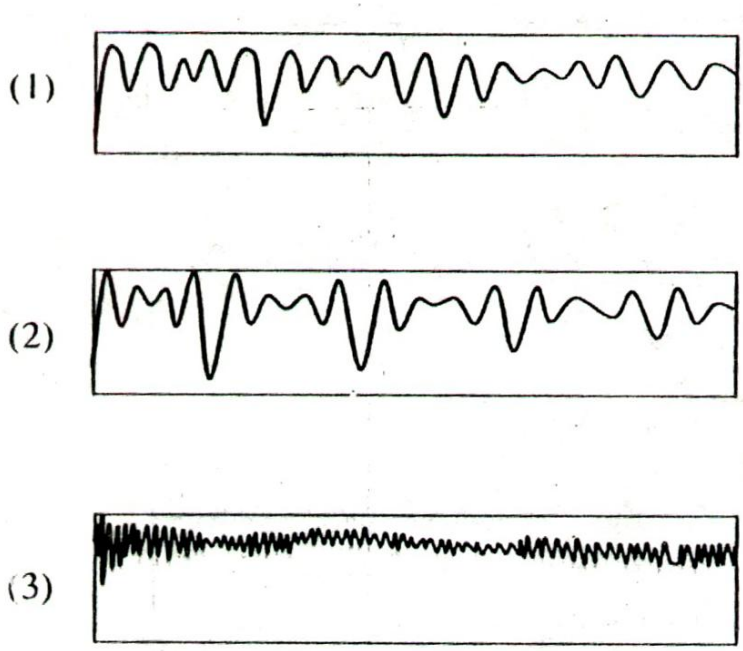
Biên độ trung bình thời gian ngắn của $d(n)$ là một hàm của k có giá trị nhỏ khi k gần chu kỳ. Hàm vi sai biên độ trung bình thời gian ngắn (AMDF) được định nghĩa như sau:

$$\gamma_n(k) = \sum_{m=-\infty}^{\infty} |x(n+m).w_1(m) - x(m+n-k).w_2(m-k)|$$

Rõ ràng rằng nếu $x(n)$ tuần hoàn với chu kỳ cơ bản P trong giới hạn của cửa sổ thì $\gamma_n(k)$ cũng tuần hoàn với chu kỳ P, việc tìm chu kỳ cơ bản của tín hiệu gốc $x(n)$ sẽ được đưa về việc tìm chu kỳ của hàm vi sai biên độ trung bình $\gamma_n(k)$. Nếu cả hai cửa sổ có độ dài như nhau ta sẽ có hàm mô phỏng giống như hàm tự tương quan. Nếu độ dài w_2 lớn hơn dộ dài w_1 thì $\gamma_n(k)$ được tính xấp xỉ theo công thức:

$$\gamma_n(k) = \sqrt{2}\beta[R_n(0) - R_n(k)]$$

với β là hằng số biến đổi từ 0.6 đến 1 với các đoạn khác nhau của tiếng nói.



Hình 3.6. Mô tả hàm vi sai biên độ trung bình
(1),(2) - Âm hữu thanh
(3) – Âm vô thanh

3.4.2. Làm trơn tín hiệu khi ghép nối

a. Phương pháp Microphonemic

Ý tưởng cơ bản của phương pháp Microphonemic là sử dụng các đơn vị có độ dài thay đổi được lấy từ tiếng nói tự nhiên. Các đơn vị này có thể là các từ, các âm tiết hay các âm vị. Từ điển mẫu sẽ được xây dựng từ những đơn vị này.

Các mẫu này được kết hợp trong trục thời gian sử dụng phương pháp PSOLA. Nếu khoảng cách formant giữa các đoạn âm thanh liên tiếp nhỏ hơn 2 Bark, thì sự kết hợp được tạo ra bởi phép nội suy từ các mẫu trên nền biên độ tuyến tính. Nếu sự khác nhau lớn hơn 2 Bark thì một mẫu trung gian phải được thêm vào bởi vì phép nội suy trên nền biên độ không đủ để thay đổi các formant.

Với phụ âm cần đặc biệt chú ý. Ví dụ, với các phụ âm dừng có thể được khôi phục trực tiếp từ các sóng tiếng nói như một biến thể của nguyên âm trong một số ngữ cảnh. Với các âm xát, độ dài mẫu khoảng 50 ms và khoảng 10 ms đối với các đơn vị được lấy ngẫu nhiên từ quá trình ghép nối của phương pháp nội suy trên.

Ưu điểm của phương pháp Microphonemic là đòi hỏi số phép so sánh thấp so với các phương pháp cơ bản. Nhưng vấn đề ở đây là làm thế nào để tối

ưu với các đoạn được lấy từ các mẫu tự nhiên và phát triển các luật để kết hợp chúng.

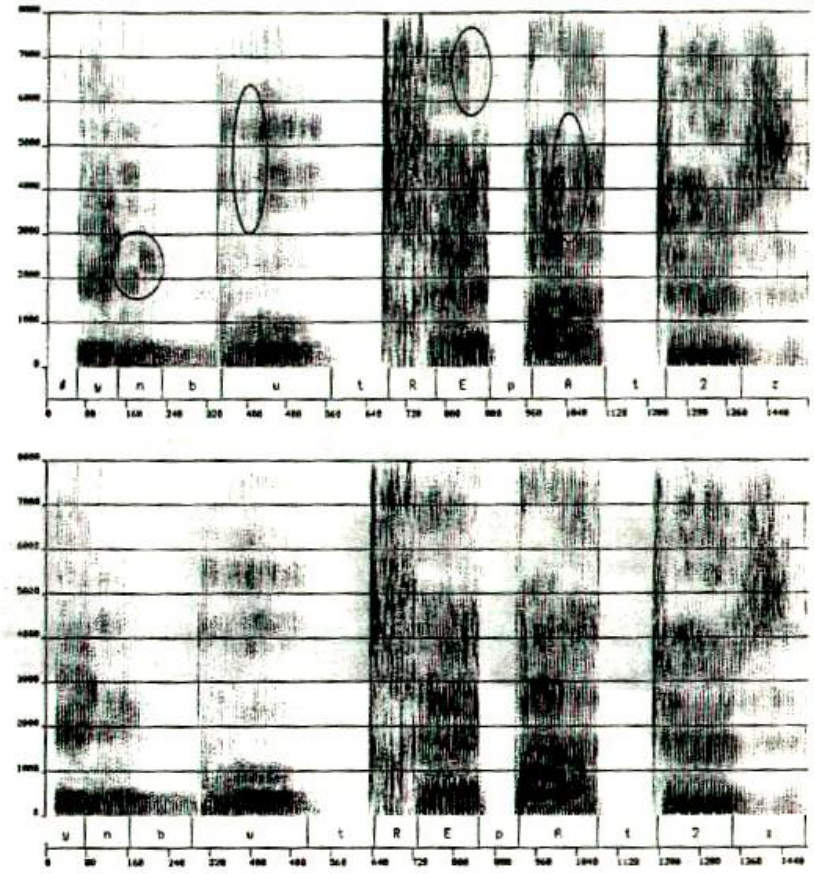
b. Mô hình hình sine

Mô hình hình sine là một mô hình thông dụng, trong đó tín hiệu tiếng nói có thể được biểu diễn bởi một tổng các sóng hình sine (thời gian, biên độ, tần số). Trong mô hình cơ sở này tín hiệu tiếng nói $s(n)$ được mô hình hoá dưới dạng tổng của L đường sine.

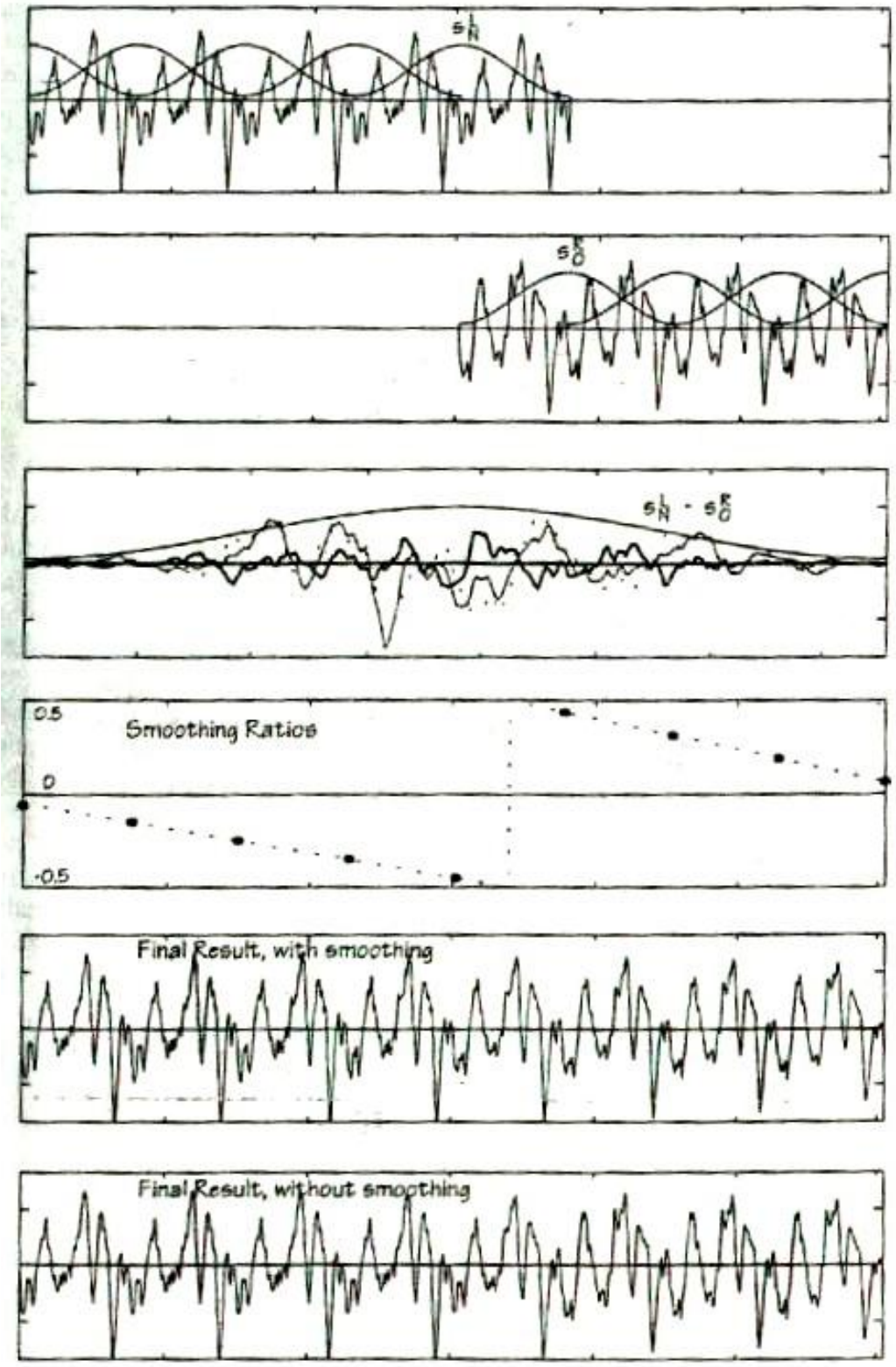
$$s(l) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \phi_l)$$

trong đó $A_l(n)$ và $\phi_l(n)$ đại diện cho biên độ và pha của mỗi đường sine thành phần với tần số góc ω_l.

Mô hình hình sine rất thích hợp khi biểu diễn các tín hiệu tuần hoàn, như nguyên âm và âm hữu thanh. Mô hình hình sine đã sử dụng thành công trong tổng hợp tiếng hát.



Hình 3.7. Ảnh phổ đã qua xử lý làm trơn
tuyến tính trên miền thời gian



Hình 3.8. Xử lý làm trơn tuyến tính miền thời gian

CHƯƠNG 4

THIẾT KẾ CHƯƠNG TRÌNH TỔNG HỢP TIẾNG VIỆT

4.1. PHÂN TÍCH GIẢI THUẬT

Như đã phân tích chi tiết trong các chương trước, với mục đích xây dựng một ứng dụng tổng hợp tiếng Việt từ văn bản, căn cứ vào đặc điểm của ba phương pháp tổng hợp tiếng nói: mô phỏng bộ máy phát âm, tổng hợp formant và tổng hợp bằng ghép nối, thì phương pháp thứ ba, phương pháp tổng hợp bằng ghép nối, là phương pháp khả thi. Vấn đề chính phải giải quyết trong phương pháp này bên cạnh chất lượng của âm tổng hợp là làm sao để kích thước dữ liệu không quá lớn.

Khi nghiên cứu tính chất âm học của tiếng nói, ta thấy rằng bất kỳ một đoạn tín hiệu tiếng nói nào, ngoài sự liên quan chặt chẽ với âm vị (được tạo nên bởi sự thay đổi dạng của tuyến âm trong quá trình phát âm) còn liên quan đến luật ngôn ngữ, trường độ, biên độ, tần số cơ bản F_0 của đoạn tín hiệu. Đối với tiếng nói không thanh điệu (như các tiếng Âu-Ấn) tần số cơ bản F_0 thường thay đổi trong các âm tiết gây nên trọng âm của từ (không làm thay đổi nghĩa) hoặc thay đổi trong câu theo từng loại câu (câu hỏi, câu trần thuật, câu cảm thán...). Tuy nhiên, trong tiếng nói có thanh điệu như tiếng Việt, khi thanh điệu của một âm tiết thay đổi sẽ dẫn tới sự thay đổi về ngữ nghĩa của từ.

Tiếng Việt có 6 thanh điệu: không dấu, huyền, sắc, nặng, hỏi, ngã. Các nghiên cứu về thanh điệu trong tiếng Việt cho thấy rằng sự thay đổi thanh điệu là kết quả của sự thay đổi tần số cơ bản của âm. Do đó nếu thay đổi được tần số cơ bản của tín hiệu theo những dạng thích hợp thì có thể tạo ra các thanh điệu từ các âm không dấu. Việc này hoàn toàn có thể thực hiện được nhờ giải thuật TD-PSOLA đã trình bày trong chương trước.

Như vậy, với việc biến đổi tần số cơ bản của một âm không dấu theo giải thuật TD-PSOLA, thanh điệu của âm tổng hợp sẽ thay đổi và ta có thêm được 5 âm khác. Điều này vô cùng quan trọng, vì nó cho phép giảm kích thước dữ liệu cần lưu trữ đi rất nhiều.

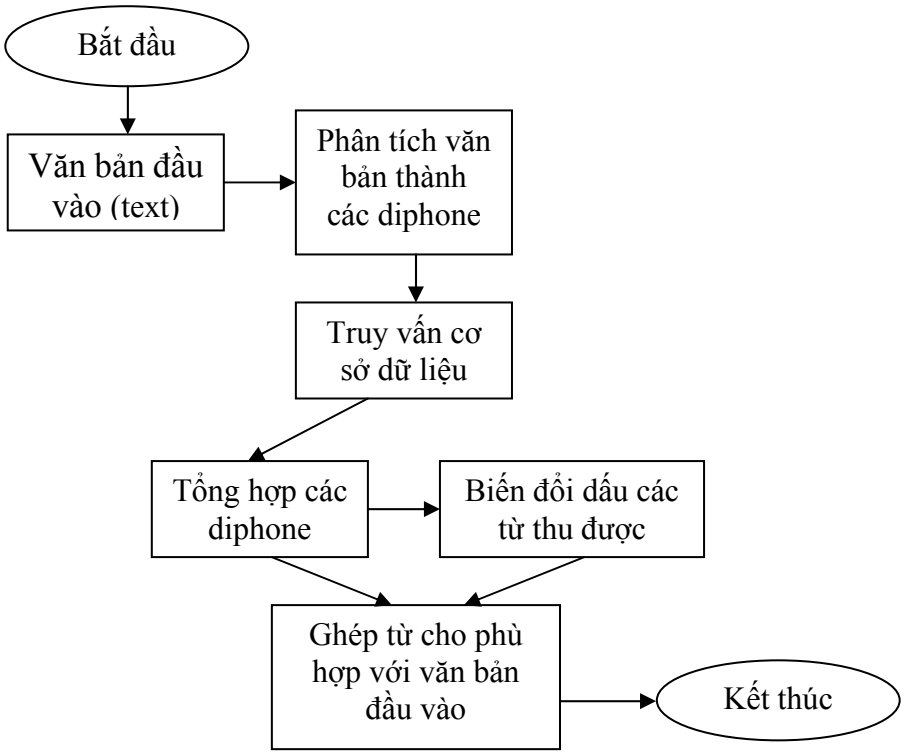
Hơn nữa, do tiếng Việt là ngôn ngữ đơn âm tiết, nên theo cách phát âm, bất kỳ từ nào cũng có thể phân tích được thành hai diphone tương ứng.

Ví dụ: từ *minh* được tạo nên từ hai diphone *mi* và *inh*.

Với những nhận xét này, để xây dựng ứng dụng tổng hợp tiếng Việt từ văn bản bằng giải thuật TD-PSOLA thì các công việc cần thực hiện là:

- Tạo cơ sở dữ liệu.
- Ghép nối các diphone.
- Biến đổi thanh điệu sau khi ghép nối.

Trình tự các bước của quá trình tổng hợp có thể mô tả bởi sơ đồ sau:



Hình 4.1. Sơ đồ tổng hợp tiếng Việt từ văn bản

Phần tiếp theo sẽ trình bày chi tiết việc thực thi và giải quyết các vấn đề liên quan tới những công việc trên.

4.2. DIPHONE TRONG TIẾNG VIỆT

Do đặc điểm đơn âm tiết của tiếng Việt nên mỗi từ luôn có thể phân tích thành hai diphone tương ứng. Với giải thuật TD-PSOLA, ta có thể tạo nên các từ có dấu từ những từ không dấu theo nguyên tắc: *biến đổi tần số cơ bản của từ không dấu một cách thích hợp*.

Ví dụ: *chào* có thể được tạo ra từ *chao* (kết hợp của hai diphone *cha* và *ao*) bằng cách biến đổi tần số cơ bản của từ *chao* để tạo thành thanh điệu tương ứng với dấu huyền.

Tuy nhiên có một số từ không tuân theo quy tắc này.

Ví dụ: từ *các* không thể tạo nên từ một từ không dấu nào, mặc dù từ *các* có thể coi như được tạo nên từ hai diphone *ca* và *ác*.

Như vậy, ngoài các từ trong tiếng Việt đều có thể áp dụng được quy tắc tổng hợp mà ta đưa ra còn có một số từ được tạo nên từ các diphone có dấu (với từ *các* là diphone *ác*).

Tóm lại:

- Mỗi từ trong tiếng Việt được tạo nên từ hai diphone, trong đó một diphone ta sẽ gọi là diphone bắt đầu và diphone kia gọi là kết thúc, được ký hiệu thêm ký tự "_" (tượng trưng cho khoảng lặng) để phân biệt.
  Ví dụ : *minh* = *_mi* + *inh_*
  _mi là diphone bắt đầu
  *inh_* là diphone kết thúc.

Tương tự như vậy: *anh* = *_a* + *anh_*

- Các diphone bắt đầu gồm một phụ âm rồi tới một nguyên âm. Ví dụ: *_nha*
- Các diphone kết thúc gồm một nhóm (một hoặc nhiều) nguyên âm rồi tới một phụ âm.
  Ví dụ: *ương_*
- Các nguyên âm có thể đứng riêng để tạo thành một diphone.
  Ví dụ: e vừa có thể là diphone bắt đầu trong từ *em* (*_e* + *em_*) nhưng cũng có thể là diphone kết thúc trong *me* (*_me* + *e_*)

Dựa vào những đặc điểm trên ta có bảng các diphone trong tiếng Việt như sau:

_a	bê	_e	hu	lo	nhơ	ơn	se	**uật**	**ượt**
a_	bi	e_	hư	lô	nhu	**ớp**	sê	**úc**	**ứt**
ác	bo	em	hy	lơ	như	**ợp**	si	**ục**	**ựt**
ạc	bô	en	_i	lu	_o	**ớt**	so	uê	ưu
ách	bơ	eng	i_	lư	o_	**ợt**	sô	ui	va
ạch	bu	eo	ia	ly	oa	pa	sơ	um	ve
ai	bư	**ép**	**ích**	ma	**oát**	pe	su	un	vê
am	ca	**ẹp**	**ịch**	me	**oạt**	pê	sư	ung	vi
an	co	**ét**	**iếc**	mê	oai	pi	ta	**uốc**	vo

ang	cô	**ẹt**	**iệc**	mi	oan	po	te	**uộc**	vô
anh	cơ	_ê	iêng	mo	oang	pô	tê	uôi	vơ
ao	cu	ê_	**iếp_**	mô	oanh	pơ	ti	uôn	vu
áp	cư	**ếch**	**iệp_**	mơ	**óc**	pu	to	uôm	vư
ạp	cha	**ệch**	**iết**	mu	**ọc**	pư	tô	**uốt**	xa
át	che	êm	**iệt**	mư	oe	py	tơ	**uột**	xe
ạt	chê	ên	iêu	my	oi	pha	tu	uông	xê
au	chi	ênh	im	na	om	phe	tư	**úp**	xi
ay	cho	**ếp**	in	ne	on	phê	tha	**ụp**	xo
ắc	chô	**ệp**	inh	nê	ong	phi	the	**út**	xơ
ặc	chơ	**ết**	**íp**	ni	oong	pho	thê	**ụt**	xu
ăm	chu	**ệt**	**ịp**	no	**óp**	phô	thi	uy	xư
ăn	chư	Êu	**ít**	nô	**ọp**	phơ	tho	uya	xy
ăng	da	ga	**ịt**	nơ	**ót**	phu	thô	uyên	_y
ắp	de	gi	iu	nu	**ọt**	phư	thơ	**uyết**	y_
ặp	dê	gia	ke	nư	_ô	qua	thu	**uyệt**	yêm
ắt	di	ghe	kê	nga	ô_	que	thư	**uýt**	yên
ặt	do	ghê	ki	nghe	**ốc**	quê	tra	**uỵt**	**yết**
ấc	dô	ghi	kha	nghê	**ộc**	qui	tre	_ư	**yệt**
ậc	dơ	go	khe	nghi	ôi	quơ	trê	ư_	yêu
âm	du	gô	khê	ngo	ôm	quy	tri	**ức**	
ân	dư	gơ	kho	ngô	ôn	ra	tro	**ực**	
âng	đa	gu	khô	ngơ	ông	re	trô	ưm	
ấp	đe	gư	khơ	ngu	**ốp**	rê	trơ	ưn	
ập	đê	ha	khu	ngư	**ộp**	ri	tru	ưng	
ất	đi	he	khư	nha	**ốt**	ro	trư	**ước**	
ật	đo	hê	ky	nhe	**ột**	ro	_u	**ược**	
âu	đô	hi	la	nhê	_ơ	rơ	u_	ươi	
ây	đơ	ho	le	nhi	ơ_	ru	ua	ươn	
ba	đu	hô	lê	nho	ơi	rư	uân	ương	
be	đư	hơ	li	nhô	ơm	sa	**uất**	**ướt**	

Bảng 4.1. Các diphone trong tiếng Việt

Tổng số diphone cần có để tạo nên tất cả các từ tiếng Việt là 389 diphone, trong đó có 61 diphone kết thúc có dấu trước (được in đậm trong bảng).

