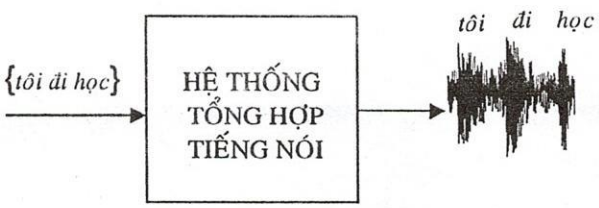
TỔNG HỢP MỘT SỐ ÂM TIẾT TIẾNG VIỆT DỰA VÀO CÁC CHU KỲ CAO ĐỘ

Lê Tiến Thường, Lê Ngọc Phú
Trường Đại học Bách khoa – Đại học Quốc gia Tp. HCM
(Bài nhận ngày 31 tháng 8 năm 2004, hoàn chỉnh sửa chữa ngày 24 tháng 01 năm 2005)

TÓM TẮT: Bài báo giới thiệu một phương pháp thực hiện việc tổng hợp một số âm tiết tiếng Việt dựa trên các chu kỳ cao độ và quy luật thay đổi của nó. Phương pháp này đã được tác giả sử dụng để xây dựng giải thuật và viết chương trình hoàn chỉnh thực hiện việc tổng hợp một số âm tiết cơ bản trong hệ thống các âm tiết tiếng Việt. Kết quả thu được rất khả quan, những âm tiết tổng hợp được có chất lượng rất rõ ràng, từ đó cho thấy đây là một phương pháp có thể áp dụng để thực hiện tổng hợp tiếng nói tiếng Việt.

I. Giới thiệu

Hệ thống tổng hợp tiếng nói nói chung có nhiệm vụ nhận các chữ viết ở ngõ vào, thực hiện tổng hợp (*tạo ra*) và phát đến ngõ ra (*loa*) chuỗi tiếng nói tương ứng một cách rõ ràng, sao cho người nghe có thể hiểu rõ được nội dung mà hệ thống tổng hợp tiếng nói muốn truyền đạt. Do đó, việc tổng hợp tiếng nói còn được gọi là việc **chuyển chữ viết sang tiếng nói** – TTS (*Text–To–Speech*). Như vậy, hệ thống tổng hợp tiếng nói tiếng Việt sẽ nhận dữ liệu ngõ vào (*bàn phím*) là các chữ viết tiếng Việt, thực hiện tổng hợp và phát đến ngõ ra (*loa*) chuỗi tiếng nói tiếng Việt tương ứng. Điều này thể hiện rõ trong *hình 1*.



Hình 1: Nhiệm vụ tổng hợp tiếng nói

Với nhiệm vụ như vậy, khả năng ứng dụng mà hệ thống tổng hợp tiếng nói sẽ mang lại là vô cùng to lớn, đặc biệt là trong lĩnh vực Viễn thông và Công nghệ thông tin. Vì vậy, việc tìm ra một phương pháp hiệu quả nhất để thực hiện việc tổng hợp tiếng nói là một hướng nghiên cứu đã và đang thu hút rất nhiều các nhà khoa học trên thế giới [4].

Hiện tại, đã có rất nhiều đề tài nghiên cứu về phương pháp thực hiện tổng hợp tiếng nói của các nước trong khu vực và trên thế giới như Trung Quốc, Thái Lan, Indonexia, Nhật Bản, Anh, Pháp, Tây Ban Nha... đã được công bố. Trong đó, rất nhiều ngôn ngữ đã được tổng hợp tiếng nói thành công như tiếng Anh, tiếng Pháp, tiếng Tây Ban Nha... Chương trình tổng hợp tiếng nói của các ngôn ngữ trên đã được công bố với chất lượng có thể chấp nhận được. Một số chương trình đã xuất hiện trên thị trường thương mại. Gần đây việc nghiên cứu về đề tài này càng được thúc đẩy mạnh do sự phát triển của công nghệ máy tính và các chip xử lý số tín hiệu (DSP) [4].

Trong khi tiếng nói của các ngôn ngữ khác đang được đầu tư nghiên cứu rất mạnh do những ứng dụng rất phổ biến mà nó sẽ mang lại thì việc thực hiện tổng hợp tiếng nói đối với ngôn ngữ tiếng Việt vẫn còn là vấn đề khá mới mẻ. Điều này là do bản chất âm học của tiếng nói tiếng Việt có những đặc trưng rất khác biệt so với tiếng nói của các ngôn ngữ khác. Việc nghiên cứu thực hiện tổng hợp tiếng nói tiếng Việt đòi hỏi phải tìm ra những giải pháp riêng, thích hợp với các đặc trưng của tiếng nói tiếng Việt. Do vậy, yêu cầu phải phân tích kỹ lưỡng, chính xác các đặc trưng thì mới có thể thực hiện tổng hợp tiếng nói tiếng Việt một cách hiệu quả nhất [8].

