THIẾT BỊ TỔNG HỢP VĂN BẢN TIẾNG VIỆT SANG TIẾNG NÓI DỰA TRÊN MÔ HÌNH MARKOV ẨN

IMPLEMENTATION OF HIDDEN MARKOV MODEL (HMM) - BASED VIETNAMESE TEXT-TO-SPEECH DEVICE

Nguyễn Hồng Thắng, Bùi Trọng Tú, Huỳnh Hữu Thuận
Trường Đại học Khoa học Tự nhiên TP.HCM

TÓM TẮT

Trong bài báo này tác giả sẽ trình bày thiết kế của một thiết bị cầm tay cho phép chuyển đổi văn bản tiếng Việt sang tiếng nói. Thiết bị có thể được sử dụng để hỗ trợ những người khuyết tật mất khả năng nói có thể giao tiếp dễ dàng và tự nhiên hơn với những người xung quanh. Thuật toán tổng hợp tiếng nói được dùng dựa trên mô hình Markov ẩn (Hidden Markov Model - HMM)[1].

Từ khóa: *TTS, tiếng Việt, HMM*

ABSTRACT

In this paper, an implementation of Vietnamese Text-To-Speech (TTS) device is presented. The device helps to interact with people with speech disability easier. The speech synthesis algorithm implemented in the device is based on Hidden Markov Model (HMM).

Keywords: *TTS, Vietnamese synthesis, HMM*

I. GIỚI THIỆU

Trong cuộc sống hằng ngày, người khuyết tật khả năng nói có thể giao tiếp với những người xung quanh bằng ngôn ngữ cử chỉ, tuy nhiên việc giao tiếp này không tự nhiên và sẽ gây khó khăn cho những người bình thường không hiểu ngôn ngữ cử chỉ. Chính vì lí do này, một thiết bị tổng hợp tiếng nói từ văn bản tiếng Việt được trình bày trong bài báo này có thể giúp việc giao tiếp của người khuyết tật trở nên dễ dàng hơn. Sử dụng thiết bị này, người dùng sẽ nhập nội dung giao tiếp dưới dạng văn bản qua một bàn phím ảo hiển thị trên màn hình cảm ứng được tích hợp trên thiết bị, dữ liệu sau đó sẽ được xử lí bằng thuật toán tổng hợp để tạo âm thanh (tiếng nói) và được phát ra loa. Do tốc độ tổng hợp tiếng nói của thiết bị có độ trễ thấp nên việc giao tiếp thông qua máy khá tự nhiên.

Thuật toán tổng hợp tiếng nói sử dụng trong thiết bị được dựa trên dự án mã nguồn mở HTS Engine [3]. Nền tảng phần cứng của thiết bị là bo mạch phát triển Raspberry Pi có cấu hình được liệt kê trong Bảng 1 và được tích hợp thêm các giao tiếp ngoại vi như LCD cảm ứng và loa ngoài.

Hình 1: Thiết bị tổng hợp tiếng nói.

Bảng 1: Cấu hình của bo mạch phát triển Raspberry Pi

Vi xử lí	**SoC Broadcom BCM2835 với CPU ARM1176JZF-S xung nhịp 700 MHz**
Bộ nhớ RAM	256 MB
Giao tiếp	GPIO, UART, I^2C, SPI
Công suất	300 mA (1.5 W)
Kích thước	85.60 mm × 56 mm
Hệ điều hành	Arch Linux ARM, Debian GNU/Linux, Raspbian OS, …
Giá	$25

II. THUẬT TOÁN TỔNG HỢP TIẾNG NÓI

Trong lĩnh vực tổng hợp tiếng nói từ văn bản, đã có nhiều phương pháp được đề xuất và thực hiện như: phương pháp tổng hợp ghép nối, tổng hợp formant, tổng hợp dựa trên HMM, … Trong đó thuật toán tổng hợp theo phương pháp ghép nối đang được sử dụng phổ biến ở thời điểm hiện tại do phương pháp này có độ phức tạp không cao. Tuy nhiên, phương pháp tổng hợp dựa trên HMM đang phát triển mạnh do có ưu điểm là dễ dàng thay đổi giọng đọc và không cần cơ sở dữ liệu lớn như các phương pháp tổng hợp khác. Trong tổng hợp tiếng nói từ văn bản, dù sử dụng phương pháp tổng hợp nào, văn bản đầu vào đều phải được qua quá trình chuẩn hóa, tức là chuyển đổi các kí hiệu, số, từ viết tắt, tên riêng, tiếng nước ngoài, … thành dạng đầy đủ, chính xác trước khi đưa vào thuật toán tổng hợp [2].