4.3. XÂY DỰNG CƠ SỞ DỮ LIỆU

Trong phương pháp tổng hợp tiếng nói bằng ghép nối thì cơ sở dữ liệu là phần vô cùng quan trọng. Chất lượng của dữ liệu ảnh hưởng trực tiếp tới chất lượng của tiếng nói tổng hợp. Khi xây dựng cơ sở dữ liệu cho quá trình tổng hợp tiếng nói bằng ghép nối diphone cần thực hiện các công việc sau:

- Thu âm (thu một số từ làm mẫu).
- Tách các diphone từ mẫu thu được.
- Lưu trữ diphone.

4.3.1. Thu âm

a. Quá trình thu âm

Để thu được các mẫu âm thanh là diphone thì cần phải thu một từ, sau đó tách ra các diphone. Vì chất lượng dữ liệu thu được ảnh hưởng trực tiếp đến tiếng nói tổng hợp nên các từ cần được thu trong một câu. Trong câu ngoài từ cần thu còn có thêm các từ đứng trước và đứng sau để làm cho việc phát âm từ cần thu tự nhiên hơn.

Ví dụ như:

Chào ***anh*** đi

Chào ***bắc*** đi

Trong các câu này, từ cần thu là "*anh*" và "*bắc*". Các từ "*chào*" và "*đi*" chỉ thêm vào để làm cho việc phát âm từ "*anh*" hay "*bắc*" tự nhiên hơn. Trong trường hợp câu thu được có ba từ mà vẫn không cho tín hiệu ổn định thì phải tăng số lượng từ trong câu lên 5 hay 7 từ, chẳng hạn như "chúc mừng ***anh*** năm mới". Với mỗi từ cần thu ta sẽ thu nhiều lần để có thể chọn được mẫu với chất lượng tốt.

Trong khi xây dựng cơ sở dữ liệu ta còn có thể giảm số lượng diphone cần thu căn cứ vào các diphone đồng âm và cách phát âm của từng địa phương.

Chẳng hạn như "*iên_*" có thể dùng như "*yên_*". Với cách phát âm của người Hà Nội thì diphone "*_do*" có thể dùng như "*_gio*"...

Toàn bộ quá trình thu được thực hiện tại trung tâm MICA, trường đại học Bách Khoa Hà Nội, với một giọng nữ người Hà Nội.

b. Xử lý sau khi thu

Sau khi thu, âm thanh được lưu lại dưới dạng file *.wav PCM không nén, mono, 16 bit , tần số lấy mẫu là 16000Hz. Khuôn dạng này cho phép lưu lại tín hiệu tiếng nói với chất lượng khá tốt.

Nhiễu trong quá trình thu được loại bỏ bằng phần mềm Cool Edit 2000 Pro. Các mẫu thu được không phải lúc nào cũng có biên độ như mong muốn (to

quá hoặc nhỏ quá) nên sau khi thu, tín hiệu mẫu có thể phải điều chỉnh biên độ. Công việc này cũng được thực hiện bằng phần mềm Cool Edit.

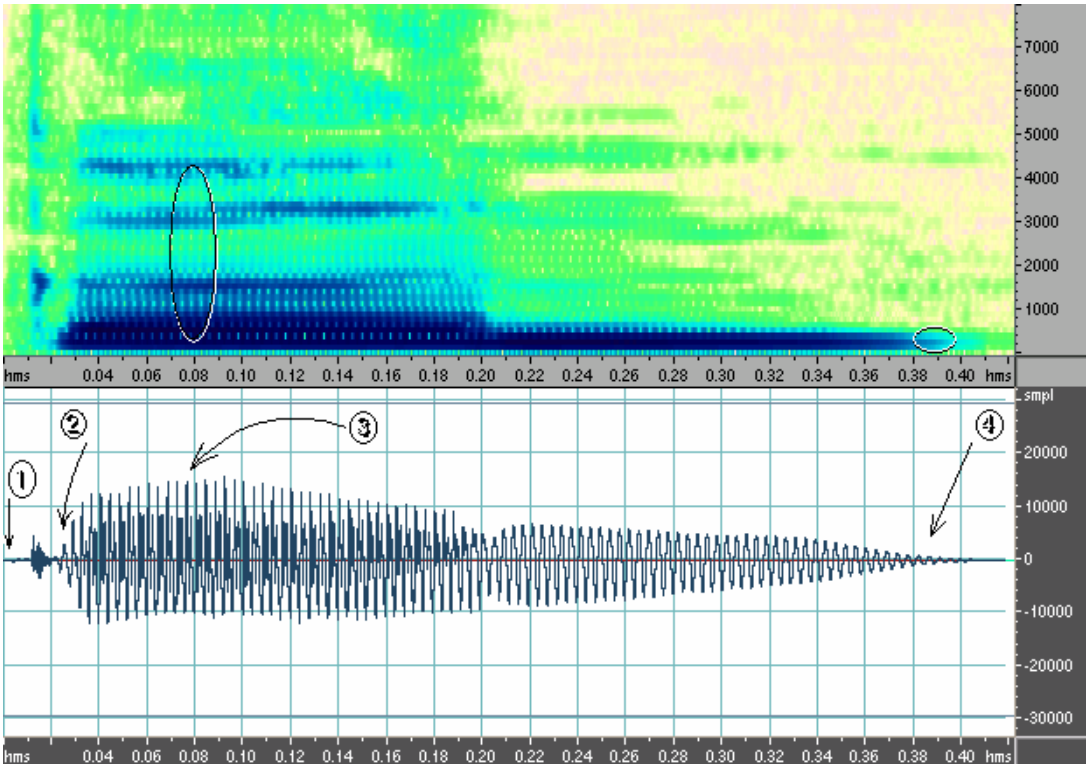
4.3.2. Tách diphone

Sau khi có được các từ mẫu, công đoạn tiếp theo là phải tách từ đã có để thu được các diphone mong muốn. Đây là công việc khó nhất khi xây dựng dữ liệu.

Do đặc điểm từ đơn âm tiết của tiếng Việt nên từ mẫu được tạo nên từ hai diphone, diphone bắt đầu và diphone kết thúc. Để có được các diphone này, ta phải cắt ra các đoạn tín hiệu từ mẫu đã có. Các diphone bắt đầu được cắt từ phần bên trái của mẫu và các diphone kết thúc được cắt từ phần bên phải của mẫu. Các điểm cắt được xác định trực tiếp bằng mắt. Điểm cắt phải thoả mãn điều kiện:

- Nằm trong phần tín hiệu ổn định.
- Điểm cắt bên phải của diphone bắt đầu và điểm cắt bên trái của diphone kết thúc phải nằm tại đỉnh cao nhất trong một chu kỳ của phần tín hiệu tương ứng với nguyên âm (gần như tuần hoàn). Điều này bảo đảm rằng hai diphone dùng để ghép nối không bị lệch pha nhiều.

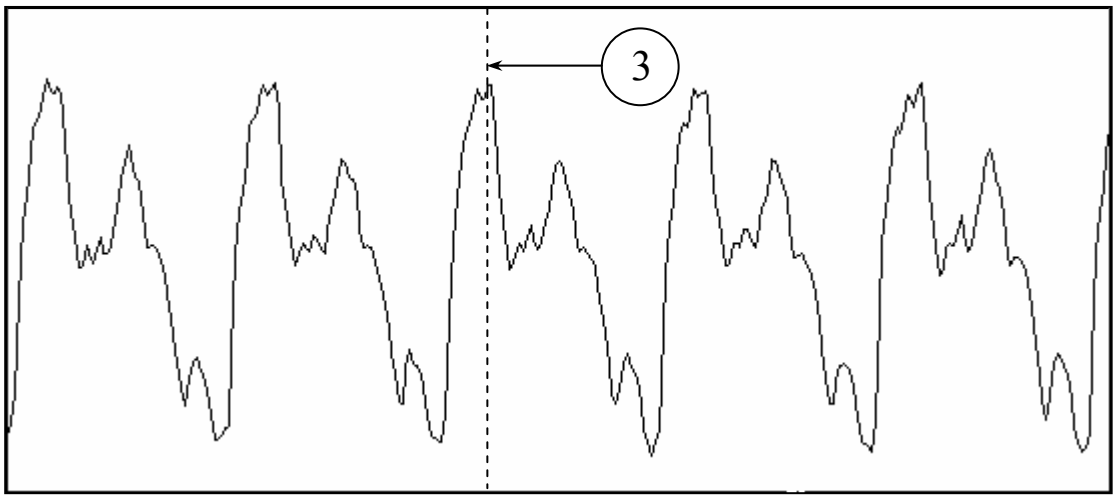
Chi tiết về việc cắt các diphone từ một mẫu được mô tả trong Hình 4.2.



Hình 4.2. Tách diphone "_cơ" từ mẫu "cơn"

Trong hình 4.2, phần bên dưới là dạng thời gian của từ "*cơn*", từ từ mẫu này ta có thể có được hai diphone là "*_cơ*" và "*ơn_*". Muốn cắt đực diphone "*_co*", do "*_co*" là diphone bắt đầu nên ta sẽ cắt ở phần bên trái của mẫu.

Điểm (1) tương ứng với điểm bắt đầu của âm "*c*", điểm (2) là điểm bắt đầu của âm "*o*", và điểm (3) nằm giữa âm "*o*" được coi là điểm kết thúc của diphone "*_co*". Đoạn nằm từ điểm (1) tới điểm (3) chính là đoạn cần lấy.



Hình 4.3. Điểm cắt bên phải của diphone "*_cơ*"

Điểm (3) trong hình 4.2 nên chọn tại vị trí tương ứng với đường cắt trong hình 4.3 (cắt tại đỉnh cao nhất của một chu kỳ trong đoạn tín hiệu tương ứng với nguyên âm).

Sau khi thực hiện tách diphone từ âm mẫu, ta có các đoạn tín hiệu tương ứng với các diphone bắt đầu và kết thúc. Công việc tiếp theo là lưu trữ các diphone này trong cơ sở dữ liệu.

4.3.3. Lưu trữ dữ liệu

Để tiện cho việc đọc dữ liệu, tất cả các diphone được lưu trong một file dữ liệu. Do các diphone thu được là tín hiệu âm thanh 16 bit nên mỗi mẫu âm thanh tương ứng với một số nhị phân 2 byte, như vậy ta có thể dùng file nhị phân để lưu dữ liệu. Hơn nữa, do khuôn dạng âm thanh các diphone như nhau nên chỉ cần lưu trữ phần dữ liệu của file wav tương ứng (không cần lưu trữ header). Theo cách này có thể tiết kiệm được 44 byte (là kích thước header của file wav) cho một diphone. File dữ liệu cần được tổ chức để việc quản lý, cập nhật dữ liệu thuận tiện và dễ dàng.

Ngoài dữ liệu âm thanh cần lưu trữ, khi ghép nối các diphone theo giải thuật TD-PSOLA còn cần phải xác định đoạn tín hiệu tuần hoàn và chu kỳ của nó nên dữ liệu về mỗi diphone còn có thêm các thông tin về độ dài chu kỳ đầu

tiên của đoạn tín hiệu tuần hoàn (tương ứng với phần hữu thanh trong diphone) và điểm chuyển tiếp giữa phần vô thanh và hữu thanh.

Dữ liệu của một diphone có khuôn dạng như sau:

Byte	Ý nghĩa
0 – 1	**Độ dài của chu kỳ đầu tiên**
2 – 3	**Vị trí của điểm chuyển tiếp giữa phần vô thanh và hữu thanh của diphone**
4 – 7	Chưa dùng
7 – Cuối	**Dữ liệu các diphone : 16 bit, mono, tại 16 kHz**

Bảng 4.2. Cấu trúc dữ liệu cho một diphone

Khi tổng hợp, các diphone cần ghép nối được xác định theo dạng văn bản của từ tương ứng, vì vậy mỗi diphone được lưu trữ với một tên, việc tìm kiếm các diphone trong cơ sở dữ liệu được thực hiện theo tên này.

File cơ sở dữ liệu có cấu trúc chi tiết mô tả trong bảng 4.3.

	Phần	Byte		Ý nghĩa
	1	0 ÷1		Số lượng các diphone
Header	2	**2 ÷5**		**Tên của diphone đầu tiên**
		6 ÷9		**Vị trí của diphone đầu tiên**
		10 ÷13		**Tên của diphone thứ hai**
		14 ÷17		**Vị trí của diphone thứ hai**
		...		**... ... ...**
		8*(N-1)+2 ÷ 8*(N-1)+5		**Tên của diphone thứ N**
		8*(N-1)+6 ÷ 8*(N-1)+9		**Vị trí của diphone thứ N**
		8*N+2 ÷11999		**Rỗng**
Dữ liệu	3	diphone 1	12000 ÷ ...	Dữ liệu của diphone có cấu trúc như bảng 4.2
		...	...	... ...
		diphone N	... ÷ ...	Dữ liệu của diphone có cấu trúc như bảng 4.2

Bảng 4.3. Cấu trúc lưu trữ của file cơ sở dữ liệu

File dữ liệu gồm hai phần:

- **Phần header** (12000 byte): lưu thông tin về số lượng các diphone, tên diphone và vị trí lưu trữ tín hiệu diphone đó trong file. Thông tin về tên và vị trí của mỗi diphone chiếm 8 byte (4 byte dành cho tên, 4 byte dành cho vị trí lưu trữ). Như vậy, phần header cho phép lưu thông tin cho gần 1500 diphone. Hiện tại, số lượng diphone là 389. Phần còn lại dùng để lưu các thông tin bổ sung.
- **Phần dữ liệu**: lưu dữ liệu cho các diphone theo bảng 4.2.

Với bộ dữ liệu hiện có (389 diphone), file dữ liệu có kích thước khoảng 2,37 MB.

4.4. PHÂN TÍCH VĂN BẢN THÀNH CÁC DIPHONE

Dựa trên cơ sở dữ liệu đã có, văn bản (tiếng Việt) được phân tích thành các từ, qua đó xác định các diphone tương ứng để tổng hợp. Do từ được tạo nên bởi các diphone không dấu bằng cách biến đổi tần số cơ bản, trong khi đó tần số cơ bản phụ thuộc vào loại câu (trần thuật, hỏi...), nên việc xác định từ luôn đi kèm với việc xác định sự biến đổi tần số cơ bản của từ.

4.4.1. Phân tích văn bản tiếng Việt thành các từ

Quá trình phân tích văn bản thành các từ gồm hai thao tác: Xác định câu trong văn bản và xử lý câu.

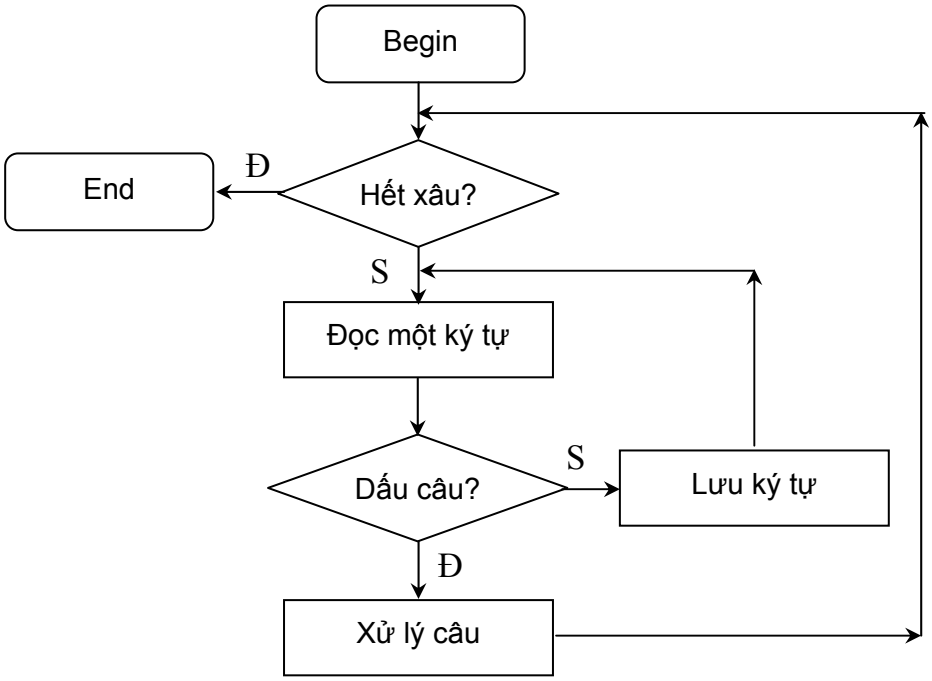
a. Xác định câu trong văn bản

Câu trong văn bản được ngăn cách với nhau bởi các dấu câu. Các dấu câu được cho trong bảng 4.4. Cần chú ý rằng khái niệm "*câu*" ở đây nhằm chỉ các loại câu khác nhau (trần thuật, hỏi...) để xác định sự biến đổi của tần số cơ bản và có thể không chặt chẽ về ngữ pháp.

Loại dấu câu	Cách viết
Dấu chấm	.
Dấu phẩy	,
Dấu chấm phẩy	;
Dấu hai chấm	:
Dấu chấm than	!
Dấu chấm hỏi	?
Các dấu ngoặc	( ) [ ] { }

Bảng 4.4. Các loại dấu câu

Do chương trình chỉ xét các văn bản dưới dạng text nên toàn bộ văn bản được coi như một xâu ký tự.Các câu được xác định theo lưu đồ thuật toán sau:



Hình 4.4. Lưu đồ thuật toán xác định câu trong văn bản

b. Xử lý câu

Sau khi được xác định, câu được phân loại để xử lý. Với mục đích thử nghiệm tổng hợp câu, báo cáo này chỉ chia câu làm ba loại:

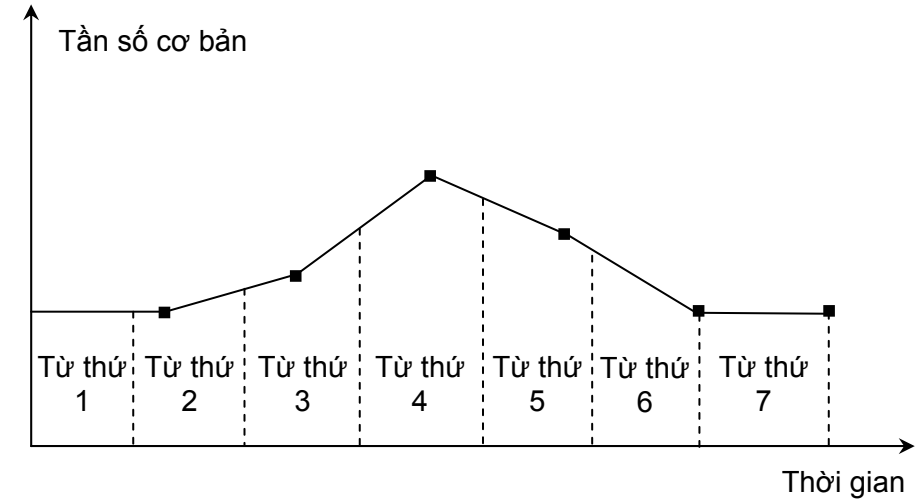
- Loại 1 (câu trần thuật): tương ứng với các dấu: “.”, “;” “)”, “]”, “}”
- Loại 2 (câu hỏi): tương ứng với dấu câu: “?”
- Loại 3 (câu hơi lên giọng ở cuối câu): dấu “,”, “!”

Sự biến đổi các thông số của tín hiệu tiếng nói tổng hợp phụ thuộc vào từng loại câu. Vấn đề này được trình bày chi tiết trong mục 4.6.2.