II. Cơ sở thực hiện

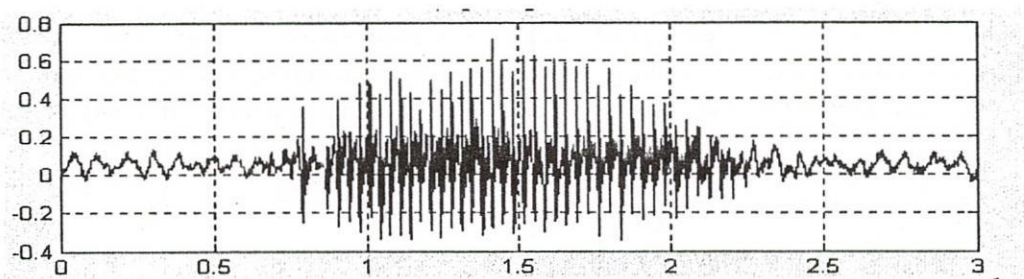
Việc tổng hợp tiếng nói nói chung có thể được thực hiện dựa trên nhiều cách khác nhau như là thu âm trước, tổng hợp formant, kết nối... Tuy nhiên hiện nay phương pháp đang phát triển mạnh nhất và được áp dụng nhiều nhất là phương pháp kết nối. Phương pháp này có đặc điểm là tốn dung lượng bộ nhớ lớn để lưu trữ dữ liệu các đơn vị cơ bản của tiếng nói (*vấn đề này hiện nay có thể chấp nhận được*) nhưng chất lượng tiếng nói tổng hợp được rất tốt (*hiên nay, đây là vấn đề được ưu tiên nhất*) [1]. Ý tưởng tổng quát của phương pháp này là tiếng nói hoàn chỉnh của một từ được kết nối từ các đơn vị ngữ âm khác nhau. Các đơn vị ngữ âm này có thể được chọn lựa tùy theo từng trường hợp và từng ngôn ngữ khác nhau, từ đó sẽ tạo ra các phương pháp riêng nhưng các phương pháp này có chung đặc điểm là tiếng nói được tổng hợp bằng cách kết nối các đơn vị ngữ âm lại với nhau [5]. Xét về mặt cấu trúc thì một âm tiết tiếng Việt được cấu tạo từ những thành phần nhỏ hơn như sau:

Bảng 1: Cấu trúc âm tiết tiếng Việt

Thanh điệu			
Âm đầu	Vần		
	Âm đệm	Âm chính	Âm cuối

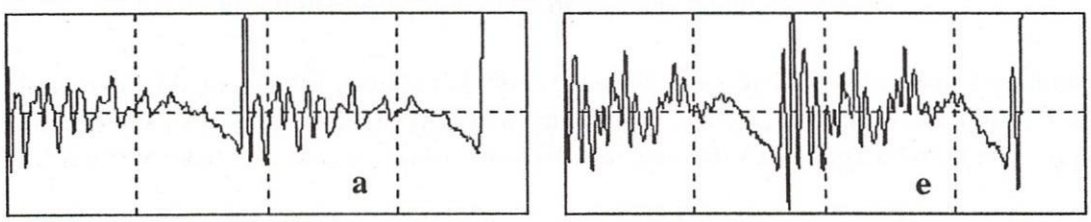
Chi tiết về đặc điểm, chức năng và cấu trúc từng thành phần trong âm tiết tiếng Việt được trình bày rõ trong tài liệu tham khảo [1]. Ở đây bài báo tập trung nghiên cứu bản chất ngữ âm học của các âm chính (*do các nguyên âm tiếng Việt đảm nhiệm*) và ảnh hưởng các thanh điệu lên các nguyên âm đó. Từ đó tìm ra một giải thuật thực hiện việc tổng hợp tiếng nói các nguyên âm tiếng Việt.

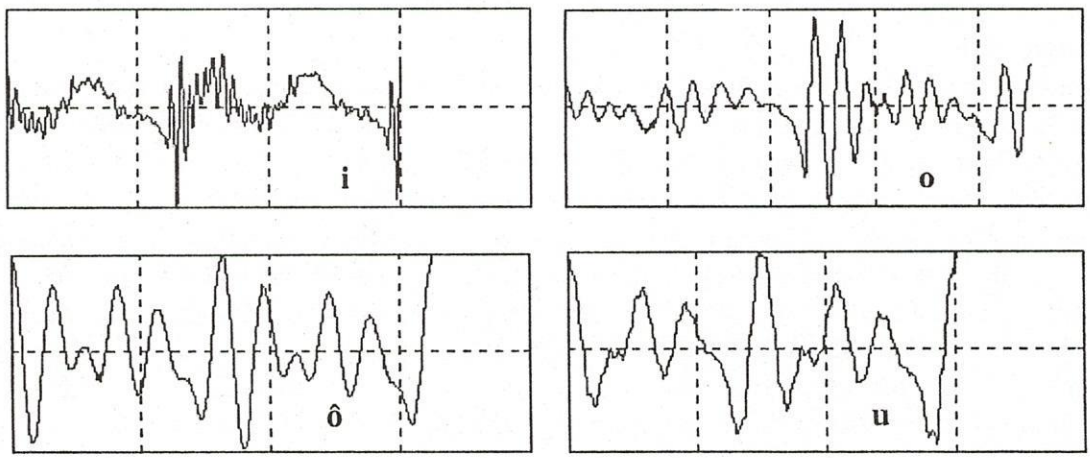
Bản chất của tiếng nói là do sự dao động của dây thanh đới (*xuất phát từ luồng không khí phát ra từ phổi*), từ đó tạo ra sự dao động (*nén hoặc giãn*) của luồng không khí ngay trước miệng của người nói. Kết quả là tạo ra sự chênh lệch cục bộ về áp suất không khí. Nếu dùng micro để nhận biết sự chênh lệch này thì tiếng nói sẽ được thể hiện dưới dạng các dao động về điện. Dao động này được thể hiện trong hình vẽ sau (*hình 2*)



