



(12) **BẢN MÔ TẢ GIẢI PHÁP HỮU ÍCH THUỘC BẢNG ĐỘC QUYỀN
GIẢI PHÁP HỮU ÍCH**

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)
CỤC SỞ HỮU TRÍ TUỆ**

(11)



2-0002551

(51) G10L 13/00; G06F 3/16
2020.01

(13) **Y**

(21) 2-2014-00077 (22) 02/04/2014
(45) 25/01/2021 394 (43) 27/10/2014 319A
(73) Đại học Quốc gia thành phố Hồ Chí Minh (VN)
Phường Linh Trung, quận Thủ Đức, thành phố Hồ Chí Minh
(72) Huỳnh Hữu Thuận (VN); Vũ Hải Quân (VN); Phạm Minh Nhựt (VN).

(54) **THIẾT BỊ TÍCH HỢP CHUYỂN VĂN BẢN TIẾNG VIỆT SANG TIẾNG NÓI DỰA
TRÊN XỬ LÝ TÍN HIỆU SỐ VÀ MÔ HÌNH MARKOV ẨN**

(57) Giải pháp hữu ích đề cập đến thiết bị tích hợp chuyển văn bản tiếng Việt sang tiếng nói dựa trên thuật toán xử lý tín hiệu số và mô hình Markov ẩn bao gồm:

i) màn hình tinh thể lỏng (LCD-Liquid Crystal Display) cảm ứng hỗ trợ bàn phím ảo tích hợp mạch điều khiển có chức năng hiển thị nội dung nhập trên màn hình và lưu trữ nội dung nhập dưới dạng mã hexa vào một vùng nhớ tạm, và sau đó chuyển dữ liệu nhập này đến bo mạch Raspberry Pi qua đường giao tiếp;

ii) đường giao tiếp giữa màn hình tinh thể lỏng và bo mạch Raspberry Pi là các đường giao tiếp nối tiếp thông dụng như UART (Universal asynchronous receiver transmitter) hay SPI (Serial Peripheral Interface) giúp truyền dữ liệu văn bản đã được nhập từ màn hình tinh thể lỏng sang cho vi xử lý lưu trữ và dùng làm đầu vào để tổng hợp tiếng nói từ văn bản tiếng Việt (TTS-Text to Speech);

iii) bo mạch Raspberry Pi với vi xử lý là một DSP (Digital Signal Processor) lõi ARM11 xung nhịp 700MHz có tích hợp khối phần cứng xử lý dấu chấm động (Vector Floating Point Unit) để tổng hợp tiếng nói từ văn bản tiếng Việt thực hiện tổng hợp tiếng nói từ văn bản (TTS-Text to Speech) dựa trên mô hình Markov ẩn (HMM - Hidden Markov Model); và

iv) loa tích hợp trên bo mạch dùng để phát dữ liệu âm thanh sau khi đã tổng hợp xong.

Lĩnh vực kỹ thuật được đề cập

Giải pháp hữu ích thuộc lĩnh vực điện tử viễn thông, công nghệ thông tin, cụ thể đề cập đến thiết bị tích hợp chuyển văn bản tiếng Việt sang tiếng nói dựa trên xử lý tín hiệu số và mô hình Markov ẩn (HMM- Hidden Markov Model).

Tình trạng kỹ thuật của giải pháp hữu ích

Trong nhiều quốc gia đang phát triển, tỷ lệ người biết chữ thấp hơn nhiều so với tiêu chuẩn ở các quốc gia đã phát triển. Điều này là trở ngại rất lớn cho nhiều hình thức học tập truyền thống vì thông thường nội dung giáo dục được truyền đạt qua văn bản không thích hợp cho một đối tượng mù chữ hoặc những người có thị giác suy giảm. Để vượt qua những khó khăn này, một hệ thống học tập dựa trên giọng nói tương tác và điện thoại di động đã được thiết kế, thực hiện và phát triển.

Tại Việt Nam nói riêng và trên thế giới nói chung, việc giao tiếp giữa những người khuyết tật với người bình thường vẫn còn khó khăn. Những người khuyết tật (không có khả năng nói) thường giao tiếp với người bình thường thông qua ngôn ngữ cử chỉ, tuy nhiên không phải lúc nào người đối diện cũng hiểu được ngôn ngữ này. Xuất phát từ lý do này, nhằm giúp việc giao tiếp giữa người khuyết tật và người bình thường trở nên dễ dàng hơn, một số sản phẩm tổng hợp tiếng nói từ văn bản (TTS- Text to Speech) đã được ra đời. Thiết bị cho phép người dùng nhập nội dung dưới dạng văn bản và chuyển văn bản này thành tiếng nói.