Căn cứ vào sự biến đổi các thông số của tín hiệu tiếng nói, câu được phân tích thành các từ đi kèm với các thông số của từ. Các thông số của từ bao gồm:

- Sự biến đổi tần số cơ bản
- Biên độ
- Trường độ

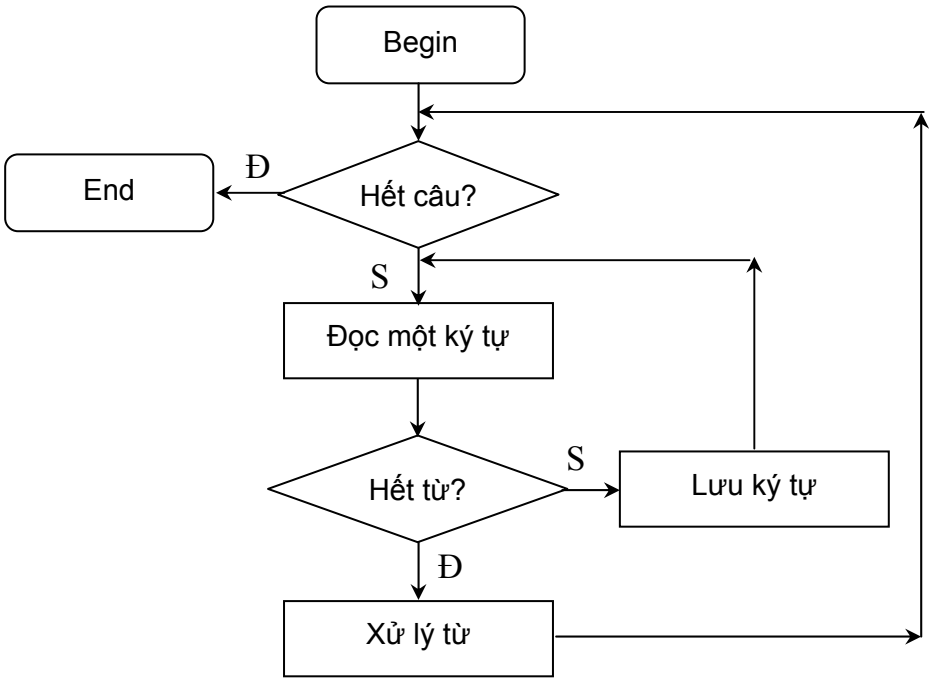
Hình 4.5 minh hoạ sự biến đổi tần số cơ bản của các từ theo sự biến đổi tần số cơ bản của câu.



Hình 4.5. Sự biến đổi tần số cơ bản của từ theo tần số cơ bản của câu

Các từ được nhấn mạnh trong câu (ví dụ từ để hỏi trong câu hỏi) có biên độ và trường độ của từ này lớn hơn các từ khác. Vấn đề này cũng được trình bày trong mục 4.6.2.

Việc tách từ trong câu được thực hiện theo lưu đồ thuật toán ở hình 4.6.



Hình 4.6. Lưu đồ thuật toán xác định từ trong câu

4.4.2. Tách từ thành các diphone

Sau khi xác định từ sẽ được xử lý bằng cách tách thành hai diphone tương ứng. Quá trình này gồm hai thao tác: chuyển từ cách biểu diễn tiếng Việt sang hiển thị theo kiểu telex và tách dạng biểu diễn telex thành hai diphone.

a. Chuyển từ biểu diễn tiếng Việt sang biểu diễn dạng telex

Để tiện xử lý về sau (sử dụng các bảng mã tiếng Việt khác nhau), trước khi tách thành hai diphone từ được chuyển thành dạng telex. Dấu của từ được viết ở cuối từ.

Ví dụ: từ *trường* được chuyển thành *truwowngf*

Việc chuyển từ dạng tiếng Việt thông thường sang dạng telex tuỳ thuộc vào loại bảng mã được sử dụng. Chương trình sử dụng bảng mã 8 bit TCVN3-ABC. Các ký tự trong tiếng Việt và mã tương ứng được cho trong phụ lục 2.

b. Tách từ thành hai diphone

Từ ở dạng biểu diễn telex được tách thành hai diphone bắt đầu và kết thúc tương ứng. Diphone bắt đầu được phân biệt bằng dấu "_" phía trước, diphone kết thúc có dấu "_" phía sau.

Ví dụ: từ *truwowngf* được tách thành hai diphone *_truw* và *uwowng_*

Mấu chốt của việc tách một từ thành hai diphone là phát hiện được vị trí bắt đầu và kết thúc của nguyên âm đầu tiên (theo chiều từ trái sang phải).

Ví dụ: nếu tìm được nguyên âm *ư (uw*) thì dễ dàng tách từ *truwowng* thành *truw* và *uwowng.*

Thuật toán xác định vị trí bắt đầu và kết thúc của nguyên âm đầu tiên được cho trong hình 4.7.

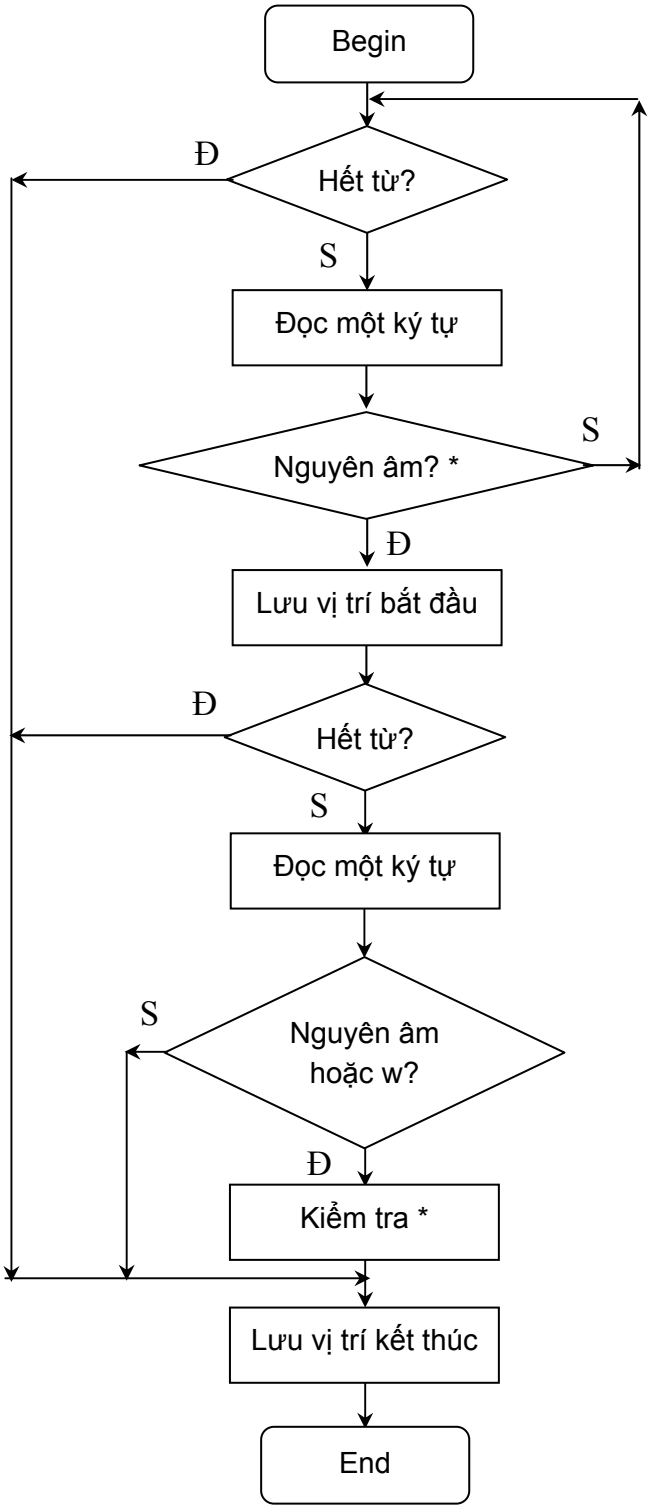
Trong lưu đồ 4.7. * ứng với quá trình kiểm tra xem hai ký tự liên tiếp có phải là *aa, aw, ee, oo, ow, uw* hay không.

Việc xác định diphone kết thúc phải đi kèm với việc xác định dấu của từ, vì có trường hợp diphone kết thúc không thể tạo thành từ diphone không dấu.

Ví dụ: từ *các* và *cạc* đều có diphone kết thúc là *ac_*, diphone này không thể tạo thành từ diphone không dấu nên phải căn cứ vào dấu của từ để xác định diphone là *acs_* hay *acj_*.

Các trường hợp này tương ứng với những diphone in đậm trong bảng 4.1.

Đa số các diphone được lưu trong cơ sở dữ liệu với tên là cách biểu diễn diphone, ví dụ diphone *an_* có tên là *an_* trong cơ sở dữ liệu, nhưng với diphone có cách biểu diễn dài, ví dụ *uwowng_,* thì tên lưu trong cơ sở dữ liệu khác với cách biểu diễn *wog_*(tên của các diphone trong cơ sở dữ liệu với kích thước 4 byte) nên cần chuyển đổi cách biểu diễn diphone phù hợp với tên trong cơ sở dữ liệu. Tên của các diphone dài được cho trong phụ lục 3.



Hình 4.7. Lưu đồ thuật toán xác định vị trí nguyên âm đầu tiên

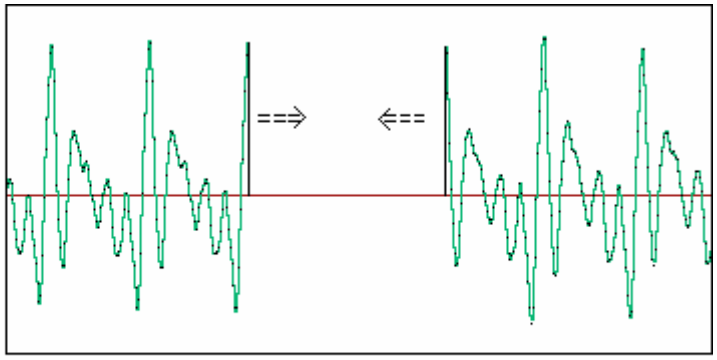
4.5. GHÉP NỐI CÁC DIPHONE VÀ ĐIỀU KHIỂN TẦN SỐ CƠ BẢN

Văn bản cần xử lý được phân tích tuần tự theo từng từ. Với mỗi từ ta xác định được diphone bắt đầu và kết thúc tương ứng. Công việc tiếp theo là ghép nối các diphone này lại và biến đổi tần số cơ bản để tạo thành tiếng nói tổng hợp.

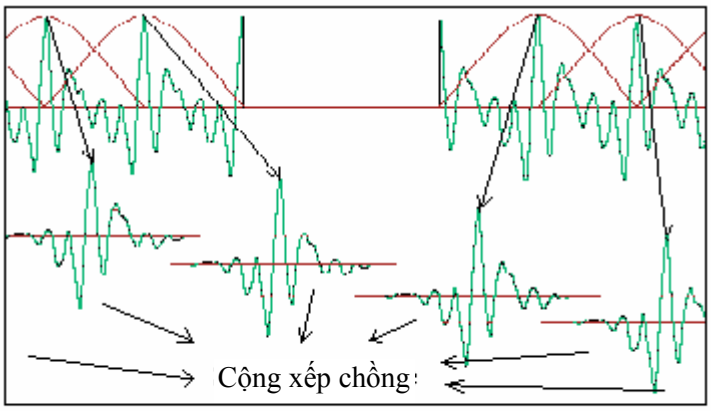
4.5.1. Ghép nối các diphone

Căn cứ vào tên của diphone bắt đầu và kết thúc, sau khi truy xuất cơ sở dữ liệu ta có được hai đoạn tín hiệu tương ứng với hai diphone này. Việc ghép nối được thực hiện giữa phần kết thúc của diphone bắt đầu và phần bắt đầu của diphone kết thúc.

Chú ý rằng các phần này đều nằm trong đoạn hữu thanh (tuần hoàn) của tín hiệu. Theo cách tách diphone đã nêu trong phần 4.3.2, các điểm ghép nối đều nằm tại các đỉnh cao nhất của mỗi chu kỳ. Sơ đồ ghép nối hai tín hiệu được cho trong hình 4.8.



Hình 4.8. Ghép nối hai diphone



Hình 4.9. Cộng xếp chồng các tín hiệu thành phần

Muốn thay đổi độ dài của tín hiệu thu được (độ dài của phần tuần hoàn), trước hết các diphone cần được phân tích thành các tín hiệu thành phần có độ dài xác định được. Sau đó, dùng TD-PSOLA cộng xếp chồng các tín hiệu thành phần lại để được một tín hiệu có độ dài mong muốn. Hình 4.9 mô tả quá trình này.

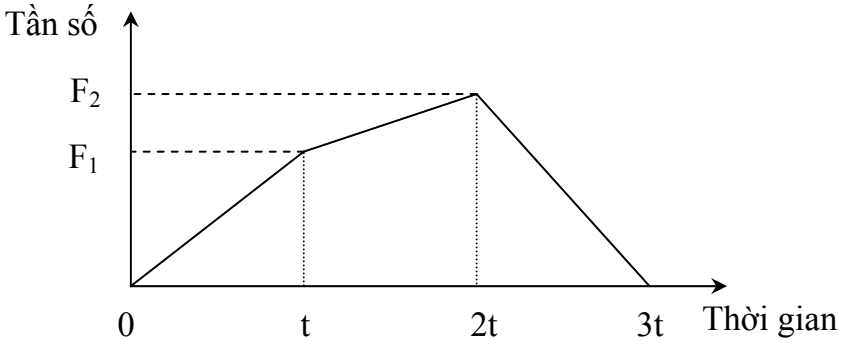
Sau khi thực hiện ghép nối, tín hiệu thu được tương ứng với một từ không dấu. Để tổng hợp được các từ có dấu, ta phải biến đổi tần số cơ bản của tín hiệu theo quy luật biến đổi tần số cơ bản của các thanh điệu trong tiếng Việt.

4.5.2. Biến đổi tần số cơ bản

Tín hiệu thu được sau khi ghép nối hai diphone có tần số cơ bản (của đoạn tín hiệu tuần hoàn) là tần số cơ bản của tín hiệu ban đầu (tín hiệu tiếng nói khi thu âm).

Để biến đổi tần số cơ bản của tín hiệu ta cần biết:

- Khoảng thời gian của tiếng nói tổng hợp
- Quá trình biến đổi của tần số cơ bản



Hình 4.10. Quá trình biến đổi tần số cơ bản của từ theo thời gian

Một cách gần đúng, có thể coi tần số cơ bản của từ biến đổi theo đường gấp khúc như hình 4.10. Tín hiểu tổng hợp được chia thành các đoạn với độ dài (thời gian) bằng nhau, tần số cơ bản trong mỗi đoạn thời gian biến thiên theo một đường thẳng. Nhờ vậy có thể xác định được tần số cơ bản tại tất cả các thời điểm.

Tín hiệu được chia thành các đoạn bằng nhau với độ dài (thời gian) nhỏ hơn, tần số cơ bản của mỗi đoạn không đổi và bằng tần số tại điểm giữa của đoạn trong quá trình biến đổi. Áp dụng TD-PSOLA, ta có thể tổng hợp được đoạn tín hiệu với tần số và độ dài cho trước.

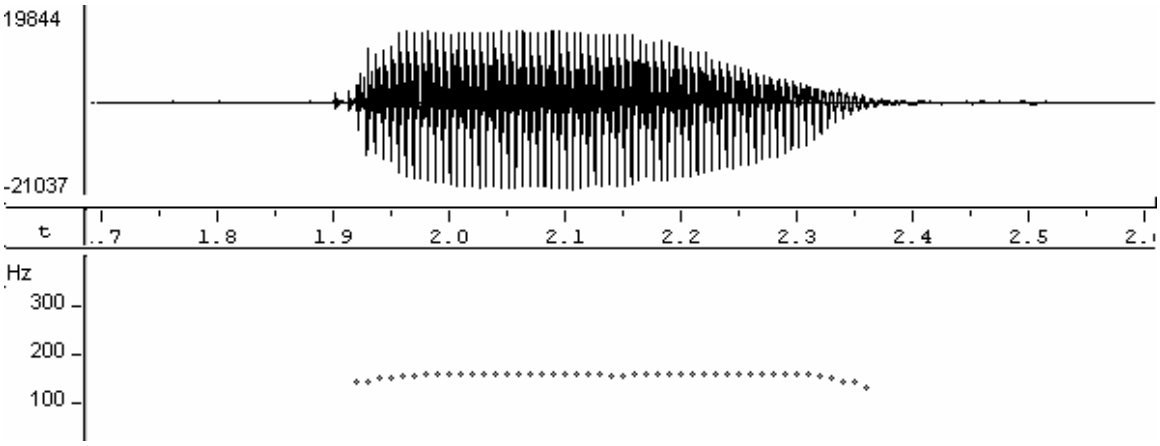
4.6. SỰ BIẾN ĐỔI THÔNG SỐ TÍN HIỆU TRONG CÁC THANH ĐIỆU VÀ CÂU

4.6.1. Biến đổi tần số cơ bản trong các thanh điệu

Trong tiếng Việt, ngữ nghĩa của một từ phụ thuộc vào thanh điệu. Khi thanh điệu thay đổi, nghĩa của từ cũng thay đổi theo. Có 6 thanh điệu trong tiếng Việt: không dấu, huyền,sắc, nặng, hỏi, ngã. Tương ứng với mỗi thanh điệu, tần số cơ bản thay đổi theo một quy luật riêng.

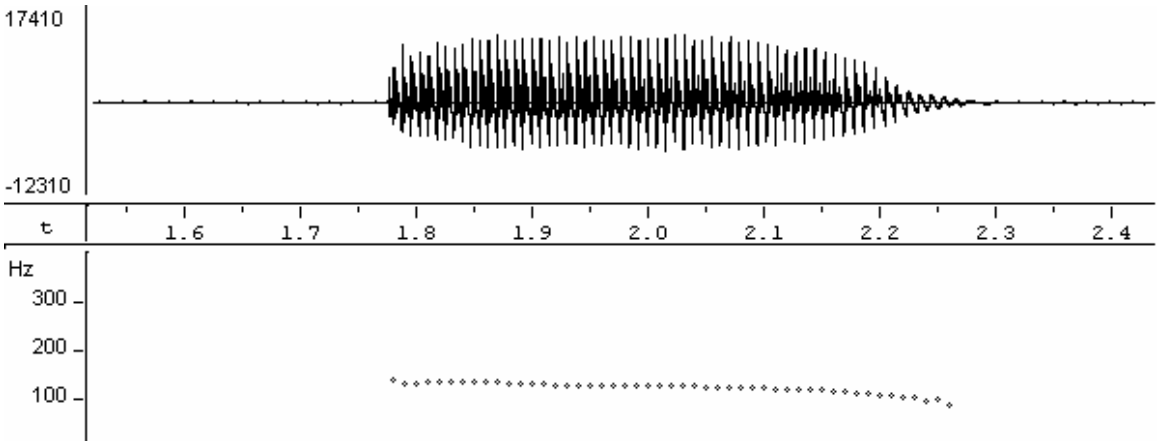
a. Không dấu

Với thanh điệu không dấu, tần số cơ bản không thay đổi.



Hình 4.11. Thanh điệu không dấu (âm a)

b. Dấu huyền



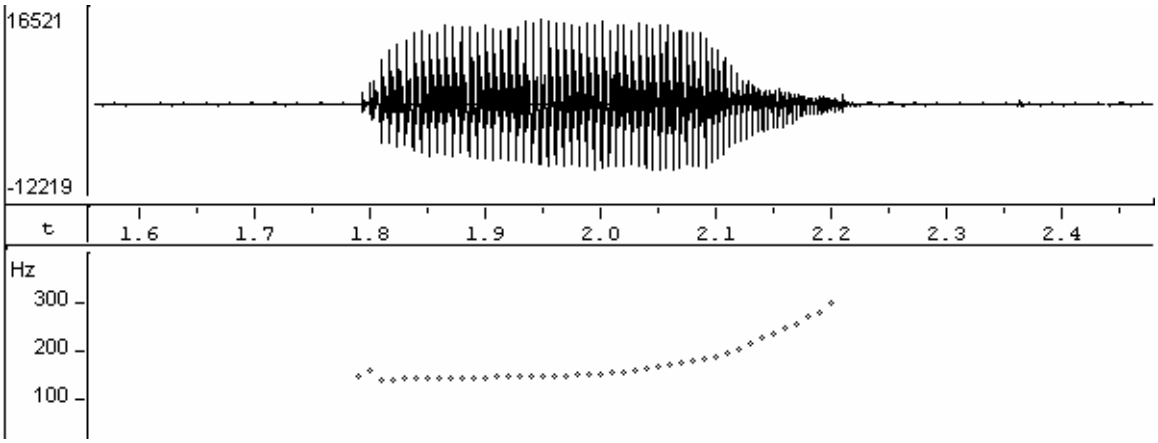
Hình 4.12. Dấu huyền (âm à)

Với dấu huyền, tần số cơ bản giảm dần.