Hình 2: Dạng sóng tiếng nói của từ 'A'

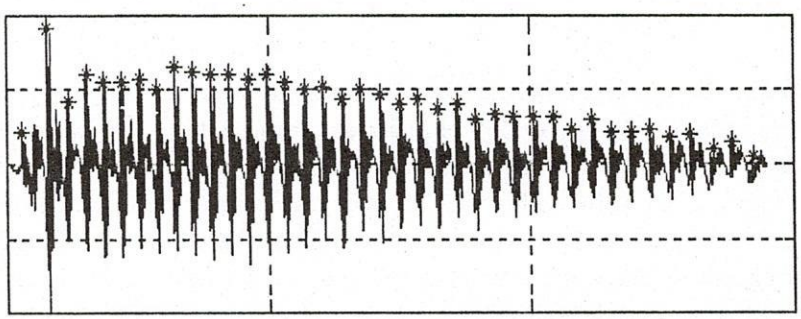
Về bản chất ngữ âm, các nguyên âm được tạo ra từ sự dao động có chu kỳ của dây thanh đới. Dựa trên cơ sở đó và bằng sự phân tích, khảo sát và thống kê, tác giả rút ra kết luận là đối với các nguyên âm tiếng Việt, chu kỳ cao độ được thể hiện rất rõ ràng, phân biệt giữa các nguyên âm khác nhau, dạng sóng 2 chu kỳ cao độ của một số nguyên âm thể hiện trong *hình 3*.





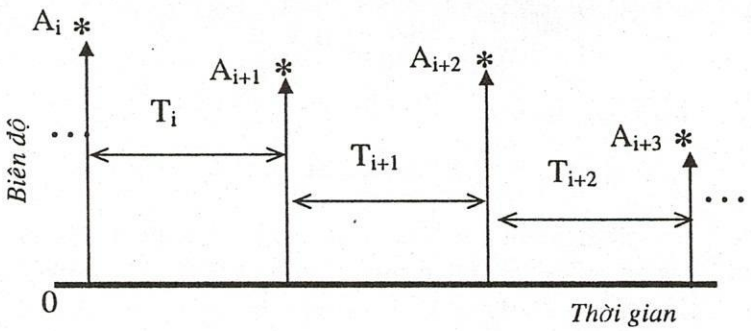
Hình 3: Dạng sóng hai chu kỳ cao độ của một số nguyên âm tiếng Việt

Bên cạnh đó, các nguyên âm tiếng Việt còn một đặc điểm quan trọng khác là sự thay đổi các đỉnh chu kỳ cao độ trong đoạn tiếng nói. Các đỉnh này có thể được xem như là cực đại địa phương trong miền thời gian trong khoảng thời gian tồn tại của một chu kỳ cao độ tiếng nói. Các đỉnh này được tác giả đánh dấu (nguyên âm O) lại trong *hình 4*:



Hình 4: Các đỉnh chu kỳ cao độ từ '0'

Cơ sở thực hiện và giải thuật chi tiết để xác định các đỉnh chu kỳ cao độ tiếng nói sẽ được trình bày trong phần sau. Bằng sự phân tích, tác giả nhận thấy các đỉnh chu kỳ cao độ có hai thông số cần phân tích. Đó là, biên độ của các đỉnh và khoảng cách liên tiếp giữa các đỉnh. Điều này được thể hiện trên (*hình 5*).