Hiện nay, các thiết bị tổng hợp tiếng nói đang được thương mại trên thị trường sử dụng phương pháp tổng hợp ghép nối, ghi lại tiếng của tất cả các từ trong tiếng Việt, sau đó phân tích văn bản đầu vào và phát lại từ tương ứng trong cơ sở dữ liệu. Do đó các thiết bị này tốn nhiều không gian lưu trữ cho các tập tin âm thanh, làm tăng giá thành sản phẩm. Bên cạnh đó, tiếng nói đầu ra được tổng hợp ở các thiết bị này không có tính tự nhiên.

Bản chất kỹ thuật của giải pháp hữu ích

Mục tiêu của giải pháp hữu ích này đề xuất thiết bị tích hợp chuyển văn bản tiếng Việt sang tiếng nói dựa trên xử lý tín hiệu số và mô hình Markov ẩn (HMM- Hidden Markov Model) với giá rẻ hỗ trợ người khuyết tật về khả năng nói có thể giao tiếp với những người xung, cho chất lượng tiếng nói đầu ra tự nhiên với kích thước cơ sở dữ liệu bé. Thiết bị này bao gồm:

i) màn hình tinh thể lỏng (LCD-Liquid Crystal Display) cảm ứng hỗ trợ bàn phím ảo tích hợp mạch điều khiển có chức năng hiển thị nội dung nhập trên màn hình và lưu trữ nội dung nhập dưới dạng mã hexa vào một vùng nhớ tạm, và sau đó chuyển dữ liệu nhập này đến bo mạch Raspberry Pi qua đường giao tiếp;

ii) đường giao tiếp giữa màn hình tinh thể lỏng và bo mạch Raspberry Pi là các đường giao tiếp nối tiếp thông dụng như UART (Universal asynchronous receiver transmitter) hay SPI (Serial Peripheral Interface) giúp truyền dữ liệu văn bản đã được nhập từ màn hình tinh thể lỏng sang cho vi xử lý lưu trữ và dùng làm đầu vào để tổng hợp tiếng nói từ văn bản tiếng Việt (TTS-Text to Speech);

iii) bo mạch Raspberry Pi với vi xử lý là một DSP (Digital Signal Processor) lõi ARM11 xung nhịp 700MHz có tích hợp khối phần cứng xử lý dấu chấm động (Vector Floating Point Unit) để tổng hợp tiếng nói từ văn bản tiếng Việt thực hiện tổng hợp tiếng nói từ văn bản (TTS-Text to Speech) dựa trên mô hình Markov ẩn (HMM - Hidden Markov Model); và

iv) loa tích hợp trên bo mạch dùng để phát dữ liệu âm thanh sau khi đã tổng hợp xong.

Mô tả vắn tắt các hình vẽ

Hình 1 thể hiện thiết bị tích hợp chuyển văn bản tiếng Việt sang tiếng nói.

Hình 2: Sơ đồ khối thể hiện nguyên lý hoạt động của giải pháp.

Mô tả chi tiết giải pháp hữu ích

Theo Hình 1, 2 thiết bị tích hợp chuyển văn bản tiếng Việt sang tiếng nói dựa trên thuật toán xử lý tín hiệu số và mô hình Markov ẩn bao gồm các bộ phận chính như sau:

Màn hình tinh thể lỏng (LCD-Liquid Crystal Display) 2 cảm ứng hỗ trợ bàn phím ảo tích hợp mạch điều khiển có chức năng hiển thị nội dung nhập trên màn hình và lưu trữ nội dung nhập dưới dạng mã hexa vào một vùng nhớ tạm, và sau đó chuyển dữ liệu nhập này đến bo mạch Raspberry Pi 1 qua đường giao tiếp 4. Màn hình LCD cảm ứng 3,5 inch, hiển thị giao diện bàn phím ảo hỗ trợ người dùng nhập văn bản.

Đường giao tiếp 3 giữa màn hình tinh thể lỏng và bo mạch Raspberry Pi 1 là các đường giao tiếp nối tiếp thông dụng như UART (Universal asynchronous receiver transmitter) hay SPI (Serial Peripheral Interface) giúp truyền dữ liệu văn bản đã được nhập từ màn hình tinh thể lỏng sang cho vi xử lý lưu trữ và dùng làm đầu vào để tổng hợp tiếng nói từ văn bản tiếng Việt (TTS-Text to Speech).

Bo mạch Raspberry Pi 1 với vi xử lý là một DSP (Digital Signal Processor) lõi ARM11 xung nhịp 700MHz có tích hợp khối phần cứng xử lý dấu chấm động (Vector Floating Point Unit) để tổng hợp tiếng nói từ văn bản tiếng Việt thực hiện tổng hợp tiếng nói từ văn bản (TTS-Text to Speech) dựa trên mô hình Markov ẩn (HMM - Hidden Markov Model).