Nếu gọi F_0 là tần số tương ứng với âm không dấu, thì sự thay đổi tần số cơ bản của dấu huyền có thể được mô tả như sau:

F_0, F_0-10, F_0-20, F_0-30, F_0-50, F_0-60

c. Dấu sắc



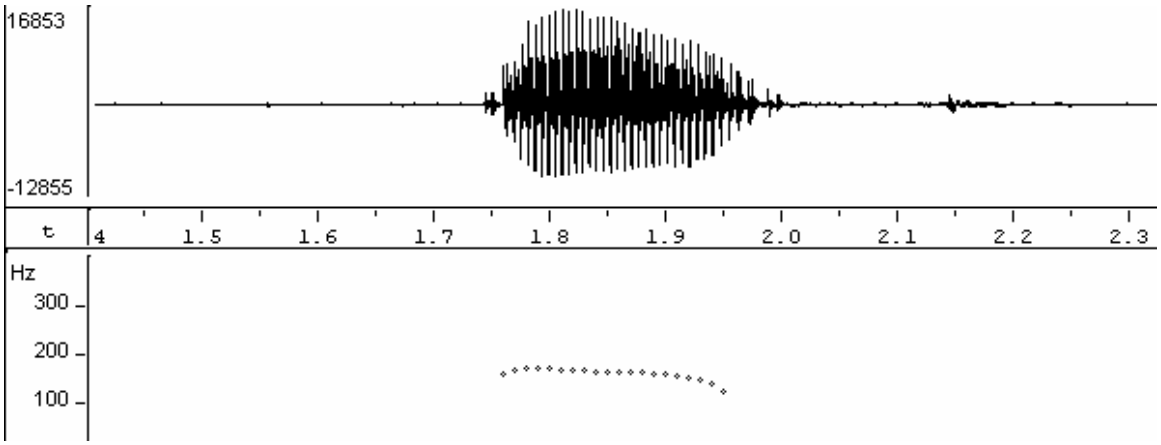
Hình 4.13. Dấu sắc (âm á)

Với dấu sắc, tần số cơ bản tăng dần.

Nếu gọi F_0 là tần số tương ứng với âm không dấu, thì sự thay đổi tần số cơ bản của dấu sắc có thể được mô tả như sau:

F_0-20, F_0-20, F_0-15, F_0-10, F_0-5, F_0+5, F_0+30, F_0+70, F_0+80

d. Dấu nặng

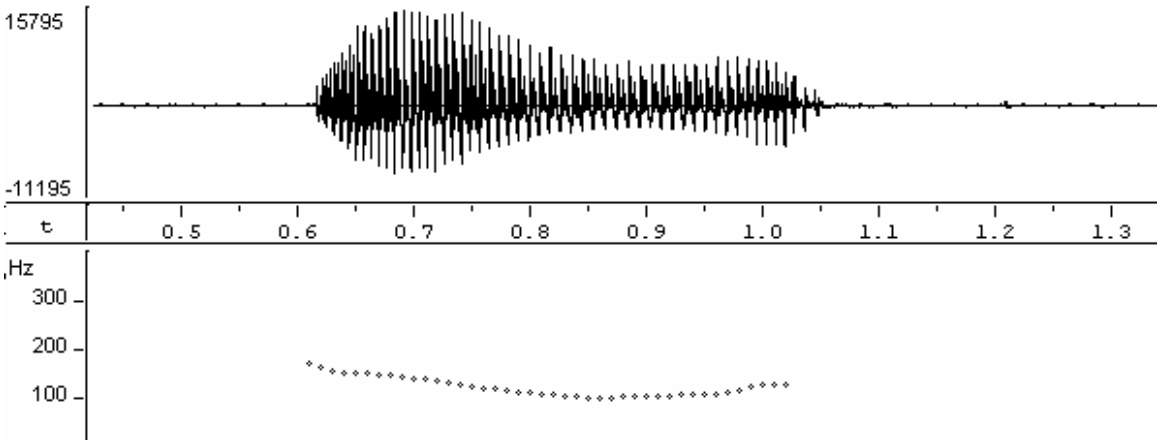


Hình 4.14. Dấu nặng (âm ạ)

Nếu gọi F_0 là tần số tương ứng với âm không dấu, thì sự thay đổi tần số cơ bản của dấu nặng có thể được mô tả như sau:

F_0, F_0, F_0-35, F_0-50, F_0-90, F_0-120, F_0-140

e. Dấu hỏi

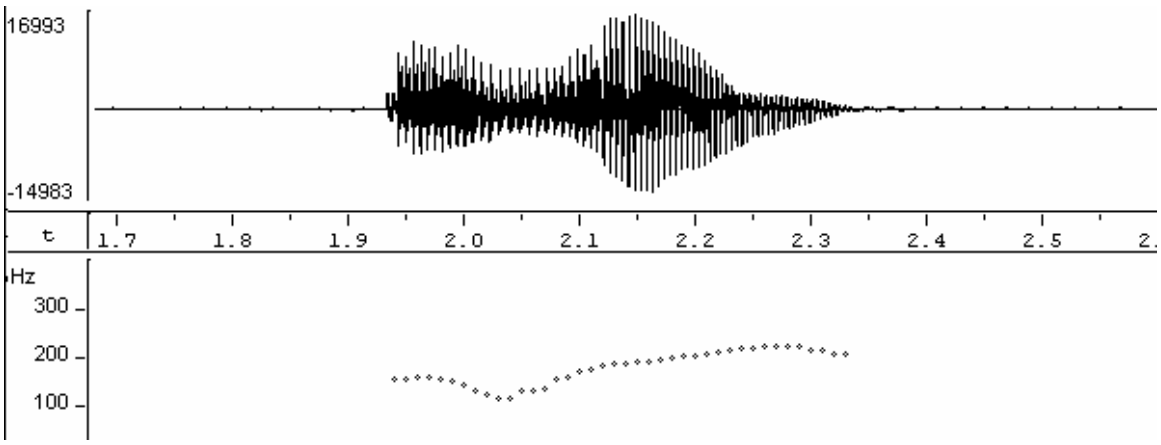


Hình 4.15. Dấu hỏi (âm ả)

Nếu gọi F_0 là tần số tương ứng với âm không dấu, thì sự thay đổi tần số cơ bản của dấu hỏi có thể được mô tả như sau:

F_0-30, F_0-15, F_0-20, F_0-35, F_0-55, F_0-70, F_0-75, F_0-85, F_0-90,F_0-95, F_0-90, F_0-80, F_0-90, F_0-30

f. Dấu ngã



Hình 4.16. Dấu ngã (âm ã)

Nếu gọi F_0 là tần số tương ứng với âm không dấu, thì sự thay đổi tần số cơ bản của dấu ngã có thể được mô tả như sau:

F_0, F_0-40, F_0+20, F_0+50, F_0+60

4.6.2. Sự biến đổi các thông số trong phát âm câu tiếng Việt

Sự thay đổi các thông số của tín hiệu tiếng nói khi phát âm một câu trong tiếng Việt khá phức tạp, vì việc phát âm này phụ thuộc vào nhiều yếu tố như loại câu (câu hỏi, câu trần thuật, câu cảm thán...), hoàn cảnh phát âm (nói chuyện, đọc,...), địa phương... Để có được những hiểu biết về việc phát âm một câu trong tiếng Việt cần có những nghiên cứu đầy đủ.

Với mục đích thử nghiệm việc ghép từ để tạo thành câu trong tiếng Việt, phần này sẽ đưa ra một số nhận xét về sự biến đổi của tín hiệu tiếng nói khi phát âm hai loại câu điển hình của tiếng Việt: câu trần thuật và câu hỏi. Những nhận xét này được rút ra qua sự so sánh với câu không có ngữ điệu.

a. Câu trần thuật

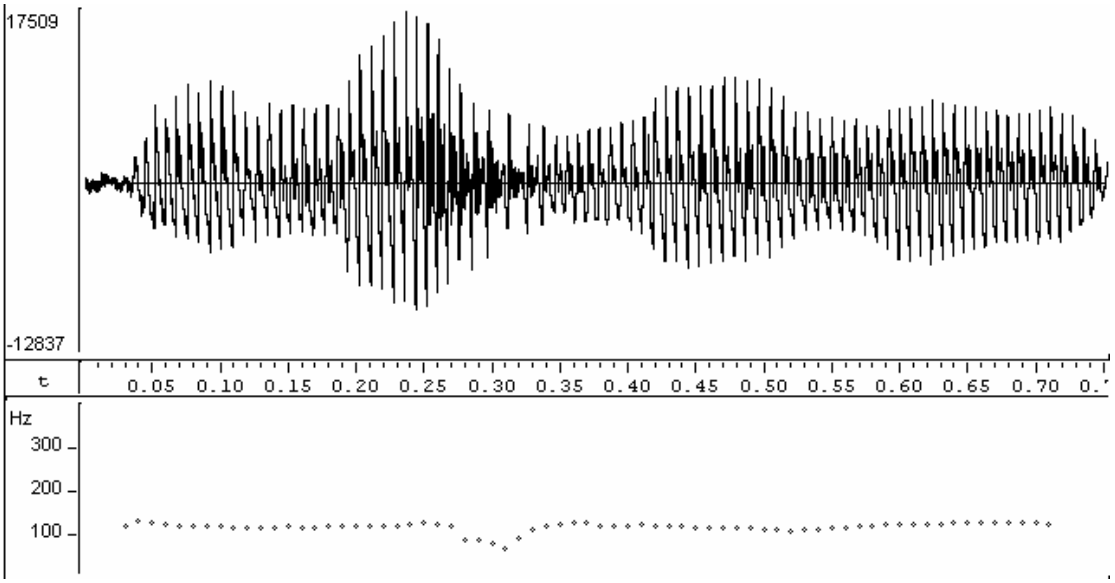
Khi phát âm câu trần thuật, tuỳ theo hoàn cảnh có thể có một số từ nào đó được nhấn mạnh. Việc xác định từ cần nhấn mạnh trong câu trần thuật liên quan tới phân tích bậc cao và không được đề cập tới ở đây. Để đơn giản, giả sử không có từ nào được nhấn mạnh rõ ràng trong câu.

Hình 4.13 và 4.14 là hình ảnh dạng sóng và tần số cơ bản của câu: *Hà Nội ngày nay* (không có ngữ điệu) và câu *Hà Nội ngày nay.* (kết thúc là dấu ".")

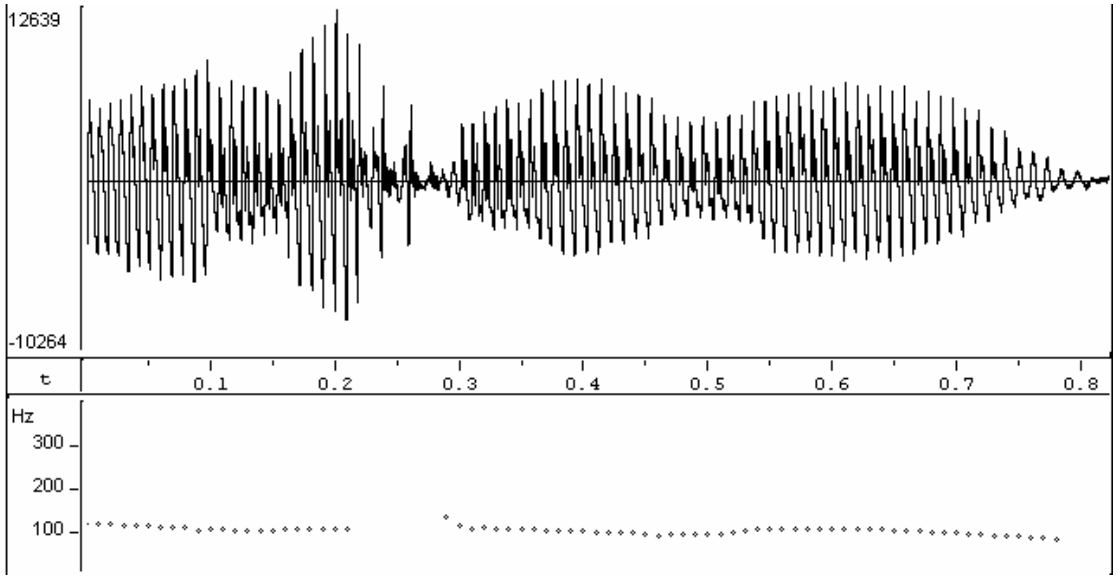
So sánh hai cách phát âm có thể rút ra các nhận xét sau:

- Về thời gian phát âm: Do không có từ nhấn mạnh nên các từ trong câu không ngữ điệu và câu trần thuật được phát âm trong khoảng thời gian gần như nhau.
- Về biên độ tín hiệu: Các từ trong câu không ngữ điệu được phát âm với biên độ tương đối đều. Biên độ các từ trong câu trần thuật giảm dần ở cuối câu.
- Về tần số cơ bản: Trong câu không ngữ điệu, tần số cơ bản của các từ (không có thanh điệu) đi theo đường nằm ngang. Tần số cơ bản của từ trong câu trần thuật giảm dần.

Như vậy, các từ trong câu trần thuật được phát âm với biên độ và tần số cơ bản giảm dần về phía cuối câu.



Hình 4.17. Câu không ngữ điệu *Hà Nội ngày nay*



Hình 4.18. Câu trần thuật *Hà Nội ngày nay.*

b. Câu hỏi

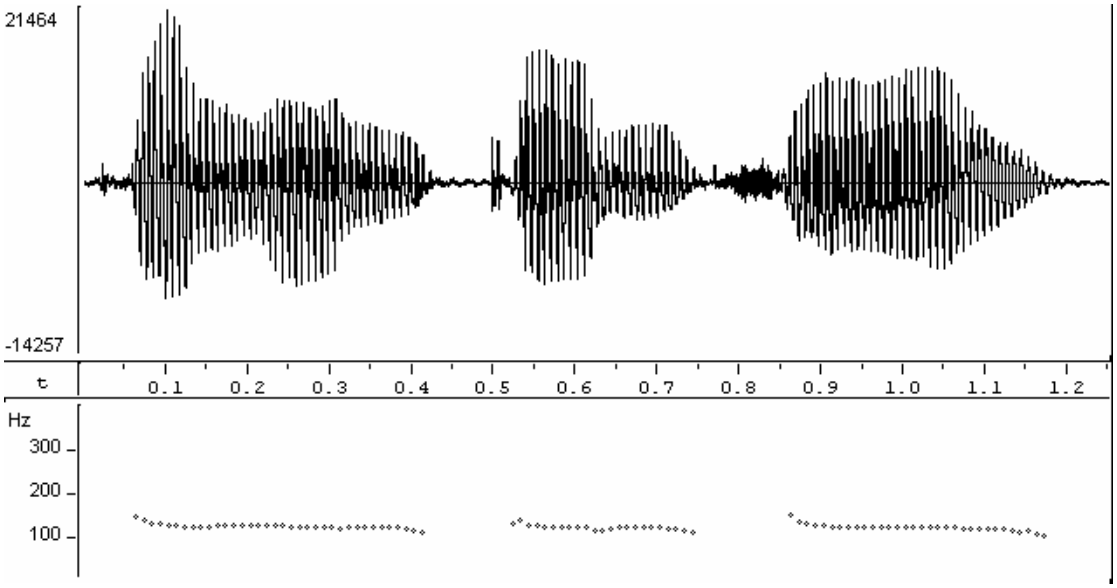
Trong câu hỏi, người nói thường nhấn mạnh vào từ cần hỏi. Những từ cần hỏi này thường không có vị trí cố định trong câu.

Ví dụ: Cùng một câu hỏi *Anh đi?* Nếu muốn hỏi về chủ ngữ (anh hoặc ai đó) thì người hỏi sẽ nhấn mạnh vào từ *anh*, nếu muốn hỏi về hành động (đi hoặc chạy) thì người hỏi sẽ nhấn mạnh vào từ *đi*.

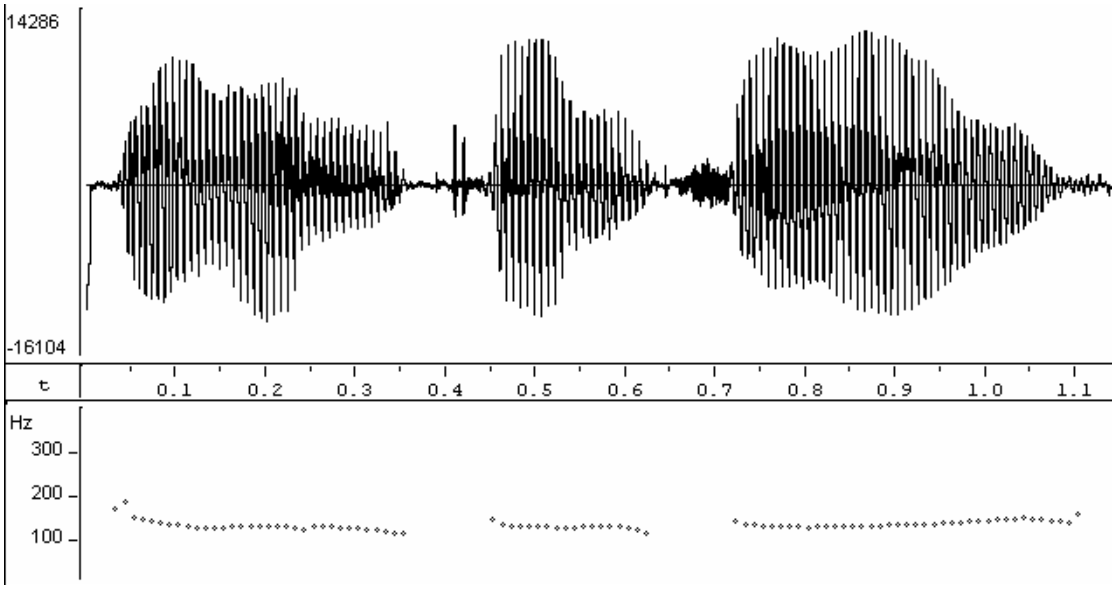
Việc xác định từ để hỏi trong câu liên quan tới việc phân tích bậc cao trong quá trình tổng hợp và không được đề cập ở đây. Để đơn giản, từ để hỏi

trong các câu thử nghiệm được coi là từ cuối câu. Câu hỏi sẽ được so sánh với câu không có ngữ điệu.

Dưới đây là hình ảnh dạng sóng và tần số cơ bản của câu: *Anh ăn chưa* (không có ngữ điệu) và câu *Anh ăn chưa?* (từ để hỏi là *chưa*)



Hình 4.19. Câu tiếng Việt không ngữ điệu: *Anh ăn chưa*



Hình 4.20. Câu hỏi *Anh ăn chưa?* với từ để hỏi *chưa*

So sánh hai cách phát âm có thể rút ra các nhận xét sau:

- Về thời gian phát âm: Các từ trong câu không ngữ điệu được phát âm trong khoảng thời gian gần như nhau. Từ để hỏi trong câu hỏi (*chưa*) được phát âm dài hơn (0.45s) các từ *anh* (0.35s) và *ăn* (0.20s) trong câu này.

- Về biên độ tín hiệu: Các từ trong câu không ngữ điệu được phát âm với biên độ tương đối đều. Từ để hỏi *chưa* trong câu hỏi được phát âm với biên độ lớn hơn từ *chưa* trong câu không ngữ điệu.
- Về tần số cơ bản: Trong câu không ngữ điệu, tần số cơ bản của các từ (không có thanh điệu) đi theo đường nằm ngang. Tần số cơ bản của từ *anh* và *ăn* trong câu hỏi không tăng dần. Tần số cơ bản của từ *chưa* trong câu hỏi tăng dần.

Như vậy, các từ để hỏi trong câu hỏi được phát âm dài hơn, với biên độ lớn hơn và tần số cơ bản tăng dần so với câu không ngữ điệu.