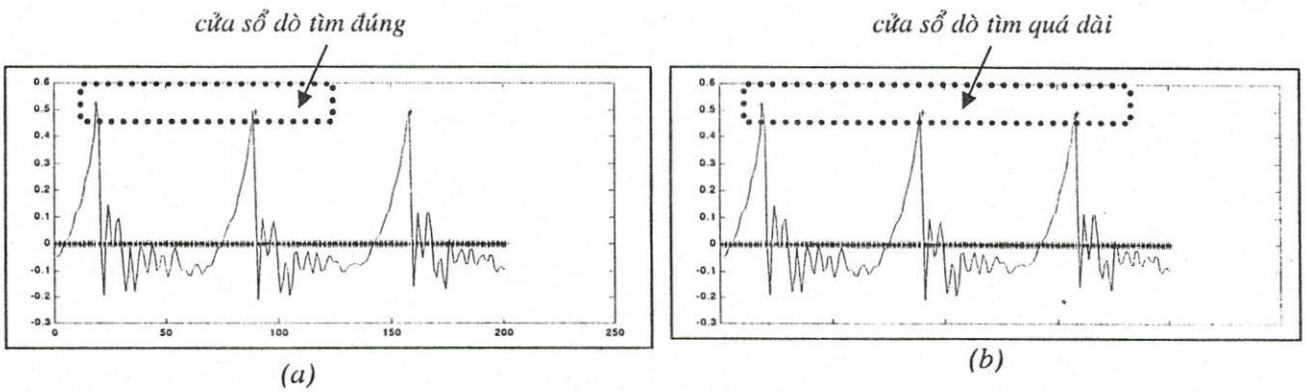
Hình 5: Bốn đỉnh chu kỳ cao độ liên tiếp trong từ 'O' (*phóng lớn*)

Trên *hình 5*, một đoạn liên tiếp bốn chu kỳ cao độ được phóng lớn. Trong đó, A_i, A_{i+1}, A_{i+2}, A_{i+3} là biên độ các đỉnh chu kỳ cao độ, T_i, T_{i+1}, T_{i+2}, là khoảng cách giữa các đỉnh chu kỳ cao độ. Việc tổng hợp tiếng nói các nguyên âm tiếng Việt được thực hiện dựa trên cơ sở các đặc trưng vừa đề cập ở trên.

III. Trích các đặc trưng

Việc tổng hợp tiếng nói được dựa trên các đặc trưng đã được phân tích, trích ra và lưu trữ sẵn trong cơ sở dữ liệu. Như đã đề cập ở trên, trong phương pháp tổng hợp này, hai đặc trưng được sử dụng để tổng hợp tiếng nói là đỉnh các chu kỳ cao độ và hai chu kỳ cao độ kế tiếp nhau trong tiếng nói.

Đối với các nguyên âm tiếng Việt, các đỉnh chu kỳ cao độ là một đặc trưng rất quan trọng. Để tổng hợp được các nguyên âm này thì việc đầu tiên là phải thực hiện xác định chính xác vị trí các đỉnh chu kỳ cao độ này. Việc đánh dấu này được thực hiện dựa trên cơ sở sau: xác định đỉnh chu kỳ cao độ đầu tiên. Đỉnh chu kỳ cao độ là điểm có biên độ cực đại trong phạm vi chu kỳ cao độ đó. Do đó, trong phạm vi tồn tại của tín hiệu tiếng nói, điểm có biên độ lớn nhất là đỉnh của một chu kỳ cao độ, đây được xem như là đỉnh của chu kỳ cao độ đầu tiên. Đỉnh chu kỳ cao độ đầu tiên này được lấy làm điểm mốc để thực hiện việc đánh dấu các chu kỳ cao độ kế tiếp về hai phía trong đoạn tiếng nói cho đến hết khoảng tín hiệu của tiếng nói. Việc đánh dấu này được minh họa trong *hình 6*:

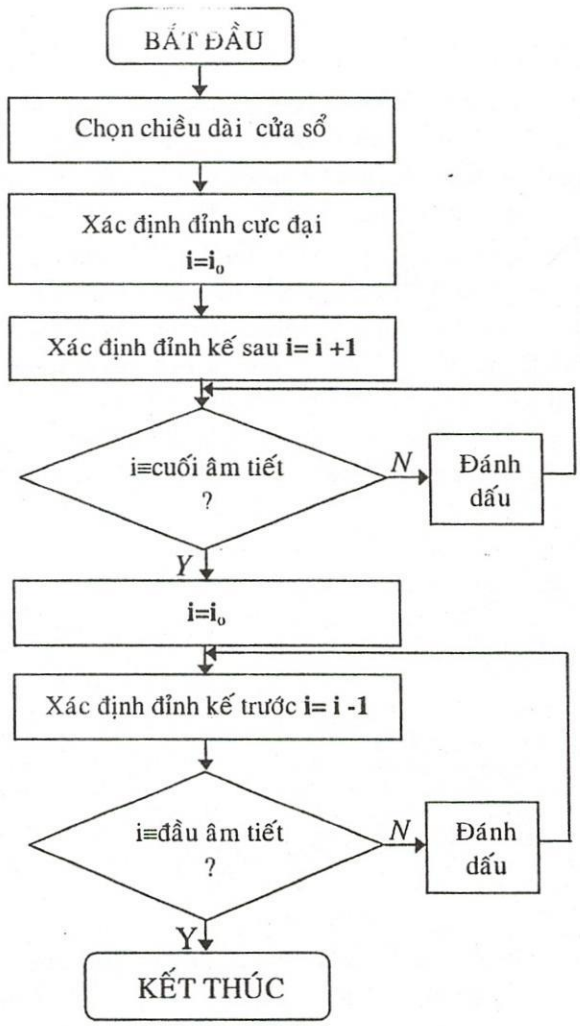


Hình 6: Xác định đỉnh chu kỳ cao độ dựa vào cửa sổ dò tìm
(a) Cửa sổ dò tìm có chiều dài hợp lý, (b) Cửa sổ dò tìm quá dài

