THIẾT BỊ ĐIỆN TỬ TỰ ĐỘNG CHÚ THÍCH ĐỊA DANH, ĐỊA ĐIỂM BẰNG GIỌNG NÓI TIẾNG VIỆT

PHẠM THANH TUYÊN
Trung tâm Nghiên cứu và Đào tạo thiết kế vi mạch
Đại học Quốc gia TP Hồ Chí Minh

Bảng mạch điện tử tự động chú thích địa danh, địa điểm bằng giọng nói dựa trên hệ thống của Trung tâm Nghiên cứu và Đào tạo thiết kế vi mạch (ICDREC) và giải thuật phiên bản 1.6 của Phòng thí nghiệm trí tuệ nhân tạo (AILAB). Giải thuật này được tối ưu hóa về tốc độ truy xuất 10 ms/từ trên hệ thống dung lượng bộ nhớ chỉ còn 64 MB bộ nhớ RAM và 256 MB bộ nhớ Flash... và được tinh chỉnh cũng như sắp xếp lại nhằm thích hợp vận hành trên hệ thống nhúng có tài nguyên nhỏ. Thiết bị là sự kết hợp giữa công nghệ chuyển đổi văn bản thành giọng nói (TTS) dành cho tiếng Việt và công nghệ định vị toàn cầu (GPS). Bộ dữ liệu được tập trung vào các địa danh, địa điểm nổi tiếng. Bảng mạch điện tử này được thực hiện bằng vi xử lý ARM chạy trên hệ điều hành Linux. Nó sẽ định hướng cho việc sản xuất các sản phẩm TTS tiếng Việt chuyên dụng có thể sử dụng trong các lĩnh vực như giáo dục, giao thông, truyền thông, dân dụng...

Từ khóa: *tổng hợp tiếng nói, chuyển đổi văn bản thành giọng nói, định vị toàn cầu, vi xử lý ARM, hệ thống nhúng.*

Giới thiệu

TTS đã trở thành một công nghệ đang rất phát triển và có nhiều ứng dụng trên thế giới, hỗ trợ nhiều ngôn ngữ khác nhau [1-5, 18] nhưng TTS dành cho tiếng Việt vẫn chưa nhiều. Mặc dù các công trình nghiên cứu về TTS tiếng Việt đã có nhiều thành quả đáng kể nhưng hầu hết các ứng dụng đều được thực hiện trên máy tính [19-24] với tài nguyên lớn trong khi trên thế giới, giải thuật TTS được áp dụng trên rất nhiều sản phẩm cầm tay chuyên dụng [13-17]. Mặt khác, công nghệ TTS trên thế giới đã phát triển đến mức cứng hóa toàn bộ giải thuật trên một chip đơn [6-11]. Do đó, việc chứng minh tính khả thi của các thuật toán tổng hợp tiếng nói trên phần cứng có tài nguyên nhỏ hơn để đi tới việc cứng hóa trên chip các giải thuật là việc làm hết sức cần thiết. Việc tích hợp TTS vào hệ thống nhúng hoặc chip đơn sẽ mang lại nhiều tiềm năng ứng dụng trong thực tế hơn rất nhiều so với việc vận hành nó trên máy tính server hoặc máy tính cá nhân. Sự khác biệt lớn nhất là hệ thống vận hành trên chip đơn có kích thước nhỏ gọn, tính linh hoạt cao, dễ dàng tích hợp vào các hệ thống có sẵn của người dùng thông qua các chuẩn giao tiếp phổ biến hiện nay như: UART, I2C hoặc SPI. Giá thành của một hệ thống nhúng cũng rẻ hơn rất nhiều (khoảng 50-60 USD) nếu đem so sánh với hệ thống trên máy tính cá nhân hoặc server.

Để chuyển từ mô hình TTS trên máy tính sang hệ thống nhúng, tác giả đã đề xuất một phần cứng nhỏ gọn; phương pháp mã hóa, sắp xếp, tinh chỉnh bộ dữ liệu dành riêng cho các hệ thống có tài nguyên nhỏ; cải thiện chất lượng tiếng nói tổng hợp đầu ra - gọi tắt là "Phương pháp tổng hợp ghép nối chuyên dụng".