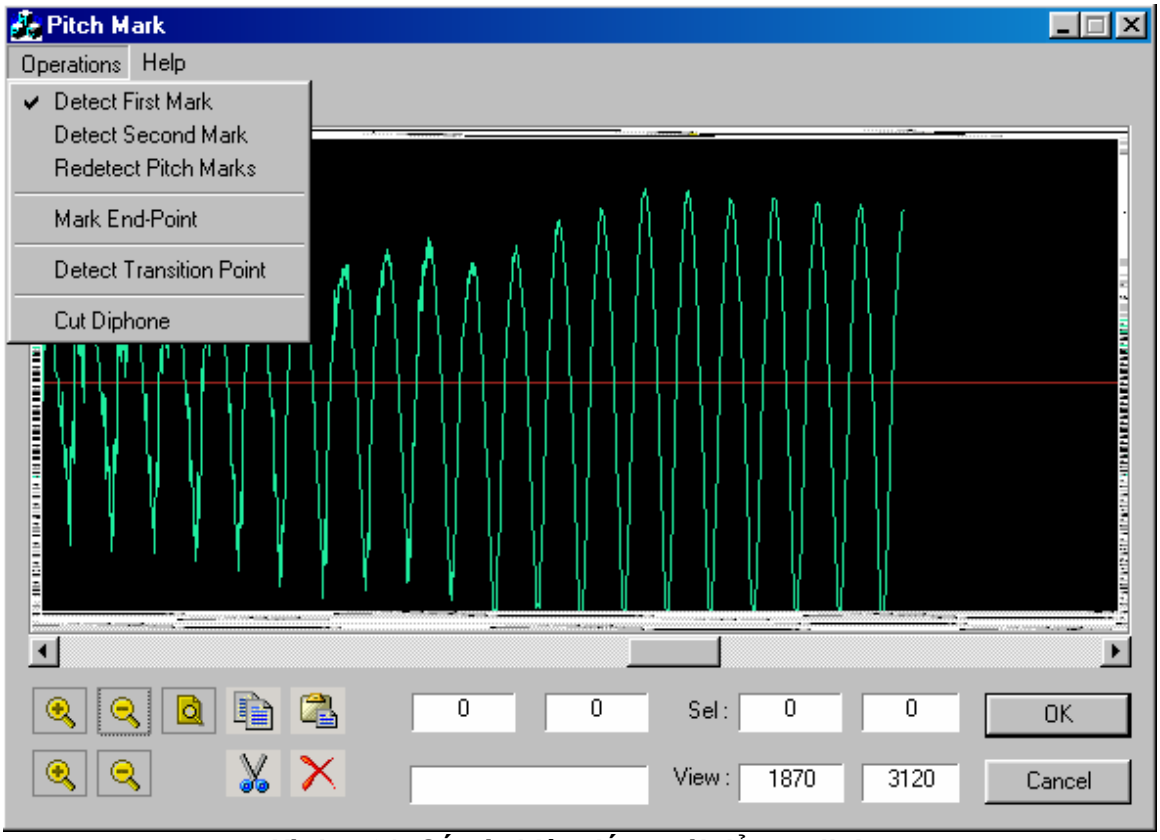
4.7. CHƯƠNG TRÌNH TỔNG HỢP TIẾNG VIỆT

Chương trình tổng hợp tiếng Việt được xây dựng với mục đích là tổng hợp tiếng nói từ văn bản tiếng Việt bằng giải thuật TD-PSOLA. Chương trình gồm các chức năng:

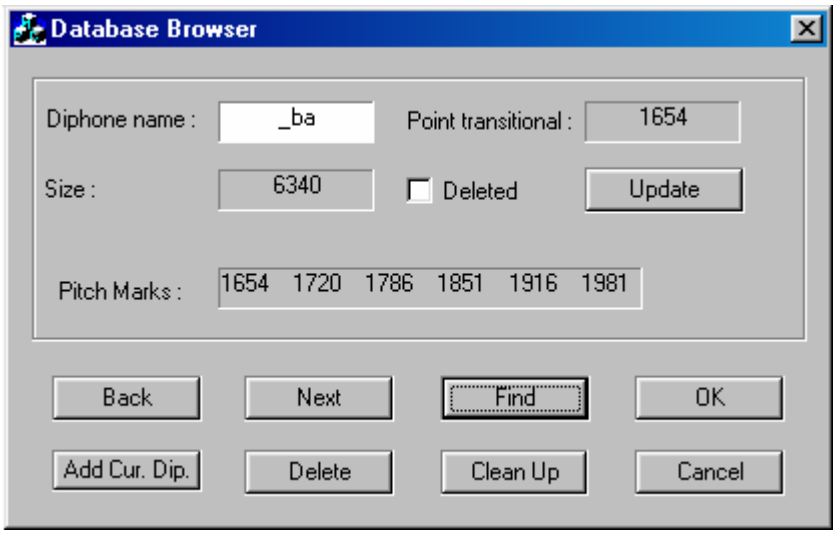
- Tách diphone từ mẫu tiếng nói có sẵn.
- Phát âm tiếng Việt được ghi trong file văn bản (*.txt)

4.7.1. Tách diphone từ mẫu tiếng nói có sẵn

Tín hiệu tiếng nói sau khi thu được lưu dưới dạng file wav. Với chức năng tách diphone, người sử dụng có thể tạo ra các diphone và lưu vào cơ sở dữ liệu.



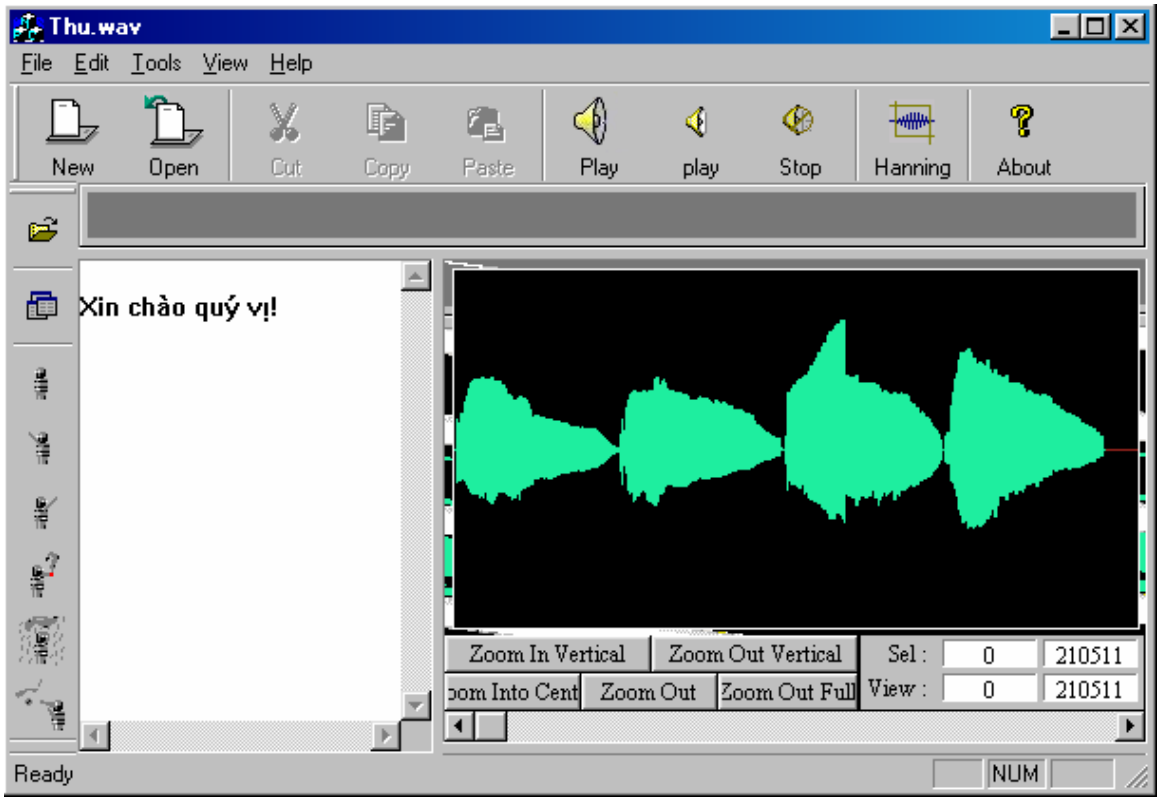
Hình 4.21. Cắt tín hiệu tiếng nói để tạo diphone



Hình 4.22. Lưu trữ diphone vào cơ sở dữ liệu

Chức năng này cho phép thêm, xoá dữ liệu trong file cơ sở dữ liệu.

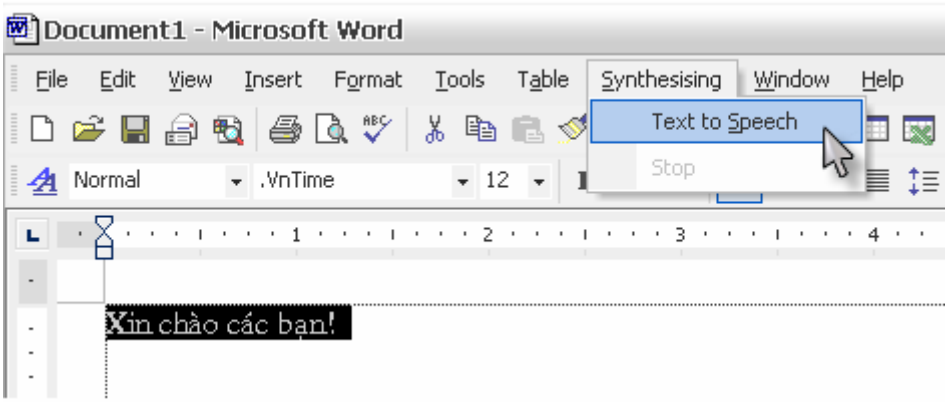
4.7.2. Phát âm tiếng Việt



Hình 4.23. Phát âm tiếng Việt

Chức năng này cho phép chương trình phát âm đoạn văn bản được ghi sẵn trong file có dạng text (*.txt). Người dùng cũng có thể gõ trực tiếp văn bản để chương trình phát âm.

Ứng dụng cũng được tích hợp vào chương trình Microsoft Word dưới dạng một menu và cho phép đọc được các văn bản thông thường, soạn thảo bằng font ABC.



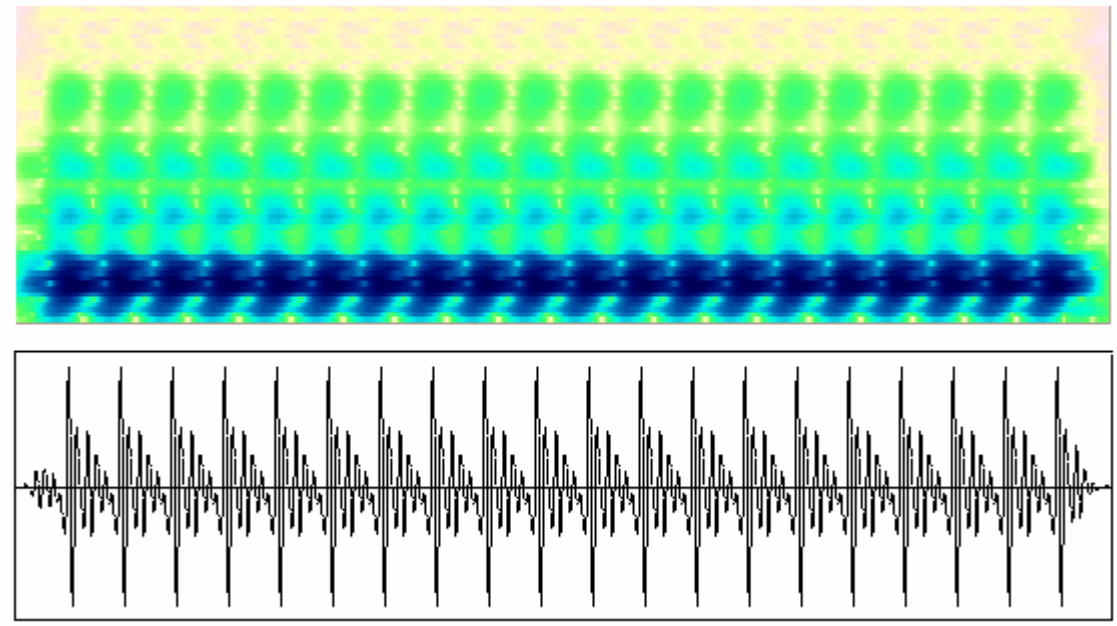
Hình 4.24. Tích hợp ứng dụng vào Microsoft Word

4.8. KẾT QUẢ ĐẠT ĐƯỢC

4.8.1. Tổng hợp các nguyên âm

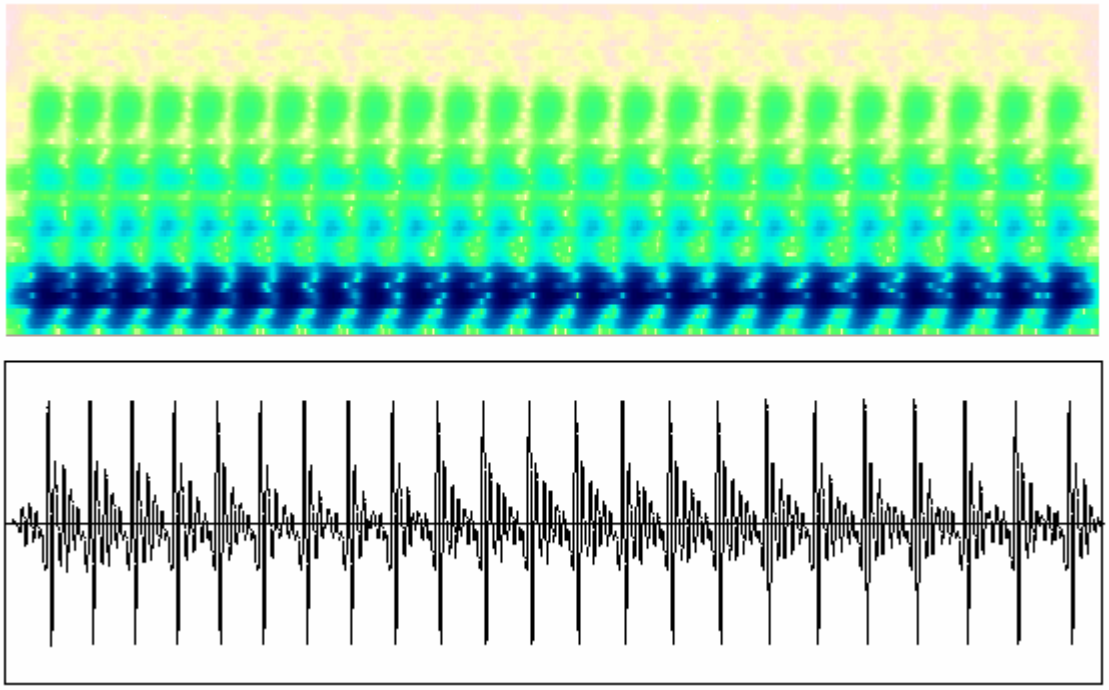
Chương trình cho phép tổng hợp các nguyên âm không dấu và có dấu với chất lượng tương đối tốt. Kết quả tổng hợp các nguyên âm như sau:

a. Nguyên âm a

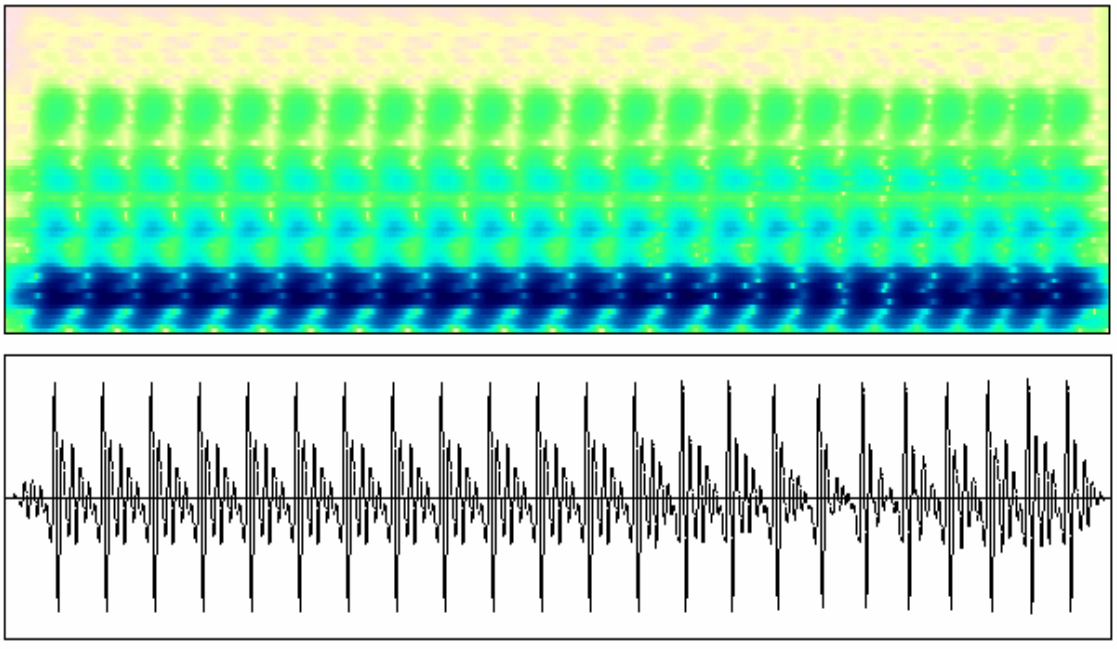


Âm **a**

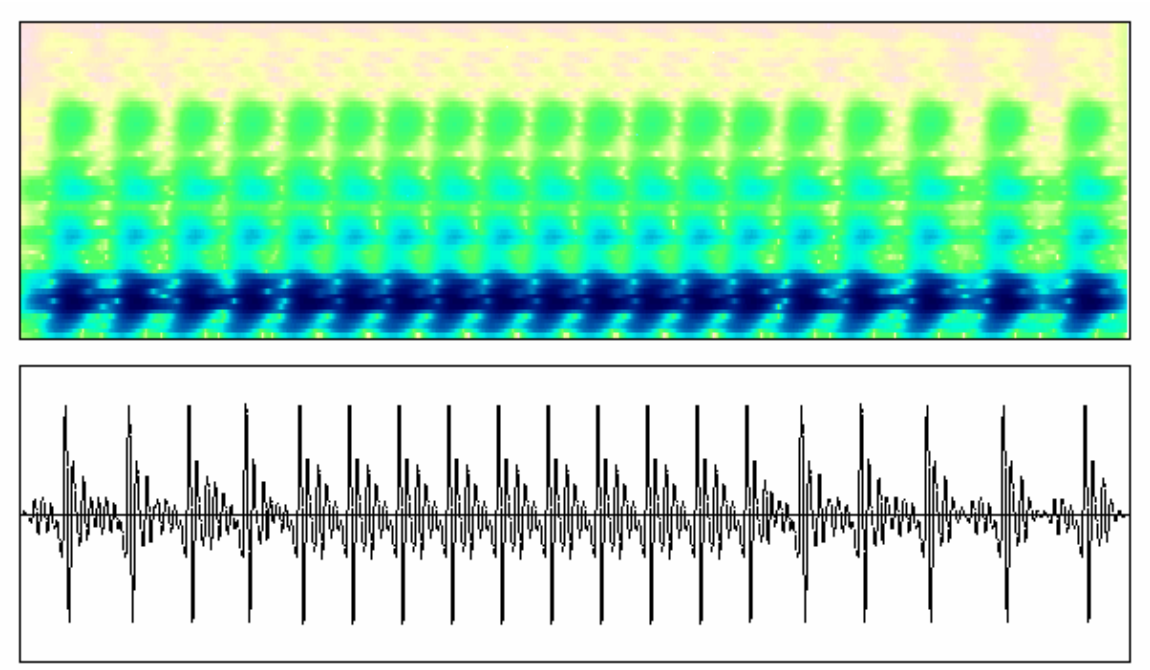
Biến đổi tần số cơ bản của nguyên âm **a** để được các âm **à, á, ạ, ả, ã**