Như vậy, khi đã xác định được đỉnh một chu kỳ cao độ, thì đỉnh các chu kỳ cao độ kế tiếp sẽ được xác định dựa trên điểm mốc vừa được xác định và một cửa sổ dò tìm. Cửa sổ này sẽ được dịch liên tiếp về hai phía trên cả đoạn tín hiệu tiếng nói để xác định hết tất cả các đỉnh chu kỳ cao độ. Đặc điểm của cửa sổ dò tìm là có chiều dài cố định. Việc xác định chiều dài cửa sổ phụ thuộc trực tiếp vào tần số cao độ của tiếng nói. Ở đây, với tần số lấy mẫu 44100 mẫu/giây, tác giả chọn cửa sổ dò tìm có chiều dài là L = 500 mẫu. Việc chọn chiều dài này dựa trên cơ sở phân tích tiếng nói của bản thân tác giả, rút ra chu kỳ cao độ (*nghịch đảo của tần số cao độ*) trong tiếng nói của mình. Chiều dài của cửa sổ dò tìm hợp lý dao động trong khoảng lớn hơn một chu kỳ cao độ và nhỏ hơn hai chu kỳ cao độ trong tín hiệu tiếng nói của người tổng hợp (*hình 6a*). Nếu chiều dài của cửa sổ dò tìm quá ngắn (*nhỏ hơn một chu kỳ cao độ*) thì tiếng nói tổng hợp được sẽ có thêm nhiều chu kỳ cao độ mới so với tiếng nói gốc, ngược lại nếu chọn của sổ dò tìm quá dài, (*lớn hơn hai chu kỳ cao độ*) – *hình 6b* thì tiếng nói tổng hợp được sẽ mất đi nhiều chu kỳ cao độ mới so với tiếng nói gốc. Cả hai trường hợp này đều làm cho tiếng nói tổng hợp có chất lượng rất xấu, có thể không nghe được. Do vậy, nếu tổng hợp tiếng nói của người khác, chiều dài này có thể sẽ thay đổi cho phù hợp với giọng nói của người đó dựa trên việc xác định chu kỳ cao độ của người tổng hợp tiếng nói.

Lưu đồ giải thuật dò tìm và đánh dấu các chu kỳ cao độ của tiếng nói được thể hiện trong *hình 7*.

Sau khi đã thực hiện việc đánh dấu đỉnh các chu kỳ cao độ, việc trích hai chu kỳ cao độ được thực hiện đơn giản bằng cách tách đoạn tiếng nói giữa hai chu kỳ cao độ kề nhau được xác định từ các đỉnh chu kỳ cao độ đã đánh dấu. Kết quả này đã được thể hiện trong *hình 2*.

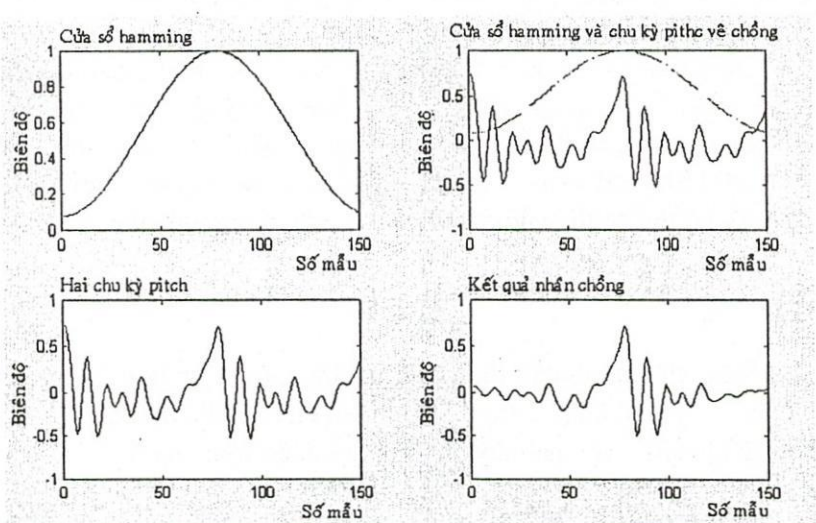


Hình 7: Lưu đồ giải thuật xác định đỉnh chu kỳ cao độ

IV. Tổng hợp tiếng nói

Việc tổng hợp tiếng nói dựa trên các đặc trưng đã được trích ra ở trên. Như đã nhận xét, dạng sóng của các nguyên âm có dạng gần tuần hoàn, chu kỳ cao độ gần như được lặp lại trong suốt chiều dài âm tiết. Do đó, tác giả đưa ra giải thuật tổng hợp các âm tiết tiếng Việt bằng cách cộng chồng lấp các đoạn tiếng nói cơ sở trong miền thời gian [6].