1. Phương pháp tổng hợp ghép nối

Theo phương pháp này, một đoạn tiếng nói sẽ được tạo ra bằng cách ghép nối các đơn vị âm thanh nhỏ hơn đã được thu âm trước tương ứng với văn bản đầu vào. Sau khi chuẩn hóa, văn bản sẽ được tách ra thành các cụm từ, rồi tiến hành chọn các cụm từ đó trong cơ sở dữ liệu tập tin âm thanh để ghép lại với nhau [5]. Các đơn vị âm thanh có thể là một câu, một cụm từ, một từ. Phương pháp này cho ra âm thanh có chất lượng tương đối tốt nhưng đòi hỏi không gian lưu trữ lớn để chứa được các phân đoạn âm thanh. Với mục đích tạo ra âm thanh có chất lượng tốt, tự nhiên thì cần phải chú ý đến một số vấn đề: cơ sở dữ liệu âm thanh phải được xây dựng, thiết kế một cách cẩn thận để có thể phủ tất cả các ngữ âm, ngôn điệu và những biến thể khác nhau của mỗi một đơn vị âm thanh [5]. Các phân đoạn âm thanh càng dài thì khi ghép nối lại với nhau sẽ tạo ra chất lượng âm thanh tự nhiên hơn, giảm tối thiểu tính không liên tục giữa các đơn vị được lựa chọn, ít tốn chi phí nối ghép.

Trong thực tế triển khai, ưu điểm của phương pháp tổng hợp ghép nối là:

- Âm thanh tạo ra có tính chất là giọng người thật,có tính tự nhiên cao. Thực tế là do việc sử dụng các đơn vị âm thanh đã được thu âm sẵn.
- Việc tính toán để lựa chọn các đơn vị âm thanh, cũng như quá trình ghép nối các đơn vị âm thanh này lại có chi phí thấp và thời gian thực hiện nhanh.

Tuy nhiên phương pháp này lại có nhưng khuyết điểm như:

- Đòi hỏi không gian lưu trữ lớn để lưu dữ liệu đã được thu âm. Thường một cơ sở dữ liệu của phương pháp ghép nối có dung lượng từ vài gigabyte trở lên.
- Quá trình chuẩn bị dữ liệu: quá trình thu

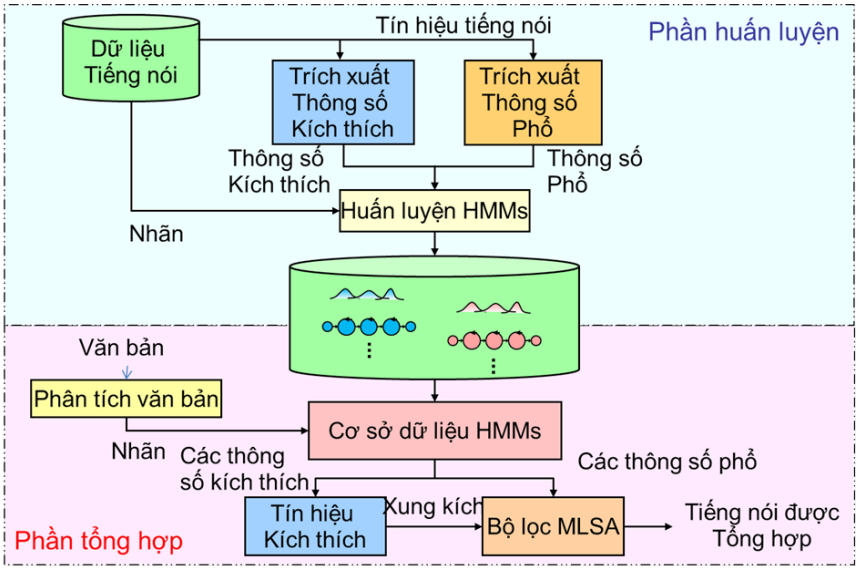
âm, phân đoạn dữ liệu và tổ chức dữ liệu tốn nhiều chi phí.