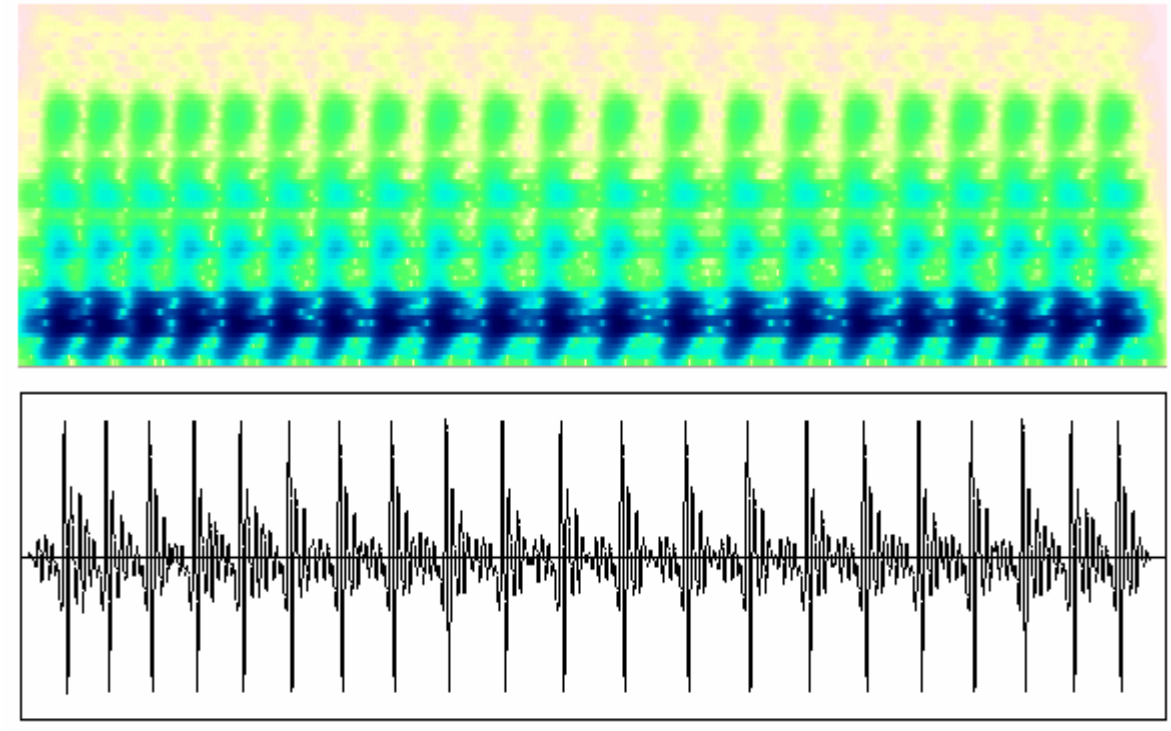
Âm **à**



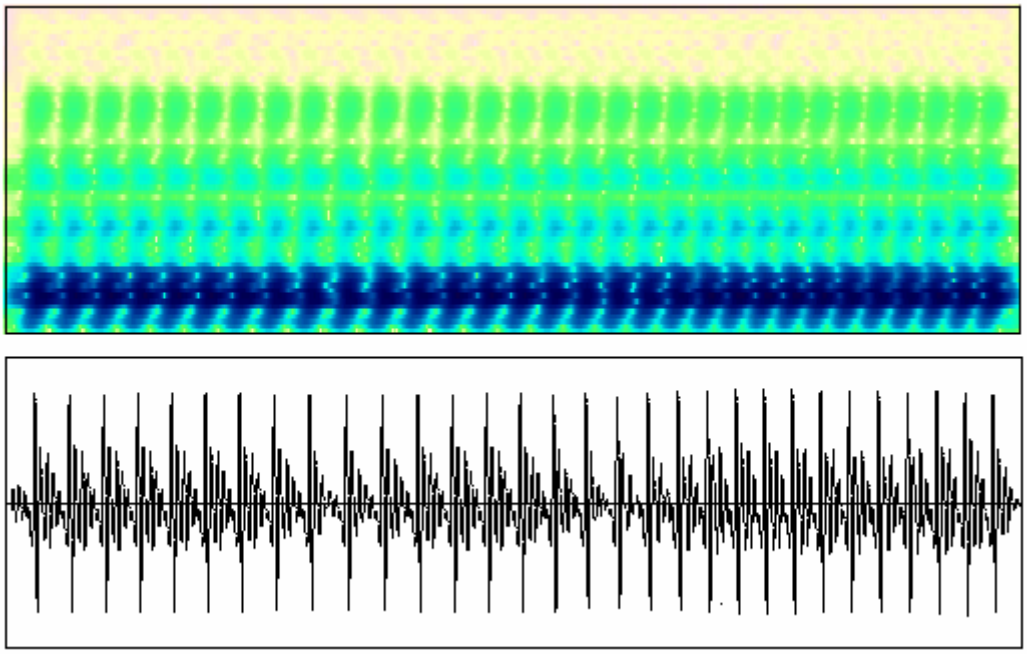
Âm **á**



Âm **ạ**

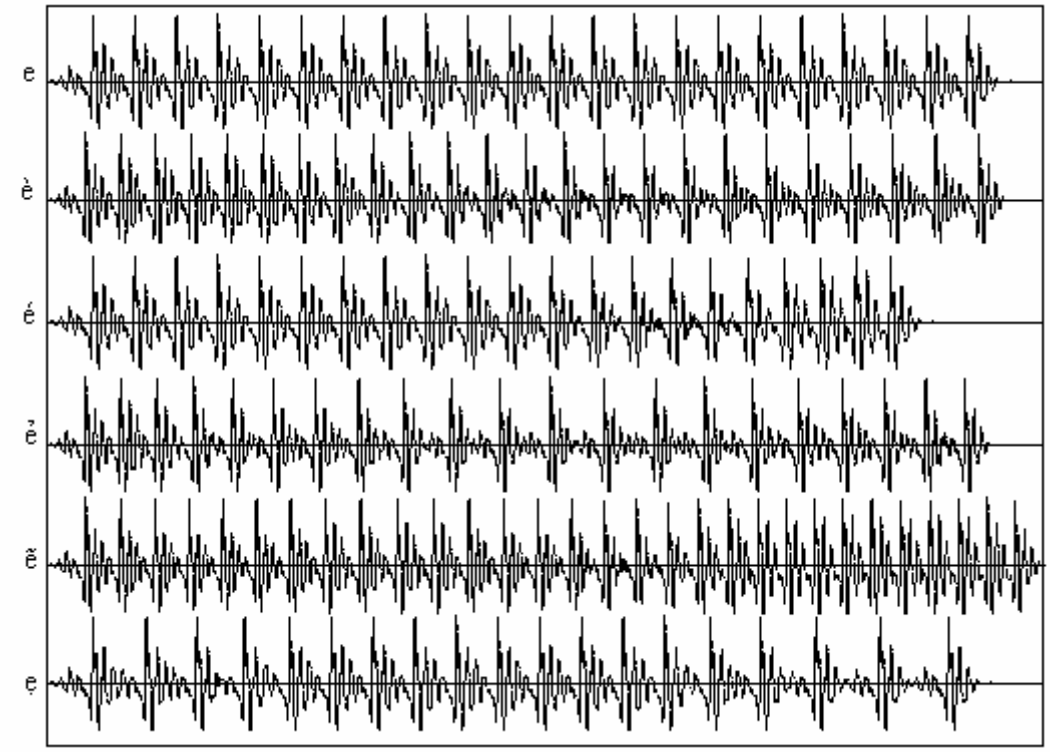


Âm **ả**

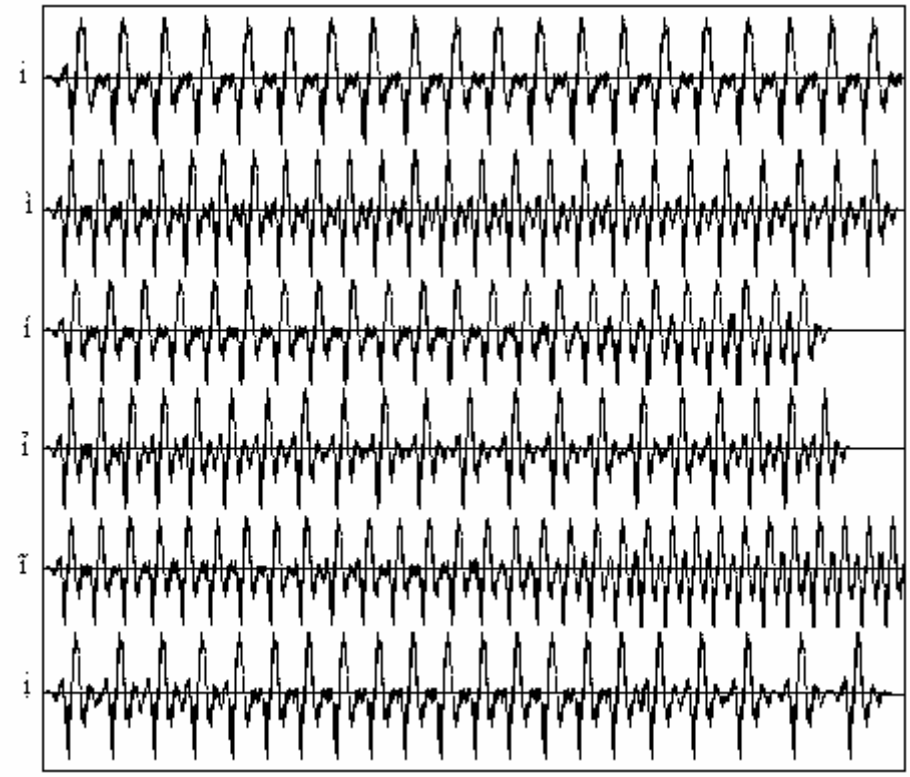


Âm **ã**

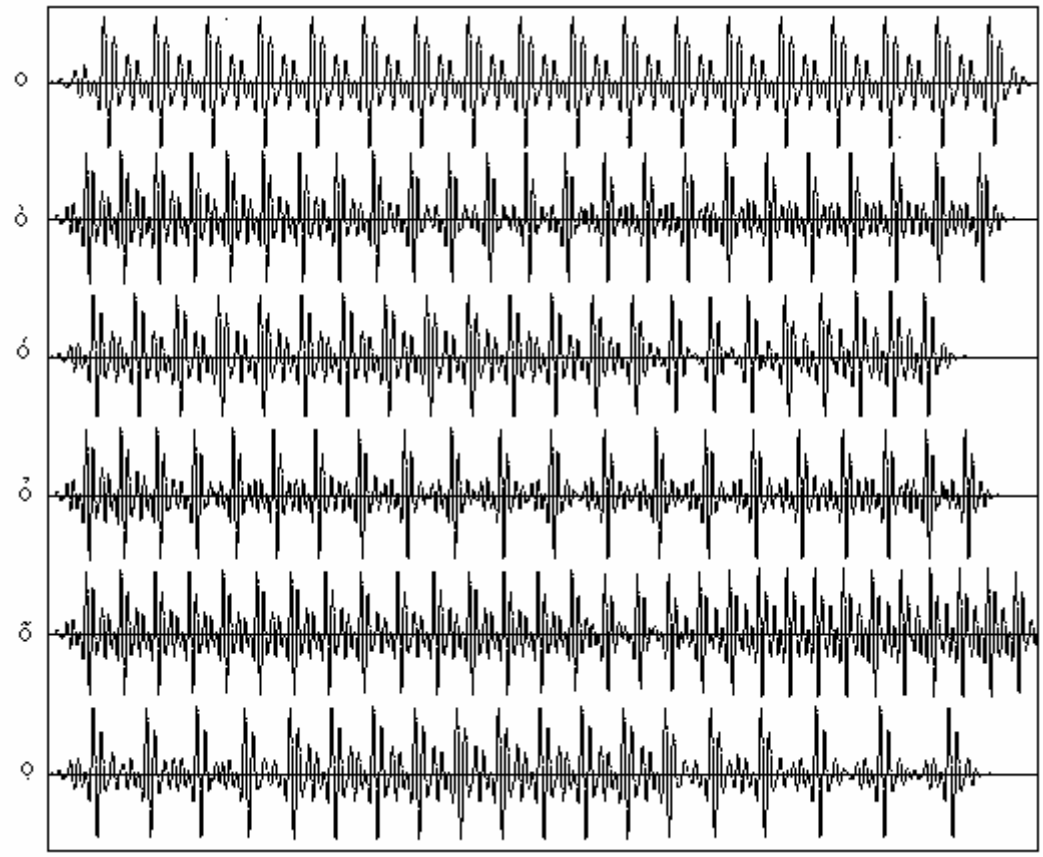
b. Các âm e, è, é, ẻ, ẽ, ẹ



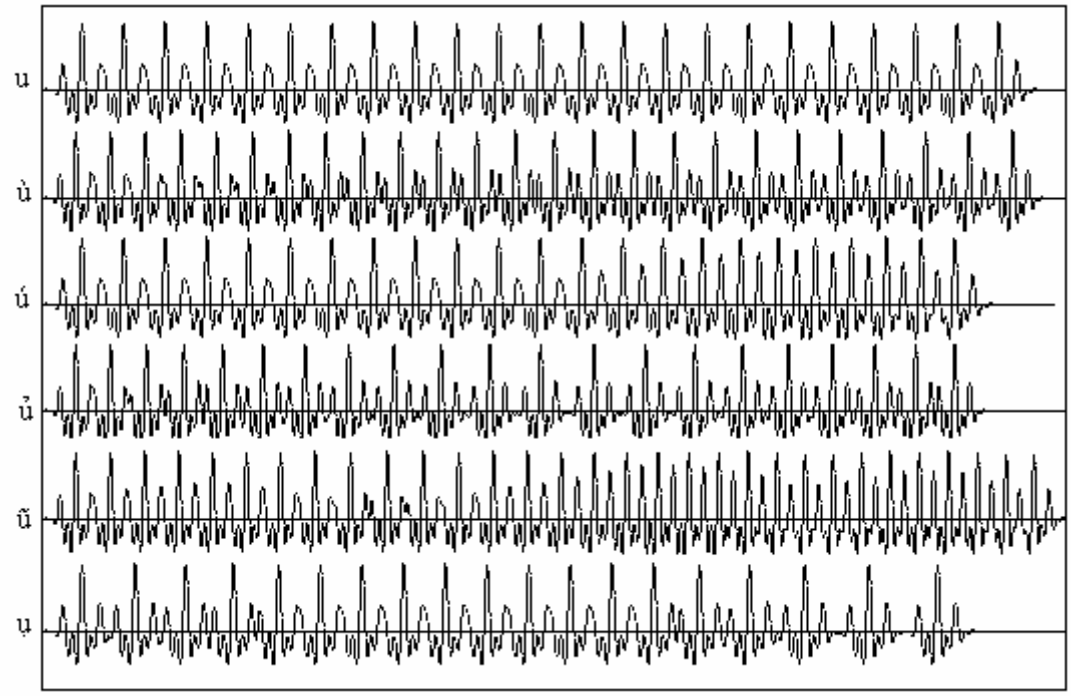
c. Các âm i, ì, í, ỉ, ĩ, ị



d. Các âm o, ò, ó, ỏ, õ, ọ



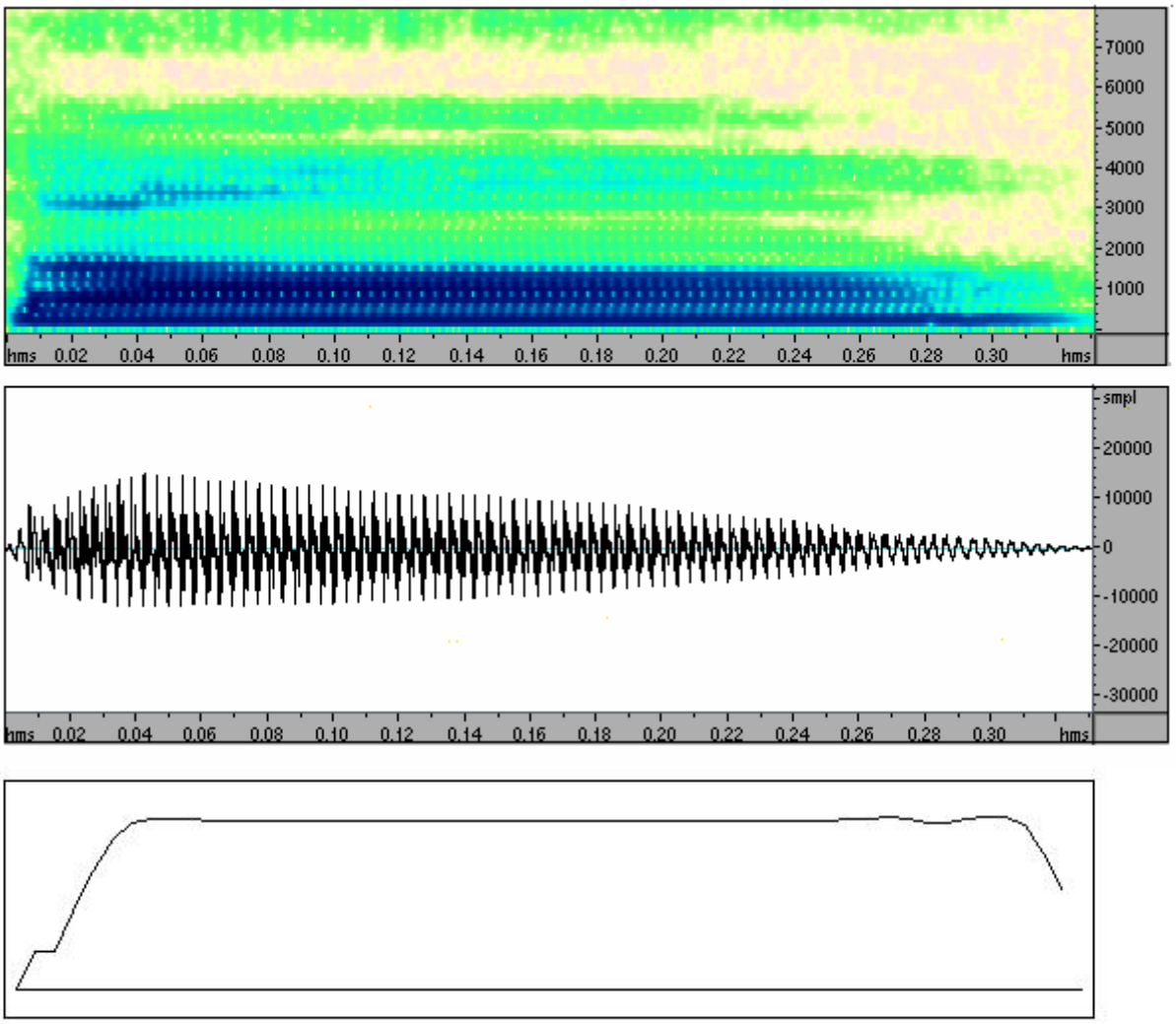
e. Các âm u, ù, ú, ủ, ũ, ụ



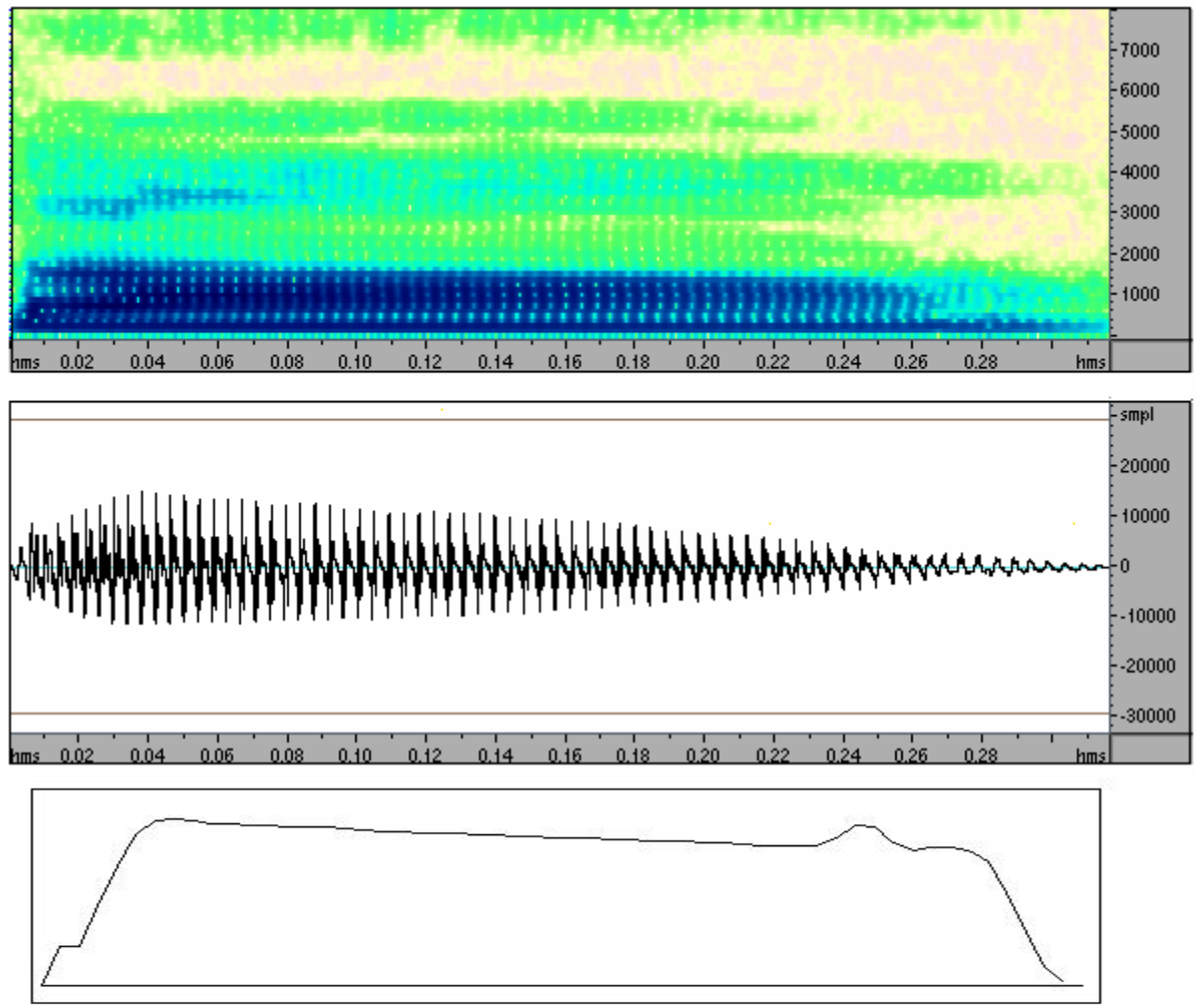
4.8.2. Tổng hợp từ

Với từ **to** được tổng hợp từ 2 diphone **_to** và **o_**, khi biến đổi tần số cơ bản ta có được các từ **tò, tó, tỏ, tõ, tọ**