Thuật toán tổng hợp tiếng nói từ văn bản (TTS-Text to Speech) dựa nhiều vào việc tính toán trên dấu chấm động tuy nhiên nhiều vi xử lý giá rẻ không được tích hợp bộ xử lý dấu chấm động (floating point unit) bằng phần cứng. Do đó, thuật toán tổng hợp tiếng nói từ văn bản (TTS-Text to Speech) trên thiết bị đã được tối ưu bằng cách tích hợp vào bo mạch 1 khối phần cứng xử lý dấu chấm động (Vector Floating Point Unit) thích hợp để chuyển tất cả các tính toán trên dấu chấm động thành các tính toán trên dấu chấm cố định (fixed point). Điều này giúp cho thời gian tổng hợp từ lúc văn bản đầu vào được nhập đến lúc tạo ra tiếng nói được rút ngắn.

Việc tối ưu dấu chấm động được thực hiện thông qua một lớp thuật toán trung gian. Mục tiêu của lớp trung gian là quy đổi thể hiện dấu chấm động sang biểu diễn bằng dấu chấm cố định. Cụ thể, việc thay thế kiểu dấu chấm động bằng dấu chấm cố định là tìm cách biểu diễn số thực bằng một số nguyên rồi sau đó mới thực hiện các phép tính toán. Công thức biến đổi một số thực thành số nguyên đại diện cho số thực dấu chấm cố định (định dạng Q20) như sau

$$y = x \times 2^{20}$$

với x: số thực cần biểu diễn

y: số nguyên dùng để biểu diễn số thực.

Thuật toán tổng hợp tiếng nói từ văn bản (TTS-Text to Speech) được dựa trên mô hình Markov ẩn (HMM) có các điểm mạnh so với các phương pháp tổng hợp theo kiểu ghép nối phổ biến hiện nay như: cơ sở dữ liệu nhỏ; không đòi hỏi kho dữ liệu ở thời điểm thực thi; dễ thay đổi giọng đọc và lĩnh vực; độ trơn, mượt của âm thanh tự nhiên.

Loa 5 tích hợp trên bo mạch Raspberry Pi 1 dùng để phát dữ liệu âm thanh sau khi đã tổng hợp xong. Ngoài ra, thiết bị còn có pin 3 để cung cấp năng lượng cho thiết bị hoạt động và cổng ra (Line out) 3.5 5' để xuất tín hiệu tiếng nói.

Thiết bị tổng hợp tiếng nói từ văn bản (TTS-Text to Speech) dựa trên mô hình Markov ẩn (Hidden Markov Model- HMM) tối ưu cho bo mạch vi xử lý giá rẻ. Văn bản được nhập vào từ người dùng qua màn hình cảm ứng 2 sẽ được đưa vào bo mạch Raspberry Pi 1 xử lý. Tại đây, văn bản sẽ được phân tích thành các tập tin gắn nhãn (label file- một định dạng chứa các thông tin đặc trưng của văn bản). Tập tin gắn nhãn và các thông số mô hình Markov ẩn (Hidden Markov Model- 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 (Pulse Code Modulation) sẽ được đưa ra loa để phát ra tiếng nói.

Lợi ích kinh tế - xã hội

Thiết bị được cấu thành dựa trên một bo mạch vi xử lý giá rẻ được kết nối với một màn hình cảm ứng. Trên màn hình cảm ứng sẽ hiển thị một bàn phím ảo cho phép người dùng nhập tiếng Việt, dữ liệu văn bản đầu vào sẽ được đưa đến vi xử lý trên bo mạch thực hiện các thuật toán chuyển văn bản thành tiếng nói. Dữ liệu đầu ra sau khi xử lý sẽ là âm thanh (tiếng nói) và được phát ra loa tích hợp sẵn trên bo mạch. Thiết bị sử dụng nguồn là pin lithium có thể sạc được. Thiết bị có kích thước nhỏ gọn (90mm x 60 mm x 21mm) cho phép người dùng có thể mang theo bên mình. Đây là một giải pháp giúp cho người khuyết tật không có khả năng nói có thể giao tiếp bình thường với mọi người xung quanh. Thiết bị khi được sản xuất sẽ có giá thành từ 1 – 1,5 triệu đồng, đây là mức giá chấp nhận được so với thu nhập trung bình ở Việt Nam.