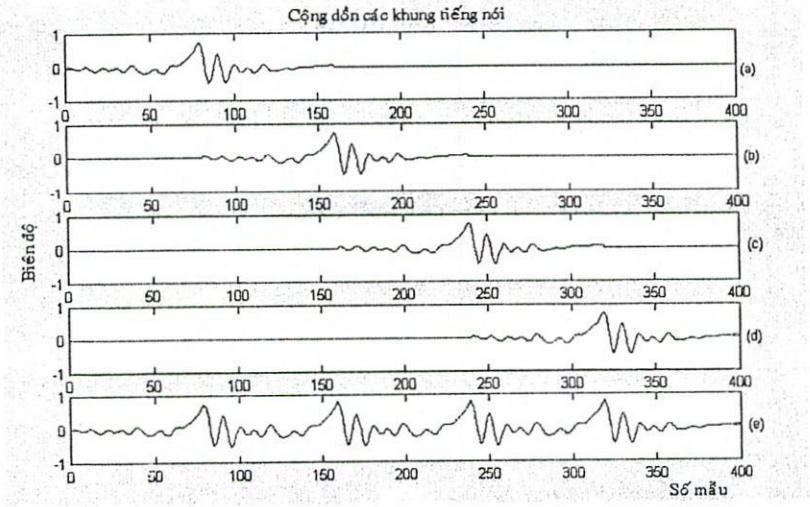
Đoạn tiếng nói cơ sở được tạo ra bằng cách nhân trực tiếp hai chu kỳ cao độ liên tiếp của tiếng nói đã được trích ra với cửa sổ hamming có cùng chiều dài được dịch khớp với đoạn 2 chu kỳ cao độ. Quá trình này được minh hoạ trong *hình 8*.



Hình 8: Nhân cửa sổ hamming với hai chu kỳ cao độ

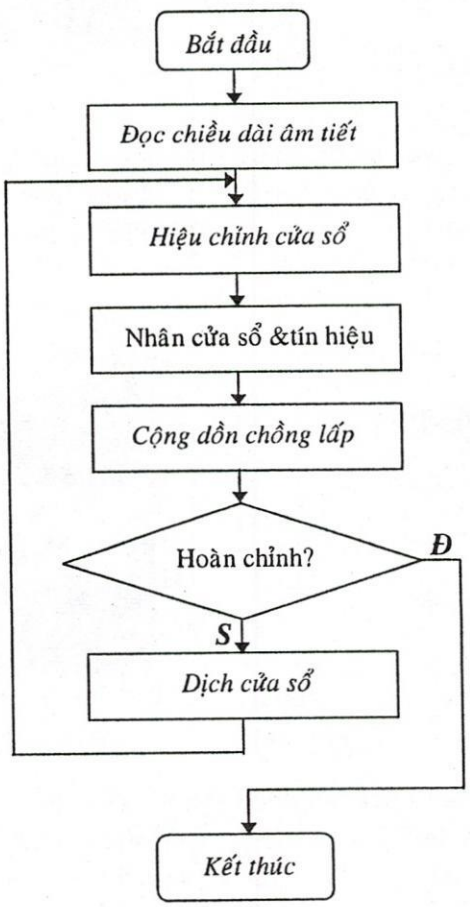
Dễ dàng nhận thấy bằng trực quan, việc nhân cửa sổ hamming có tác dụng: Giữ nguyên dạng sóng của đoạn giữa chu kỳ cao độ đồng thời làm suy hao dạng sóng hai biên của chu kỳ cao độ.

Đoạn tiếng nói cơ sở sau khi được tạo ra, sẽ được dịch liên tiếp tới vị trí đỉnh của các chu kỳ cao độ, đồng thời được nhân với hệ số là giá trị biên độ đỉnh của chu kỳ cao độ tương ứng (*đây là các đặc trưng đã được xác định và lưu trữ trước*). Quá trình này được minh họa trong *hình 9* như sau:



Hình 9: Dịch và cộng dồn các đoạn tiếng nói cơ sở.