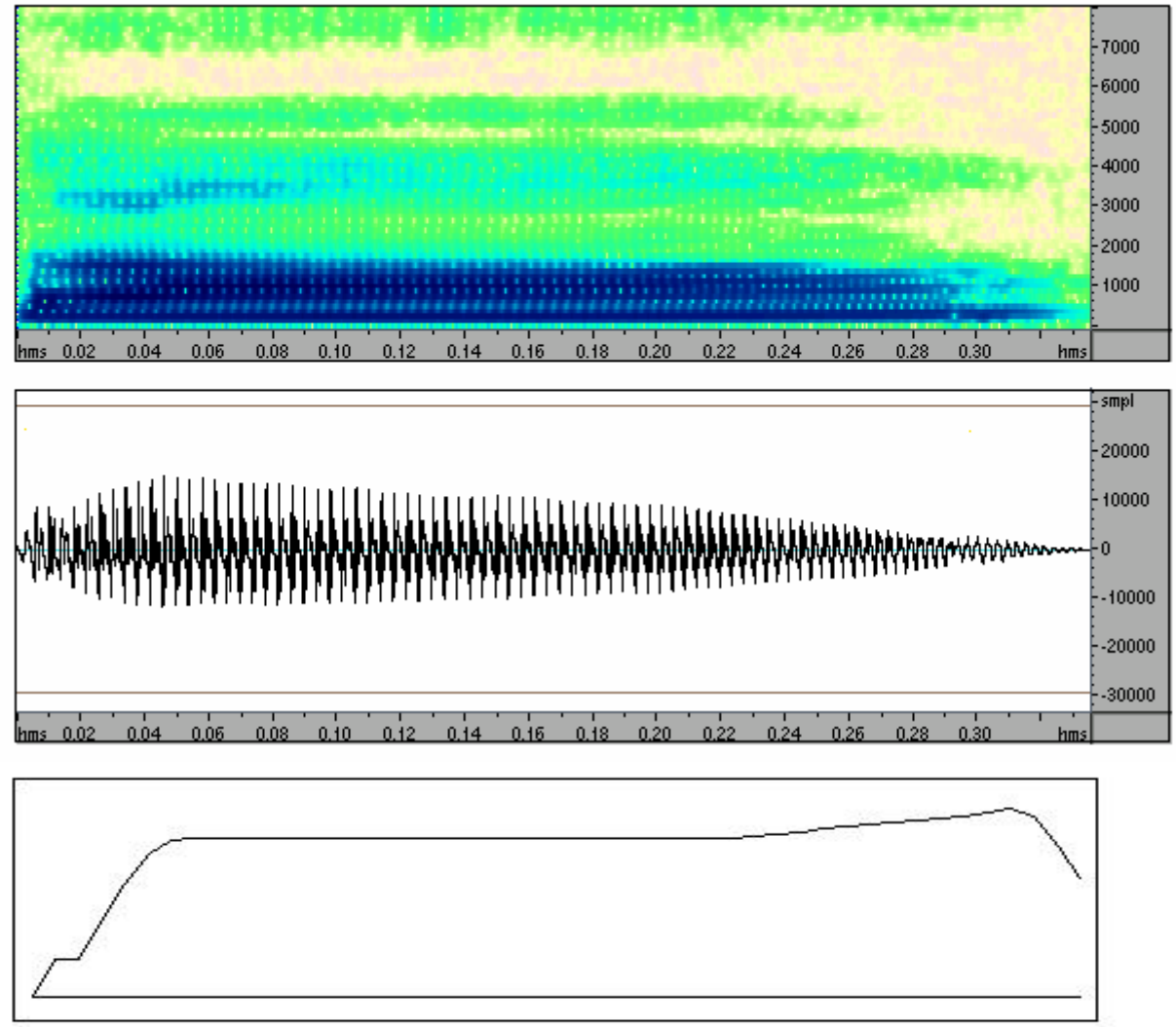
a. Từ *to*



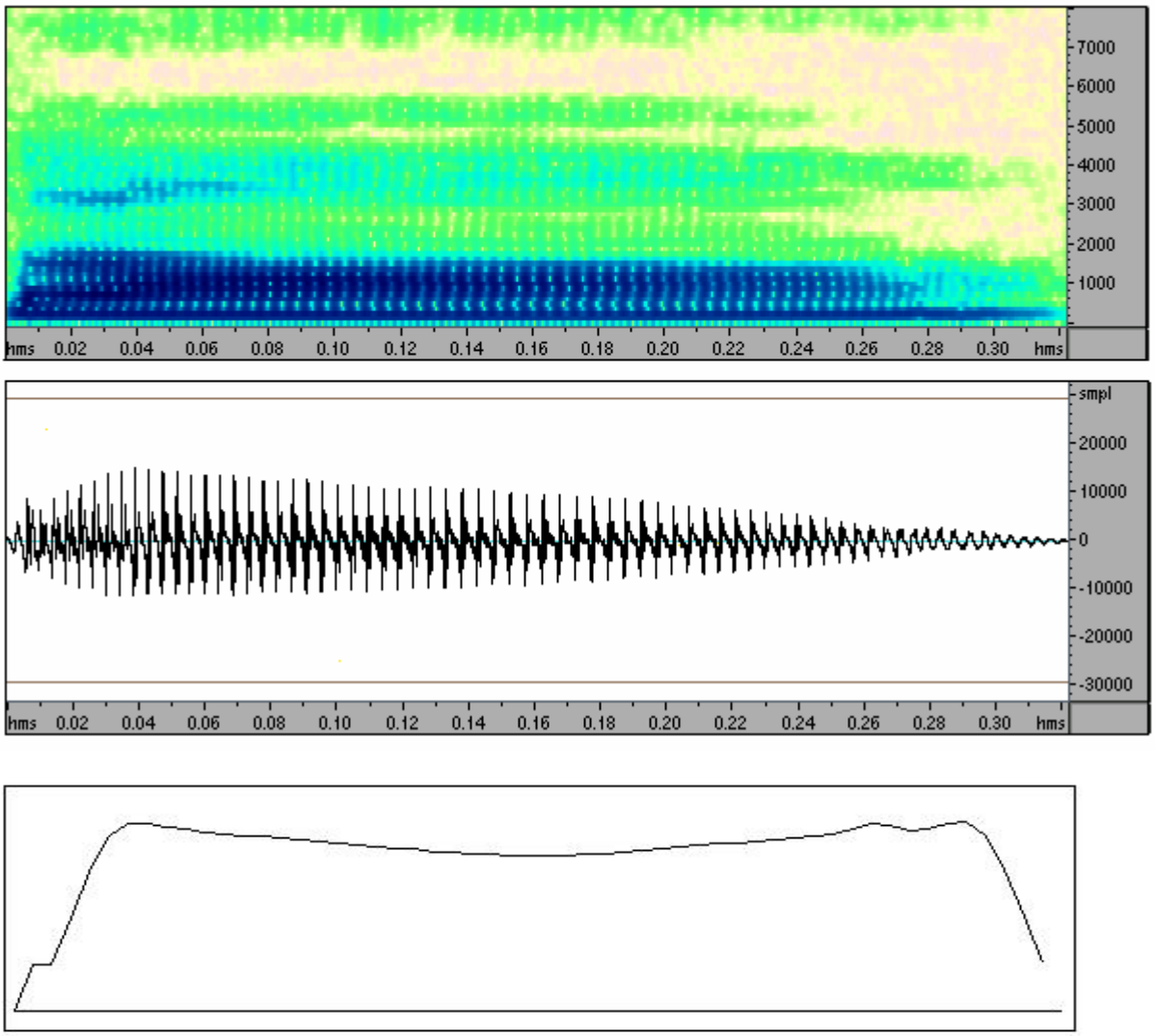
b. Từ *tò*


7000
6000
5000
4000
3000
2000
1000
hms 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 hms
smpl
20000
10000
0
-10000
-20000
-30000
hms 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 hms

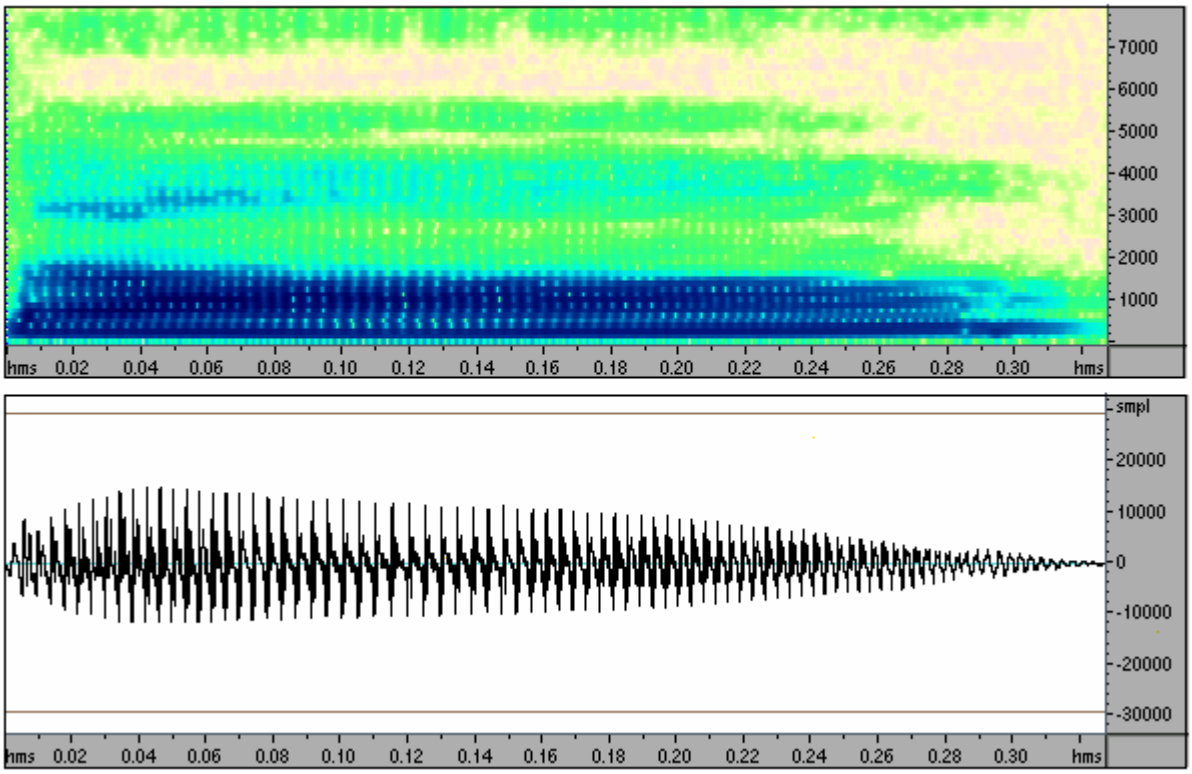
c. Từ *tó*

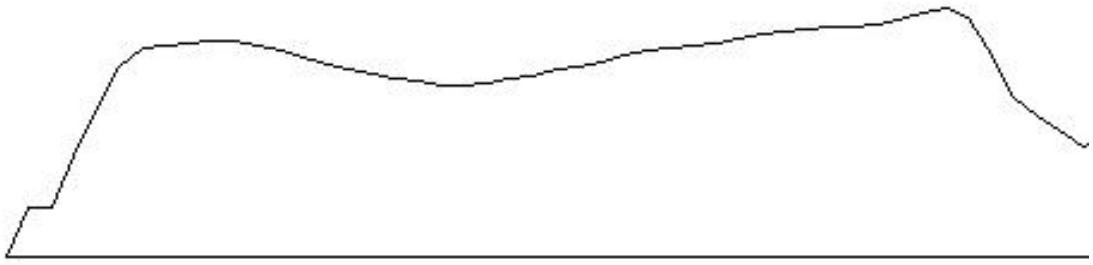

7000
6000
5000
4000
3000
2000
1000
hms 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30 hms
smpl
20000
10000
0
-10000
-20000
-30000
hms 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30 hms

d. Từ *tỏ*

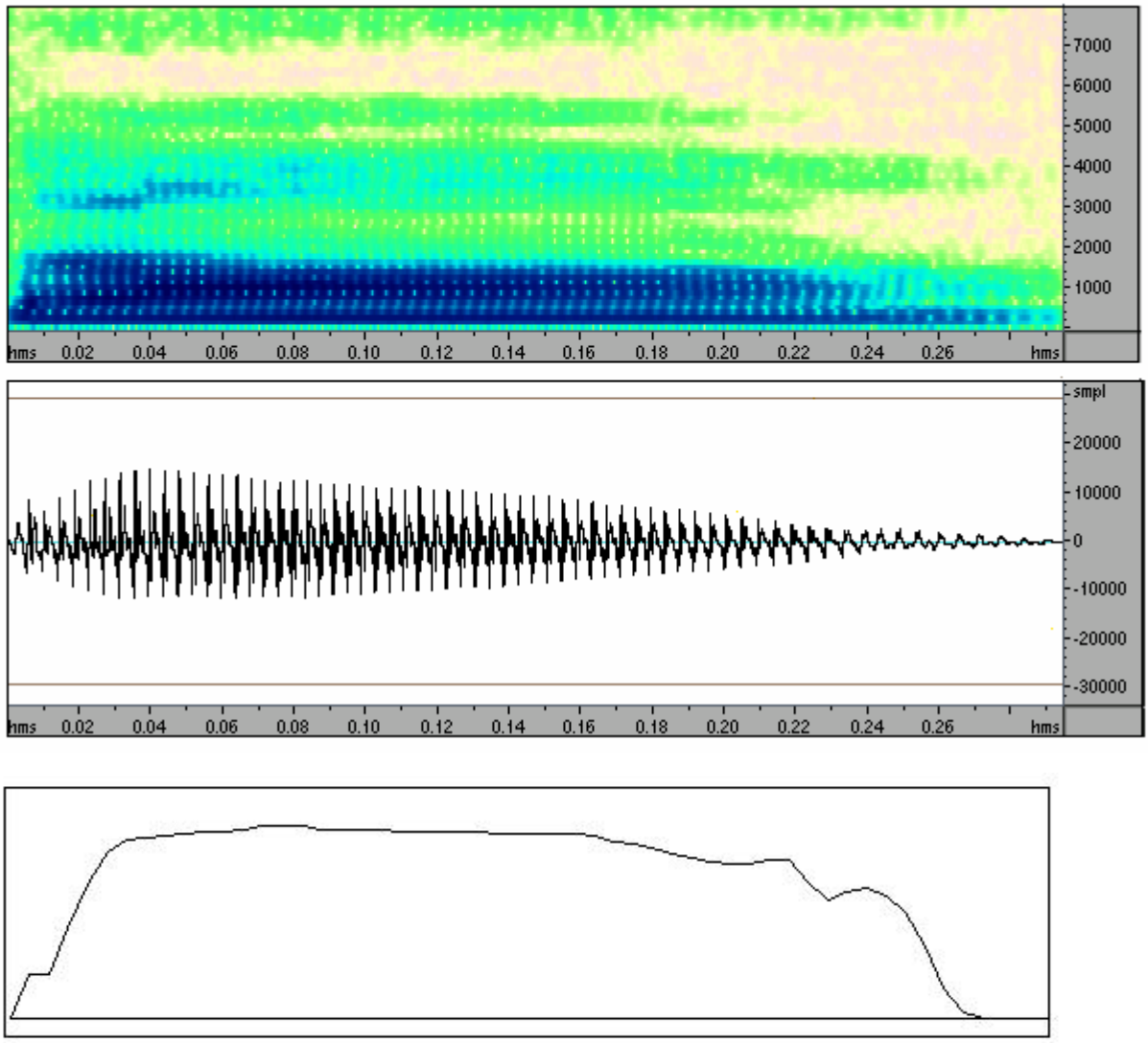

7000
6000
5000
4000
3000
2000
1000
hms 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30 hms
smpl
20000
10000
0
-10000
-20000
-30000
hms 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30 hms

e. Từ *tõ*

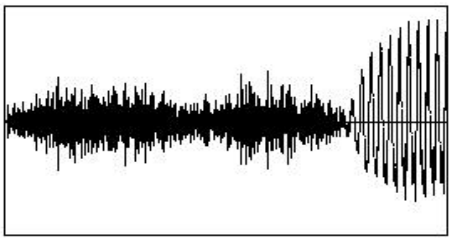

7000
6000
5000
4000
3000
2000
1000
hms 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30 hms
smpl
20000
10000
0
-10000
-20000
-30000
hms 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30 hms



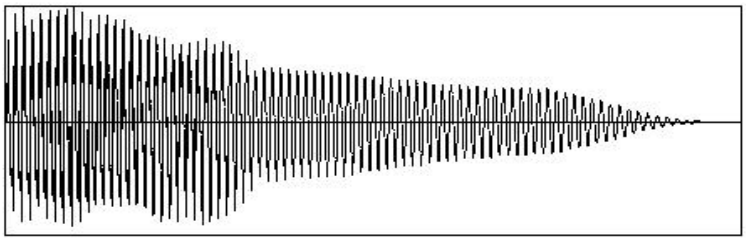
f. Từ tọ


7000
6000
5000
4000
3000
2000
1000
hms 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 hms
smpl
20000
10000
0
-10000
-20000
-30000
hms 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 hms

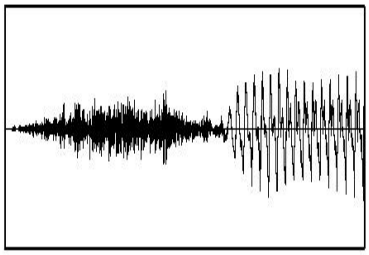
4.8.3. Tổng hợp từ "Xin chào"