YÊU CẦU BẢO HỘ

1. Thiết bị tích hợp chuyển văn bản tiếng Việt sang tiếng nói dựa trên thuật toán xử lý tín hiệu số và mô hình Markov ẩn bao gồm:

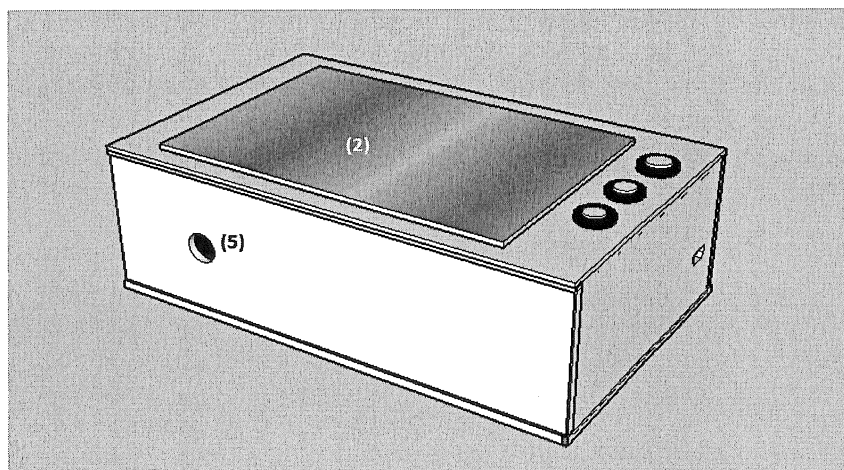
i) màn hình tinh thể lỏng (LCD-Liquid Crystal Display) cảm ứng hỗ trợ bàn phím ảo tích hợp mạch điều khiển có chức năng hiển thị nội dung nhập trên màn hình và lưu trữ nội dung nhập dưới dạng mã hexa vào một vùng nhớ tạm, và sau đó chuyển dữ liệu nhập này đến bo mạch Raspberry Pi qua đường giao tiếp;

ii) đường giao tiếp giữa màn hình tinh thể lỏng và bo mạch Raspberry Pi là các đường giao tiếp nối tiếp thông dụng như UART (Universal asynchronous receiver transmitter) hay SPI (Serial Peripheral Interface) giúp truyền dữ liệu văn bản đã được nhập từ màn hình tinh thể lỏng sang cho vi xử lý lưu trữ và dùng làm đầu vào để tổng hợp tiếng nói từ văn bản tiếng Việt (TTS-Text to Speech);

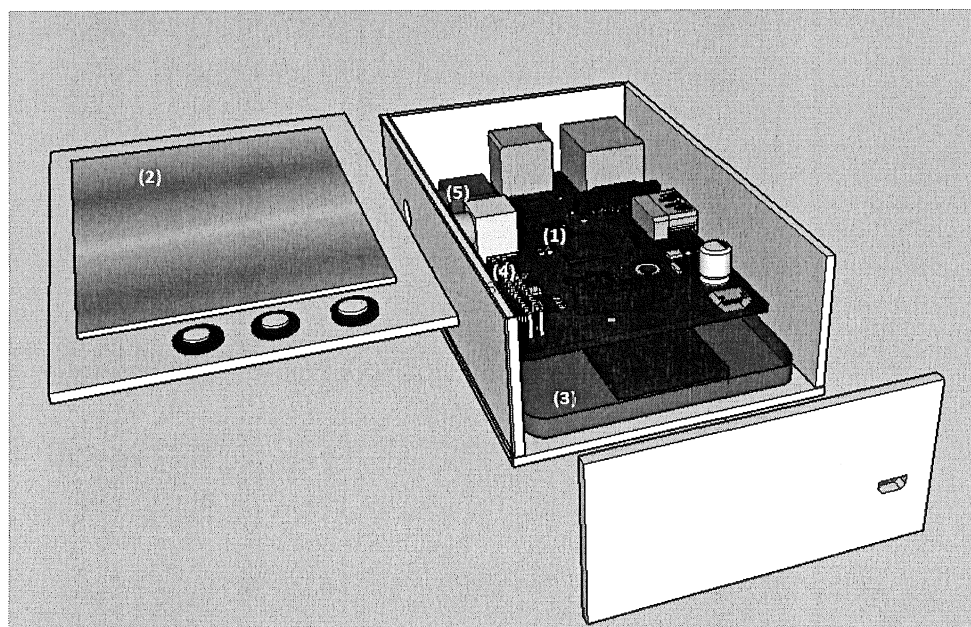
iii) bo mạch Raspberry Pi với vi xử lý là một DSP (Digital Signal Processor) lõi ARM11 xung nhịp 700MHz có tích hợp khối phần cứng xử lý dấu chấm động (Vector Floating Point Unit) để tổng hợp tiếng nói từ văn bản tiếng Việt thực hiện tổng hợp tiếng nói từ văn bản (TTS-Text to Speech) dựa trên mô hình Markov ẩn (HMM - Hidden Markov Model); và

iv) loa tích hợp trên bo mạch dùng để phát dữ liệu âm thanh sau khi đã tổng hợp xong.

PHỤ LỤC



Hình 1



Hình 2