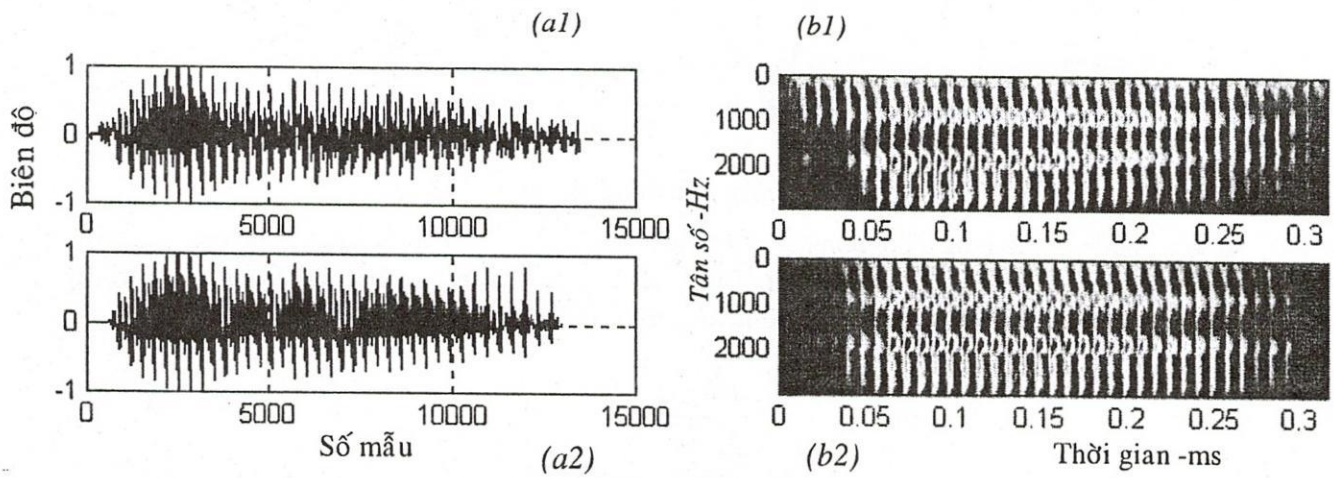
Ở đây, một vấn đề đáng quan tâm là số lần dịch và cộng dồn các đoạn tiếng nói cơ sở. Điều này được quyết định bởi chiều dài của tiếng nói. Thông số này được xác định bởi đặc trưng về vị trí các đỉnh chu kỳ cao độ. Quá trình dịch và cộng dồn được tóm tắt trong sơ đồ *hình 10* như sau:



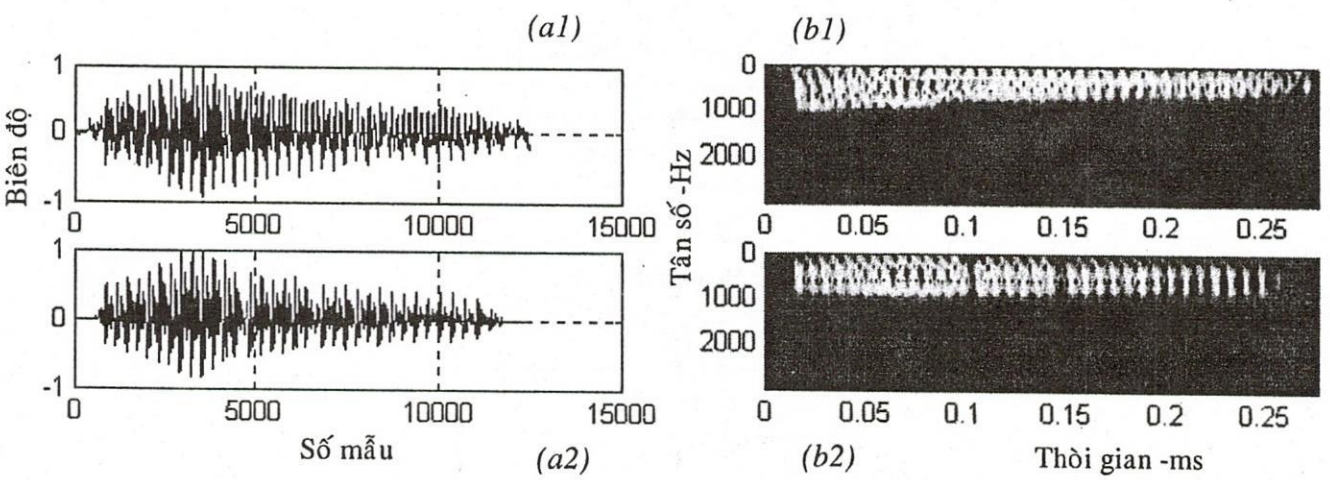
Hình 10: Lưu đồ giải thuật cộng dồn chồng lấp