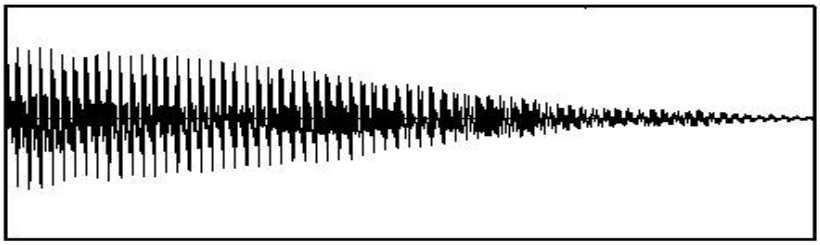
_xi



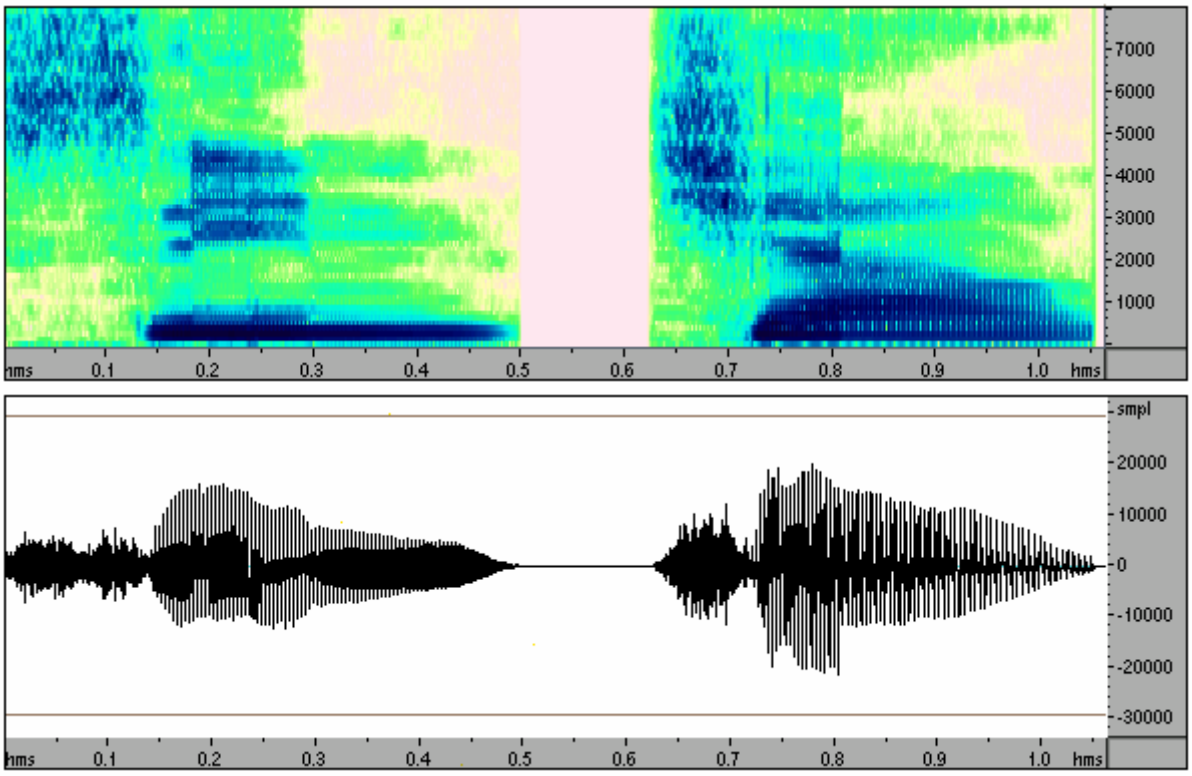
in_



_cha



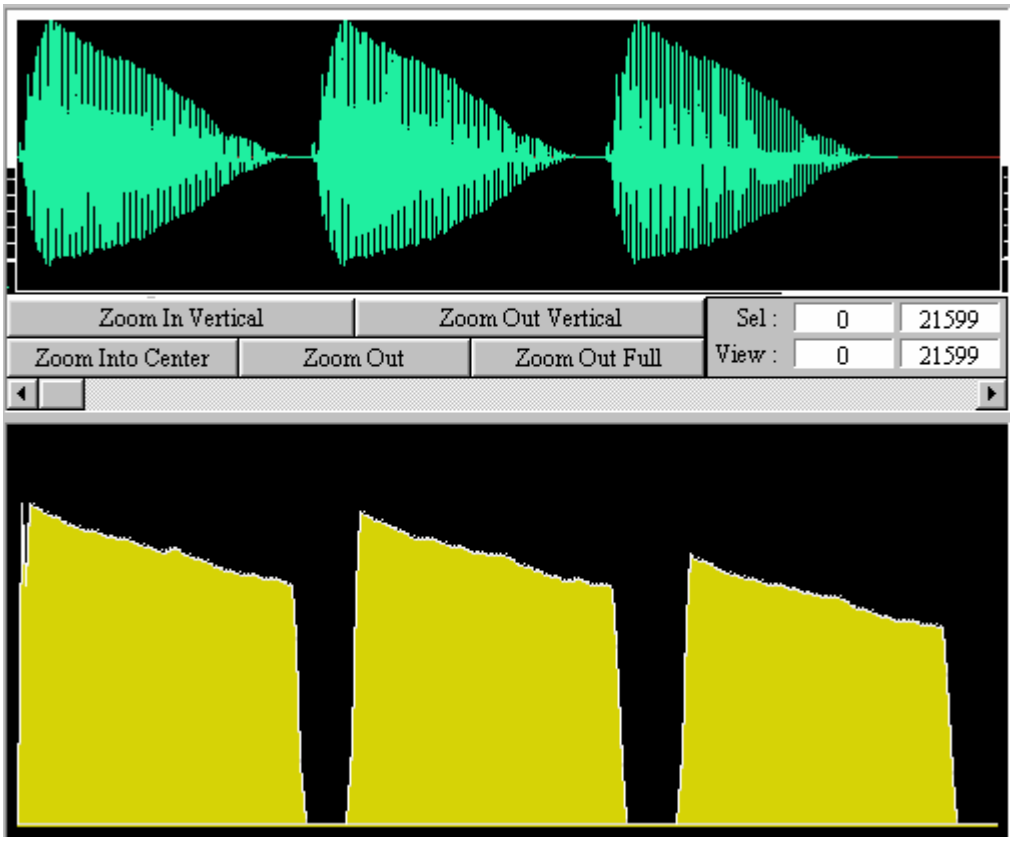
ao_



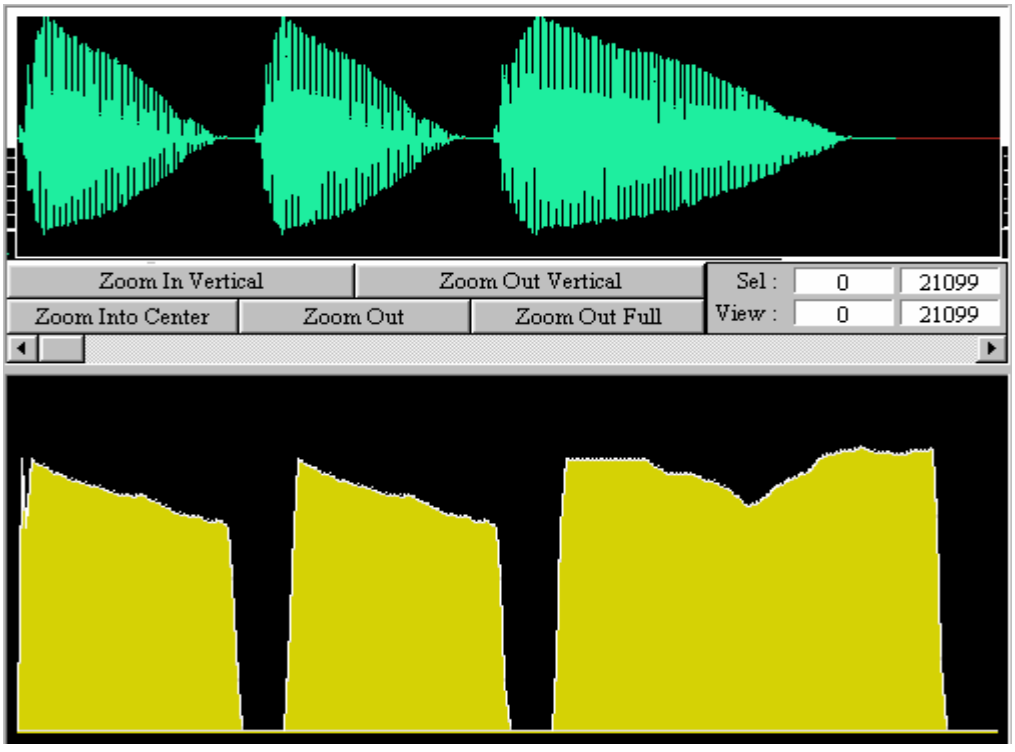
«Xin chào»

4.8.4. Tổng hợp câu

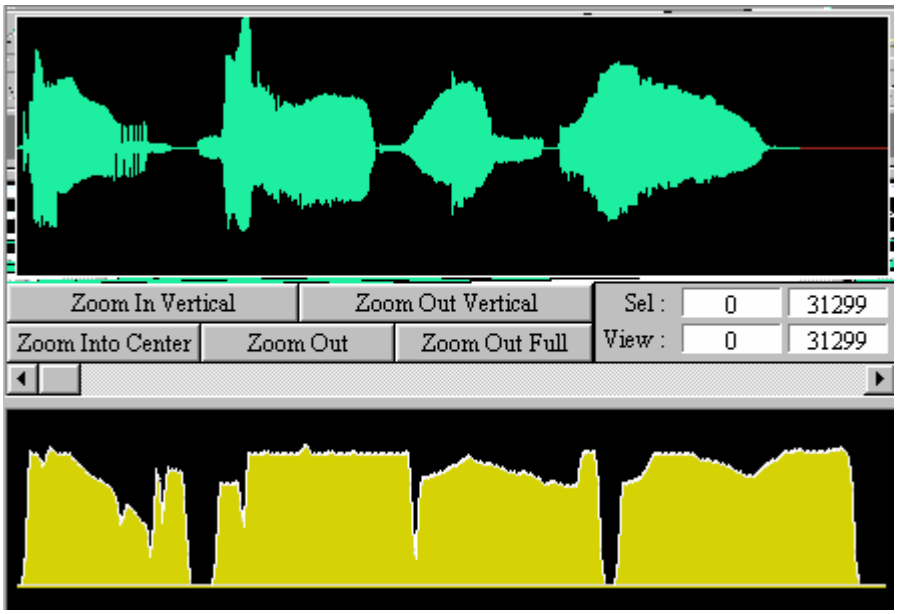
a. Câu trần thuật *Tò tò tò.*



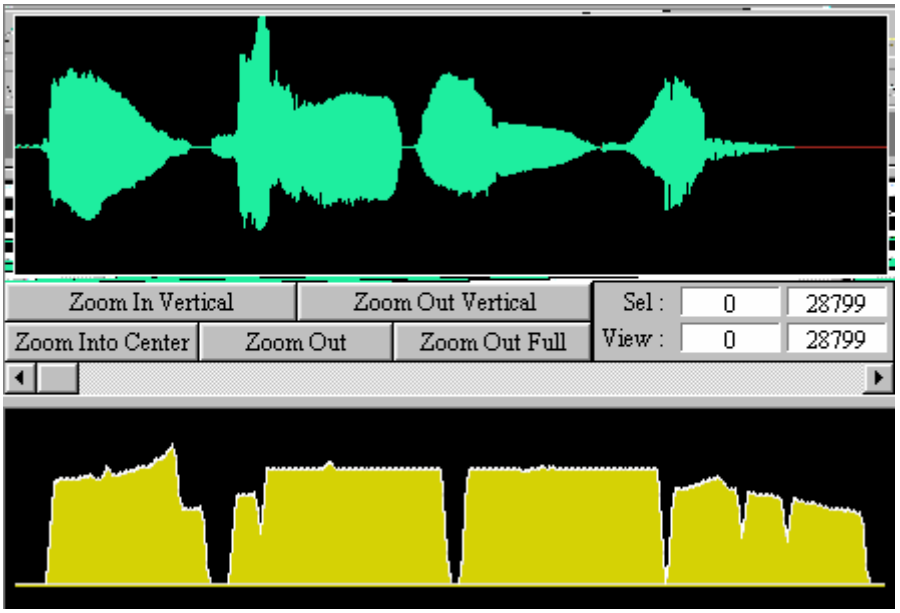
b. Câu hỏi tò tò tò?



c. Tổng hợp câu hỏi *Cậu đang làm gì?*



d. Tổng hợp câu trần thuật *Tớ đang ôn bài.*



KẾT LUẬN

1. Đánh giá kết quả

Với mục đích nghiên cứu phương pháp tổng hợp tiếng Việt bằng cách ghép nối các diphone, sau khi xây dựng chương trình và quan sát kết quả thu được, có thể rút ra một số nhận xét như sau:

a. Biến đổi tần số cơ bản tạo ra các thanh điệu

Kết quả tổng hợp với những từ có thể tạo nên từ các từ không dấu cho thấy có thể tạo nên các thanh điệu trong tiếng Việt tương ứng với dấu huyền, sắc, nặng, hỏi, ngã từ thanh điệu không dấu bằng cách biến đổi tần số cơ bản một cách thích hợp. Các từ tổng hợp thuộc nhóm này có chất lượng khá tốt. Ngoài ra, điều này còn cho phép giảm số lượng diphone, từ đó giảm kích thước của cơ sở dữ liệu.

Với những từ không thể tạo nên từ các từ không dấu không thể áp dụng việc biến đổi tần số cơ bản như trên. Vì các diphone được thu bởi một giọng duy nhất nên giải pháp hiện tại là giữ nguyên tần số cơ bản của các diphone tương ứng. Chất lượng tổng hợp từ thuộc nhóm này chưa cao.

b. Tổng hợp các loại câu đơn giản trong tiếng Việt

Cùng với khả năng biến đổi tần số cơ bản của tín hiệu tiếng nói tổng hợp, giải thuật TD-PSOLA còn cho phép thay đổi độ dài thời gian của tín hiệu tổng hợp. Điều này được áp dụng khi tổng hợp thử nghiệm hai loại câu trong tiếng Việt là câu trần thuật và câu hỏi.

Với câu trần thuật, chương trình tổng hợp bằng cách tạo ra tín hiệu với tần số cơ bản và biên độ giảm dần ở cuối câu. Với câu hỏi, chương trình tổng hợp bằng cách tăng thời gian phát âm, biên độ và tần số của từ cuối câu (tương ứng với việc nhấn mạnh từ để hỏi). Các kết quả này mang tính chất thử nghiệm và cho chất lượng chấp nhận được.

c. Cơ sở dữ liệu diphone

Khi tổng hợp tiếng nói bằng ghép nối thì khó khăn lớn nhất số lượng từ vựng dùng để tổng hợp quá lớn. Nhưng phương pháp tổng hợp từ các diphone trong tiếng Việt rõ ràng đã khắc phục được khó khăn này.

Sau khi xây dựng cơ sở dữ liệu diphone hoàn chỉnh, kích thước của cơ sở dữ liệu này là 2.37 MB (2 495 572 byte) với 389 diphone (xấp xỉ 6482byte/1diphone) . Kết quả lưu trữ này có thể thấy rõ nếu so sánh với một cơ sở dữ liệu gồm các từ tiếng Việt trong bảng sau:

	CSDL Diphone	**CSDL Từ**
Kích thước 1 mẫu	6482 byte / 1 diphone	13000 byte / 1từ
Số lượng mẫu	389 diphone	> 5000 từ
Kích thước cơ sở dữ liệu	2.37 MB	> 62 MB

2. Phương hướng phát triển đề tài

Để tổng hợp tiếng Việt với chất lượng tốt, với những kết quả đã đạt được, đề tài này có thể phát triển tiếp để giải quyết các vấn đề sau:

- Việc tổng hợp các từ tiếng Việt từ những diphone có dấu cho kết quả chưa cao. Vấn đề này cần được nghiên cứu kỹ hơn vì số lượng từ được tạo ra trực tiếp từ các diphone có dấu không ít.
- Sự thay đổi các thông số của tín hiệu tiếng nói trong các loại câu khác nhau cũng là một vấn đề khá quan trọng. Nếu nắm được đầy đủ đặc trưng biến đổi tín hiệu tiếng nói của các loại câu thì chất lượng tiếng nói tổng hợp sẽ được nâng cao.
- Tuy đã có một cơ sở dữ liệu diphone đầy đủ nhưng công việc thu và tách diphone từ các mẫu tiếng nói vẫn được thực hiện hoàn toàn thủ công, chính vì vậy các diphone có chất lượng chưa cao. Trong tương lai cần có các phương pháp xây dựng cơ sở dữ liệu thích hợp để khắc phục nhược điểm này. Ngoài ra có thể bổ sung thêm các giọng khác vào cơ sở dữ liệu (giọng nam, người già, trẻ em...) để tăng tính đa dạng.
- Đề tài này mới chỉ nghiên cứu về tổng hợp mức thấp mà chưa nghiên cứu nhiều về tổng hợp mức cao nên mới chỉ xử lý các văn bản đơn giản thành các diphone tương ứng phục vụ cho tổng hợp tín hiệu tiếng nói. Các nghiên cứu đầy đủ về tổng hợp mức cao chắc chắn sẽ giúp các ứng dụng tổng hợp tiếng Việt trở nên hoàn thiện hơn.

PHỤ LỤC

1. Phụ lục 1: Bảng các diphone tiếng Việt

_a	bê	_e	hu	lo	nhơ	ơn	se	**uật**	**ượt**
a_	bi	e_	hư	lô	nhu	**ớp**	sê	**úc**	**ứt**
ác	bo	em	hy	lơ	như	**ợp**	si	**ục**	**ựt**
ạc	bô	en	_i	lu	_o	**ớt**	so	uê	ưu
ách	bơ	eng	i_	lư	o_	**ợt**	sô	ui	va
ạch	bu	eo	ia	ly	oa	pa	sơ	um	ve
ai	bư	**ép**	**ích**	ma	**oát**	pe	su	un	vê
am	ca	**ẹp**	**ịch**	me	**oạt**	pê	sư	ung	vi
an	co	**ét**	**iếc**	mê	oai	pi	ta	**uốc**	vo
ang	cô	**ẹt**	**iệc**	mi	oan	po	te	**uộc**	vô
anh	cơ	_ê	iêng	mo	oang	pô	tê	uôi	vơ
ao	cu	ê_	**iếp_**	mô	oanh	pơ	ti	uôn	vu
áp	cư	**ếch**	**iệp_**	mơ	**óc**	pu	to	uôm	vư
ạp	cha	**ệch**	**iết**	mu	**ọc**	pư	tô	**uốt**	xa
át	che	êm	**iệt**	mư	oe	py	tơ	**uột**	xe
ạt	chê	ên	iêu	my	oi	pha	tu	uông	xê
au	chi	ênh	im	na	om	phe	tư	**úp**	xi
ay	cho	**ếp**	in	ne	on	phê	tha	**ụp**	xo
ắc	chô	**ệp**	inh	nê	ong	phi	the	**út**	xơ
ặc	chơ	**ết**	**íp**	ni	oong	pho	thê	**ụt**	xu
ăm	chu	**ệt**	**ịp**	no	**óp**	phô	thi	uy	xư
ăn	chư	Êu	**ít**	nô	**ọp**	phơ	tho	uya	xy
ăng	da	ga	**ịt**	nơ	**ót**	phu	thô	uyên	_y
ắp	de	gi	iu	nu	**ọt**	phư	thơ	**uyết**	y_
ặp	dê	gia	ke	nư	_ô	qua	thu	**uyệt**	yêm
ắt	di	ghe	kê	nga	ô_	que	thư	**uýt**	yên
ặt	do	ghê	ki	nghe	**ốc**	quê	tra	**uỵt**	**yết**
ấc	dô	ghi	kha	nghê	**ộc**	qui	tre	_ư	**yệt**
ậc	dơ	go	khe	nghi	ôi	quơ	trê	ư_	yêu
âm	du	gô	khê	ngo	ôm	quy	tri	**ức**	

ân	dư	gơ	kho	ngô	ôn	ra	tro	**ực**	
âng	đa	gu	khô	ngơ	ông	re	trô	ưm	
ấp	đe	gư	khơ	ngu	**ốp**	rê	trơ	ưn	
ập	đê	ha	khu	ngư	**ộp**	ri	tru	ưng	
ất	đi	he	khư	nha	**ốt**	ro	trư	**ước**	
ật	đo	hê	ky	nhe	**ột**	ro	_u	**ược**	
âu	đô	hi	la	nhê	_ơ	rơ	u_	ươi	
ây	đơ	ho	le	nhi	ơ_	ru	ua	ươn	
ba	đu	hô	lê	nho	ơi	rư	uân	ương	
be	đư	hơ	li	nhô	ơm	sa	**uất**	**ướt**	

Chú thích: Các diphone in đậm là các diphone có dấu.

2. Phụ lục 2: Bảng mã TCVN3-ABC của các ký tự tiếng Việt

Ký tự	Mã TCVN3	Ký tự	Mã TCVN3
à	97	ò	223
á	184	ó	227
ả	182	ỏ	225
ã	183	õ	226
ạ	185	ọ	228
ă	168	ô	171
ằ	187	ồ	229
ắ	190	ố	232
ẳ	188	ổ	230
ẵ	189	ỗ	231
ặ	198	ộ	233
â	169	ơ	172
ầ	199	ờ	234
ấ	202	ớ	237
ẩ	200	ở	235
ẫ	201	ỡ	236
ậ	203	ợ	238
è	204	ù	239
é	208	ú	243
ẻ	206	ủ	241
ẽ	207	ũ	242
ẹ	209	ụ	244
ê	170	ư	173
ề	210	ừ	245
ế	213	ứ	248
ể	211	ử	246
ễ	212	ữ	247
ệ	214	ự	249
ì	215	ỳ	250
í	221	ý	253
ỉ	216	ỷ	251
ĩ	220	ỹ	252
ị	222	ỵ	254
đ	174		

3. Phụ lục 3: Tên các diphone dài trong cơ sở dữ liệu

Diphone	Tên	Diphone	Tên	Diphone	Tên	Diphone	Tên
ác_	acs_	ạc_	acj_	ách_	a1s_	ạch	a1j_
ang_	ag_	áp_	aps_	ạp_	apj_	át	ats_
ạt_	atj_	ắc_	1cs_	ặc_	1cj_	ăng_	awg_
ắp_	1ps_	ặp_	1pj_	ắt_	1ts_	ặt_	1tj_
ấc_	2cs_	ậc_	2cj_	âng_	aag_	ấp_	2ps_
ập_	2pj_	ất_	2ts_	ật_	2ts_	_chê	_ch3
_chô	_ch5	_chơ	_ch4	_chư	_ch6	_đê	_dd3
_đô	_dd5	_đơ	_dd4	_đư	_dd6	eng_	eg_
ép_	eps_	ẹp_	epj_	ét_	ets_	ẹt_	etj_
ếch_	31s_	ệch_	31j_	ênh_	eeh_	ếp_	3ps_
ệp_	3pj_	ết_	3ts_	ệt_	3tj_	_gi	_gii
_gia	_da	_ghe	_ge	_ghê	_gee	_ghi	_gi
_hy	_hi	ích_	i1s_	ịch_	i1j_	iếc_	i2s_
iệc_	i2j_	iêng_	ieg_	iết_	i3s_	iệt_	i3j_
iêu_	ieu_	íp_	ips_	ịp_	ipj_	ít_	its_
ịt_	itj_	_khê	_kh3	_khô	_kh5	_khơ	_kh4
_khư	_kh6	_ky	_ki	_ly	_li	_my	_mi
_nghe	_nge	_nghê	_ng3	_nghi	_ngi	_ngô	_ng5
_ngơ	_ng4	_ngư	_ng6	_nhê	_nh3	_nhô	_nh5
_nhơ	_nh4	_như	_nh6	oát_	o1s_	oạt_	o1t_
oang_	oag_	oanh_	oah_	óc_	ocs_	ọc_	ocj_
ong_	og_	oong_	o2_	óp_	ops_	ọp_	opj_
ót_	ots_	ọt_	otj_	ốc_	5cs_	ộc_	5cj_
ông_	oog_	ốp_	5ps_	ộp_	5pj_	ốt_	5ts_
ột_	5tj_	ớp_	4ps_	ợp_	4pj_	ớt_	4ts_
ợt_	4tj_	_phê	_ph3	_phô	_ph5	_phơ	_ph4
_phư	_ph6	_quê	_qu3	_quơ	_qu4	_quy	_qui
_sy	_si	_ty	_ti	_thê	_th3	_thô	_th5
_thơ	_th4	_thư	_th6	_trê	_tr3	_trô	_tr5
_trơ	_tr4	_trư	_tr6	uân_	uan_	uất_	u1s_
uật_	u1j_	úc_	ucs_	ục_	ucj_	ung_	ug_
uốc_	u2s_	uộc_	u2j_	uôi_	uoi_	uốt_	u3s_
uột_	u3j_	uông_	uog_	úp_	ups_	ụp_	upj_
út_	uts_	ụt_	utj_	uyên_	u4_	uyết_	u5s_

uyệt_	u5j_	uýt_	u6s_	uyt_	u6j_	ức_	6cs_
ực_	6cj_	ưng_	uwg_	ước_	w1s_	ược_	w1j_
ươi_	woi_	ươn_	won_	ương_	wog_	ướt_	w2s_
ượt_	w2j_	ứt_	6ts_	ựt_	6tj_	_vy	_vi
_xy	_xi	yêm_	iem_	yên_	ien_	iết_	i3s_
iệt_	i3j_	yêu_	ieu_				

TÀI LIỆU THAM KHẢO

[1]. Dư Thanh Bình
Dò tìm tần số cơ bản trong xử lý tiếng nói
Đại hoc Bách Khoa Hà Nội, 2001

[2]. Lawrence R.Rabiner, Biing–Huang Juang
Foundamentals of speech recognition,
Prentice Hall, 1993

[3] Lawrence R.Rabiner, Ronald W.Schafer
Digital processing of speech signals
Prentice Hall, 1978

[4] Lê Xuân Hùng
Đồ án tốt nghiệp đại học ngành Công Nghệ Thông Tin
Đại học Bách khoa Hà Nội, 2001

[5] Nguyễn Quốc Trung
Xử lý tín hiệu và lọc số
Nhà xuất bản khoa học và kỹ thuật, 2001

[6] Quách Tuấn Ngọc
Xử lý tín hiệu số
Nhà xuất bản giáo dục, 1997

[7] Thierry Dutoit
An Introduction to Text-to-Speech Synthesis
1997

[8] Thierry Dutoit
High quality Text-to-Speech synthesis of the France language
1993

[9] Trịnh Văn Loan
Các bài giảng về xử lý tiếng nói
Đại hoc Bách Khoa Hà Nội, 1998