Như vậy, để từng bước làm chủ công nghệ đầy tiềm năng này, ICDREC quyết định khởi đầu bằng việc thiết kế và chế tạo một thiết bị chú thích địa danh, địa điểm bằng giọng nói được kết hợp giữa công nghệ TTS và GPS. Thiết bị này có thể tạo ra

A BOARD WHICH NOTIFIES GEOGRAPHIC NAME AND LOCATION AUTOMATICALLY BY VIETNAMESE SPEECH

Summary

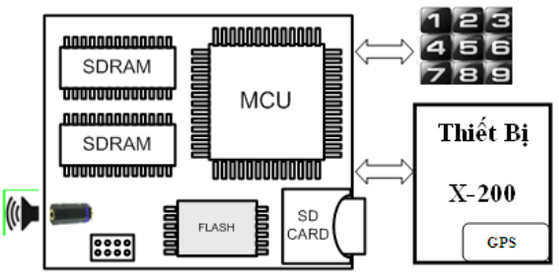
A board which notifies geographic name and location automatically by Vietnamese speech was proposed basing on ICDREC system and algorithm of Artificial Intelligence Laboratory (AILAB) version 1.6. The algorithm was optimized for the access speed approximately 100ms/1 word, 64MB RAM and 256MB Flash memory; and it was refined and rearranged to be suitable for the small embedded resource system. The board was the combination of Text To Speech (TTS) for Vietnamese and Global Positioning System (GPS) technology. Databases focused on famous places. The board was built by using ARM microprocessor, running on Linux operating sytem. It will orient people to produce dedicated Vietnamese TTS products. It can be used in many areas: education, transportation, communication and civilian applications...

***Key words:** Speech synthesis, text to speech, global positioning system, ARM microprocessor, embedded system.*

ngay dạng sản phẩm ứng dụng có thể định hướng thị trường. Mặt khác, thông qua đó để nắm bắt rõ ràng và chi tiết về giải thuật TTS tiếng Việt hướng đến mục tiêu xa hơn là cứng hóa hoàn toàn giải thuật này.

Thiết kế hệ thống

Sơ đồ tổng quát hệ thống chính với các thành phần được thể hiện trong hình 1.



Hình 1: sơ đồ khối của card TTS giao tiếp với thiết bị giám sát hành trình X-200

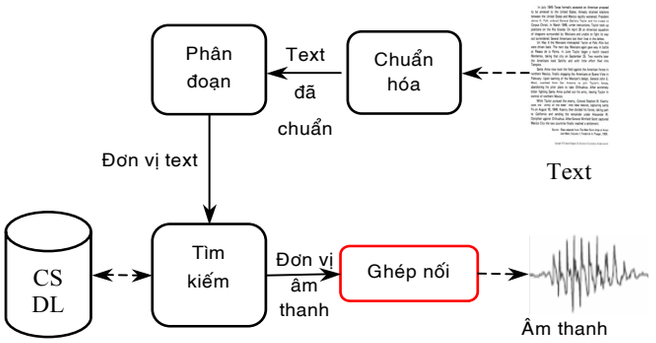
Thiết bị bao gồm 2 thành phần chính: card TTS và thiết bị giám sát hành trình (X200) có gắn module GPS. X200 sẽ liên tục gửi tọa độ GPS vị trí hiện tại của thiết bị về cho card TTS xử lý tính toán khoảng cách và xuất ra âm thanh.

Phần cứng hệ thống card TTS được tối ưu hóa với 6 thành phần chính: vi xử lý ARM9 tốc độ 180 MH_Z, bộ nhớ RAM 64 MB, bộ nhớ FLASH 256 MB, thẻ nhớ micro SD 4 GB, chip giải mã âm thanh định dạng wav/mp3 và ngõ vào ra giao tiếp với thiết bị bên ngoài qua giao thức truyền/nhận nối tiếp bất đồng bộ USART (Universal asynchronous receiver/transmitter).

Giải thuật TTS trên hệ thống nhúng

Giải thuật TTS dành riêng cho tiếng Việt phiên bản 1.6 được phát triển bởi AILAB sử dụng phương pháp tổng hợp ghép nối. 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 [19]. Thuật toán đòi hỏi tài nguyên phần cứng lớn, AILAB phải sử dụng máy chủ server để vận hành hệ thống này. Như vậy, việc chỉnh sửa và tối ưu thuật toán, sắp xếp lại bộ cơ sở dữ liệu nhằm phù hợp với hệ thống nhúng vốn có tài nguyên nhỏ nhằm tiết kiệm chi phí là việc làm tối quan trọng. Phương pháp tổng hợp ghép nối chuyên dụng được dùng cho thiết bị nhúng đề xuất 3 vấn đề chính: sắp xếp tinh chỉnh theo quy luật; mã hóa từ, cụm từ và chuẩn hóa dữ liệu.

Bộ từ điển của từ, cụm từ tiếng Việt và các từ viết tắt, từ nước ngoài có dung lượng gần 50 MB. Dung lượng này là rất lớn so với một hệ thống nhúng thông thường. Dữ liệu này chỉ có thể được lưu trữ trên RAM để tăng tốc độ truy xuất. Trong khi bộ nhớ RAM của hệ thống dự kiến chỉ có 64 MB cho toàn bộ chương trình và dữ liệu. Nếu tăng dung lượng bộ nhớ của hệ thống lên sẽ dẫn tới việc tăng giá thành và diện tích bảng mạch. Trong khi đó, bộ dữ liệu âm thanh có dung lượng gần 2,5 GB, được giải quyết bằng cách lưu trữ bộ dữ liệu này vào thẻ nhớ.