V. Kết luận, hướng phát triển

Dựa trên thuật giải như trên, tác giả đã viết được một chương trình tổng hợp tiếng nói. Các âm tiết tổng hợp được là các nguyên âm đơn đơn của tiếng Việt như *a, o, ô, ơ, i, u*... các nguyên âm kết hợp với các thanh điệu (*á, à, ả, ó, ò...*), các âm tiết kép (*ao, áo, ôi, ơi*...) – là các âm tiết kết hợp từ hai nguyên âm đơn. Các âm tiết tổng hợp được có chất lượng tốt: *rõ ràng, tự nhiên, dễ phân biệt*. Việc đánh giá chất lượng tiếng nói tổng hợp được thực hiện dựa trên việc so sánh giữa tiếng nói tổng hợp được và tiếng nói gốc. Dựa trên việc quan sát dạng sóng tín hiệu tiếng nói trong miền thời gian và đồ thị phổ trong miền tần số của tín hiệu tiếng nói của các nguyên âm 'ô' *hình 11* và 'e' hình *12* của tín hiệu gốc (*thu trực tiếp*) và tín hiệu tổng hợp, có thể nhận xét là dạng sóng và phổ của tín hiệu gốc và tín hiệu tổng hợp được là tương đối giống nhau. Ngoài hai nguyên âm này, tác giả cũng đã phân tích hầu hết các nguyên âm khác trong hệ thống các nguyên âm tiếng Việt và đều cho kết quả tương tự như vậy. Tuy nhiên, cơ sở chính để đánh giá chất lượng của tiếng nói tổng hợp được là nghe và so sánh tiếng nói gốc và tiếng nói tổng hợp được. Trong quá trình thực hiện đề tài này, tác giả đã viết chương trình thực hiện việc tổng hợp các nguyên âm đơn và một số âm kép trong tiếng Việt. Các âm tiết tổng hợp được có chất lượng tốt, có thể nghe rõ ràng, tự nhiên, chấp nhận được.



Hình 11: Dạng sóng và phổ tín hiệu tiếng nói gốc của nguyên âm 'ô'
(a1) Dạng sóng tín hiệu gốc (a2) Dạng sóng tín hiệu tổng hợp
(b1) Phổ tín hiệu gốc (b2) Phổ tín hiệu tổng hợp



Hình 12: Dạng sóng và phổ tín hiệu tiếng nói gốc của nguyên âm 'e'
(a1) Dạng sóng tín hiệu gốc *(a2)* Dạng sóng tín hiệu tổng hợp
(b1) Phổ tín hiệu gốc *(b2)* Phổ tín hiệu tổng hợp

Các kết luận sau đây được rút ra trong quá trình phân tích, khảo sát và tổng hợp tiếng nói các âm tiết tiếng Việt. Dạng sóng tiếng nói của các nguyên âm tiếng Việt có dạng gần tuần hoàn. Hai đặc trưng quan trọng nhất có thể được sử dụng để tổng hợp tiếng nói là hai chu kỳ cao độ kế tiếp và các đỉnh chu kỳ cao độ. Đây là hai đặc trưng phân biệt các âm tiết và cũng là hai đặc trưng phân biệt giọng nói của những người khác nhau. Kết luận này được rút ra bởi vì tiếng nói tổng hợp được dựa trên chu kỳ cao độ và sự thay đổi của chúng có thể được nhận dạng giọng nói của từng người một cách chính xác.

Kết quả tổng hợp một số âm tiết tiếng Việt có chất lượng tốt, cho thấy đây là một hướng nghiên cứu rất có triển vọng. Mặc dù vậy, đây cũng chỉ là kết quả cơ bản, chương trình tổng hợp tiếng nói chỉ được thực hiện trên các nguyên âm. Để thực hiện tổng hợp tiếng nói thành công, đòi hỏi phải thực hiện thành công các phụ âm và thực hiện nghiên cứu ảnh hưởng của các từ trong câu, từ đó mới tổng hợp được hoàn chỉnh một văn bản.

VIETNAMESE SPEECH SYNTHESIS BASED ON PITCH PERIODS

Le Tien Thuong , Le Ngoc Phu
University of Technology – VNU-HCM

***ABSTRACT**: The purpose of this paper is to present a method applied to the Vietnamese speech synthesis based on pitch periods and it's changes. The method is used to build up the algorithm and to syntesize some Vietnamese basis syllables. The results of the method analysis have show that this is an efficient approach for Vietnamese speech synthesis.*

TÀI LIỆU THAM KHẢO

[1] Lê Ngọc Phú "*Tổng hợp tiếng nói tiếng Việt*", Luận văn Cao học, 2003.

[2] Tu Trong Do, Timio TAKARA, "*Precise tone generation for Vietnamese text-to-speech system*". 0-7803-7663-3/03/$17.00 2003IEEE.

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

[4] History of speech synthesis - http://www.ling.su.se/staff/kemplne.htm

[5] Kurt Edward Dusterhoff, "Synthesizing fundamental frequency using models automatically *trained from data*", Doctor of Philosophy, University of Edinburgh 2000

[6] Daniel Jurafsky, James H. Martin, "*Speech and Language Processing – An introduction to natural language processing, Computational linguistics and speech recognition*". ISBN 0-13-095069-6, Prentice Hall 2000.

[7] Kurt Edward Dusterhoff, "*Synthesizing fundamental frequency using models automatically trained from data*", Doctor of Philosophy, University of Edinburgh 2000.

[8] T.LeTien, "A Study based on the continuous wavelet transform for the Vietnamese Speech Processing", Proceedings of the 1997 Internation Conference on Neural Information Processing and Intelligent Information Systems, "ICONIP'97, University of Otago, Dunedin, NewZealand, pp 1072-1075, Nov.1997.