- Khi thay đổi giọng đọc, cần phải tiến hành thu âm lại cũng như phân đoạn và tổ chức dữ liệu từ đầu.
- Việc lựa chọn các đơn vị ghép nối sẽ ảnh hưởng đến độ trơn, mượt của âm thanh.
- Do đó, phương pháp tổng hợp ghép nối thường khó triển khai trên các thiết bị nhúng, do đặc điểm của các thiết bị này là không gian lưu trữ tương đối nhỏ, phương pháp này thích hợp hơn khi triển khai theo mô hình dịch vụ client-server qua mạng Internet.

2. Phương pháp tổng hợp hmm

Do bởi những hạn chế của phương pháp ghép nối, có hai phương pháp khác được đề xuất là: tổng hợp tiếng nói bằng phương pháp formant và tổng hợp dựa trên mô hình Markov ẩn (HMM). Phương pháp formant, tức tổng hợp cộng hưởng tần số, không cần sử dụng cơ sở dữ liệu thu sẵn khi chạy, mà tổng hợp dựa trên một mô hình âm thanh [6]. Tuy nhiên, tiếng nói được tổng hợp theo phương pháp này có độ tự nhiên thấp. Phương pháp HTS có nhiều ưu điểm do có thể thay đổi đặc tính giọng nói bằng việc thay đổi thông số của mô hình Markov ẩn(HMM) mà không cần một cơ sở dữ liệu quá lớn như các phương pháp khác [2]. Và đây cũng là giải pháp được chọn cho thiết kế được trình bày trong bài báo.

Hệ thống HTS tổng quát gồm hai phần như được thể hiện ở Hình 2: phần huấn luyện và phần tổng hợp.



Hình 2: Mô hình hệ thống tổng hợp tiếng nói dựa trên HMM.

Trong phần huấn luyện thì dữ liệu tiếng nói và các đoạn văn bản của các dữ liệu tiếng nói đó được dùng để trích ra các tham số phổ và tham số kích thích. Các tham số này sẽ được mô hình hóa dùng HMM phụ thuộc ngữ cảnh.

Ở phần tổng hợp, từ một chuỗi label phụ thuộc ngữ cảnh thì một chuỗi HMM được hình thành bằng cách ghép nối các mô hình HMM tương ứng với các label đó. Sau đó thông qua các thuật toán tạo tham số, các thông số kích thích và các thông số phổ sẽ được tính ra từ chuỗi HMM đó, và các thông số này tiếp theo sẽ được đưa vào bộ lọc tổng hợp (MLSA hoặc MGLSA) để tổng hợp ra tiếng nói [2].

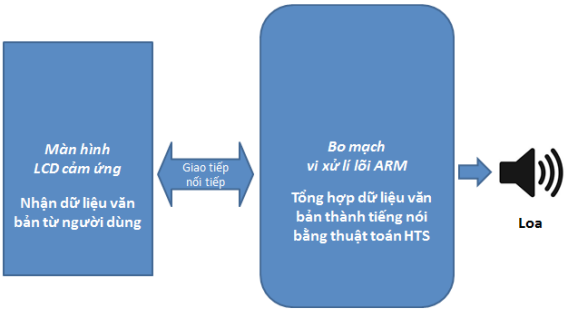
III. NGUYÊN LÍ HOẠT ĐỘNG CỦA THIẾT BỊ

Phần cứng của thiết bị tổng hợp văn bản

tiếng Việt thành tiếng nói gồm ba thành phần chính:

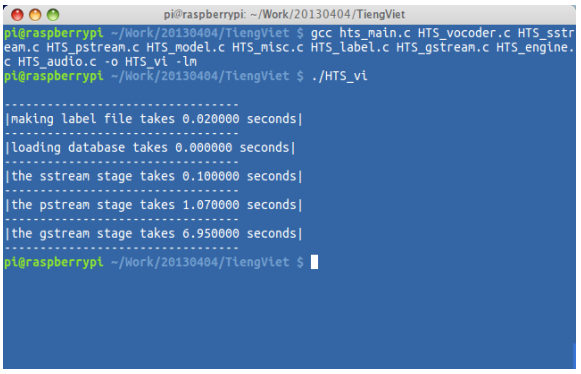
- Bo mạch Raspberry Pi với VXL chính là SoC BCM2835xung nhịp 700MHz có tích hợp khối phần cứng xử lí dấu chấm động (Vector Floating Point Unit) thích hợp để thực hiện thuật toán tổng hợp tiếng nói HTS.
- Màn hình LCD hỗ trợ cảm ứng.
- Loa tích hợp

Văn bản được nhập vào từ người dùng qua màn hình cảm ứng sẽ được đưa vào bo mạch Raspberry Pi thông qua giao tiếp nối tiếp RS232. Tại đây, văn bản sẽ được phân tích thành các chữ label -một định dạng chứa các thông tin đặc trưng của văn bản. Label và các thông số HMM đã được huấn luyện trước sẽ là đầu vào của quá trình tạo ra mã PCM (Pulse Code Modulation) tương ứng với văn bản đầu vào, bao gồm các bước như tạo ra các thông số excitation, lọc tổng hợp. Các dữ liệu PCM sẽ được đưa ra loa để phát ra tiếng nói. Hình 3 cho thấy lưu đồ các khối chức năng của thiết bị.



Hình 3: Sơ đồ khối của thiết bị tổng hợp tiếng nói.

Hình 4 cho thấy thời gian chạy thuật toán tổng hợp tiếng nói HMM trên bo mạch phát triển Raspberry Pi. Văn bản đầu vào là một câu văn tiếng Việt với nội dung: "Một hai ba bốn năm sáu bảy tám chín mười", thời gian phát của đoạn văn bản này (tức độ dài của tập tin âm thanh) sau khi được tổng hợp là khoảng 7 giây. Thực nghiệm cho thấy thời gian tổng hợp tổng cộng là khoảng 8 giây và thời gian xử lí chủ yếu là ở phần gstream.



Hình 4: Kết quả chạy thuật toán TTS tiếng Việt trên bo mạch Raspberry Pi.

Từ kết quả thực nghiệm trên, có thể thấy được thời gian trung bình để tổng hợp một từ tiếng Việt trên thiết bị là khoảng hơn 1 giây. Đối với văn bản đầu vào càng dài thì thời gian tổng hợp càng lâu. Để khắc phục điều này, phần mềm trên thiết bị sẽ tạo ra vùng đệm xử lí cho từng câu, trong khi hệ thống tiến hành tổng hợp câu trước thì người dùng có thể nhập câu tiếp để giảm thời gian chờ từ lúc người dùng nhập xong đoạn văn bản đến khi tiếng nói ứng với đoạn văn bản đó được phát ra loa, nhờ đó tăng trải nghiệm người dùng của thiết bị.

IV. TỔNG KẾT

Giải pháp sử dụng bo mạch Raspberry Pi là trung tâm xử lí, chạy thuật toán tổng hợp tiếng nói. Bo mạch Raspberry Pi thuộc dự án mã nguồn mở dành cho giáo dục nên giá thành rất thấp (khoảng 550 nghìn đồng) tuy nhiên vẫn đáp ứng tốt yêu cầu phần cứng cho thuật toán TTS. Điều này tạo ra một giải pháp hỗ trợ người không nói được với giá thành thấp, chỉ khoảng 1 – 1,5 triệu đồng. Hiện tại so với mặt bằng giá của các thiết bị cùng loại bán ra trên thị trường giải pháp được đề xuất có giá thành rẻ hơn trong khi chất lượng tiếng nói sau khi được tổng hợp vẫn được đảm bảo. Đồng thời, các sản phẩm đang được thương mại chưa sử dụng thuật toán tổng hợp tiếng nói dựa trên mô hình Markov ẩn.

TÀI LIỆU THAM KHẢO

[1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura,"Speech parameter generation algorithms for HMM-based speechsynthesis," Proc. of ICASSP 2000, vol.3, pp.1315–1318, June 2000

[2] Takayoshi Yoshimura, "Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based Text-to-Speech Systems", Doctoral Dissertation,

Department of Electrical and Computer Engineering, Nagoya Institute of Technology, January 2002

[3] http://hts.sp.nitech.ac.jp/

[4] http://en.wikipedia.org/wiki/Raspberry_Pi

[5] Vũ Hải Quân, Cao Xuân Nam, "Tổng hợp tiếng nói tiếng Việt, theo phương pháp ghép nối cụm từ", chuyên san "Các công trình nghiên cứu, phát triển và ứng dụng CNTT&TT",tập V-1, số 1, tháng 04/2009.

[6] Lê Hồng Minh, "Tổng hợp formant âm tiết tiếng Việt", Tạp chí Bưu chính Viễn thông, Số 179, 2002, tr. 41-44