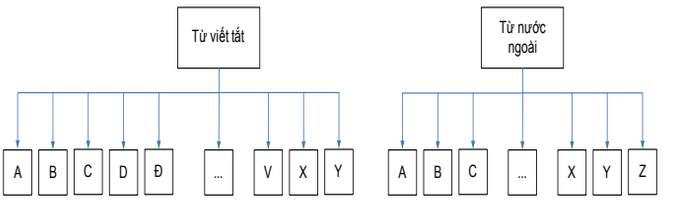


Hình 2: giải thuật TTS dành riêng cho tiếng Việt sử dụng phương pháp tổng hợp ghép nối

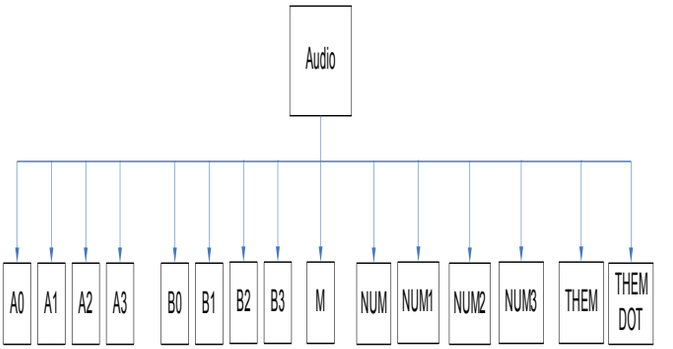
Các giai đoạn cần phải được tối ưu và có giải thuật mới ở đây gồm có: chuẩn hóa, phân đoạn và tìm kiếm. Chuẩn hóa văn bản là quá trình tiền xử lý văn bản theo một định dạng chuyên biệt dành riêng cho việc phân tích từ loại và cú pháp. Bước này thực hiện hai nhiệm vụ cơ bản là xóa bỏ các ký tự dư thừa, ký tự không có ý nghĩa phát âm như các dấu '?', '!'... và chuyển các từ viết tắt, từ tiếng nước ngoài, số, ngày tháng,… thành từ tiếng Việt đọc được.

Trong [19], toàn bộ văn bản sẽ duyệt lần lượt qua các bộ chuẩn hóa: số, ngày tháng, tiếng nước ngoài, từ viết tắt… Như vậy, nếu có bao nhiêu bộ chuẩn hóa sẽ có bấy nhiêu lần đọc toàn bộ văn bản. Việc chuẩn hóa theo phương pháp này không thể áp dụng được vì sẽ tốn thời gian xử lý cực lớn với một file văn bản dài trên một hệ thống có tài nguyên nhỏ, tốc độ thấp. Trong hệ thống nhúng sẽ không thực hiện theo cách đó. Đầu tiên, toàn bộ văn bản sẽ được chuyển thành dạng viết hoa. Sau đó mỗi từ, cụm từ của văn bản sẽ được duyệt qua các bộ xử lý này. Do vậy, hệ thống sẽ chỉ tốn thời gian đọc file văn bản 1 lần. Thời gian xử lý sẽ nhanh hơn.

Từ điển từ viết tắt, tiếng nước ngoài và file âm thanh được chia nhỏ ra thành từng thư mục, từng file theo cấu trúc sau (hình 3 và 4):



Hình 3: cấu trúc thư mục từ viết tắt và từ nước ngoài



Hình 4: cấu trúc thư mục âm thanh

Việc chia nhỏ và sắp xếp theo quy luật sẽ giúp truy xuất quy đổi 1 từ viết tắt hoặc từ nước ngoài nhanh hơn. Thay vì phải truy xuất tối đa tất cả các trường hợp A-Z thì vi xử lý chỉ phải truy xuất 1 từ duy nhất dựa vào ký tự đầu tiên của từ đó do từ viết tắt và tiếng nước ngoài không có dấu.

Trong [19] bộ từ điển từ viết tắt được sắp xếp theo quy luật: Từ viết tắt_Nghĩa_Ngữ cảnh trái_Ngữ cảnh phải.

Ví dụ: CHXHCNVN#CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM#TOÀN QUYỀN NƯỚC_TẠI THUỴ ĐIỂN TRỊNH

Bảng 1: nghĩa của từ viết tắt và ngữ cảnh

Từ	Nghĩa	Ngữ cảnh trái	Ngữ cảnh phải
CHXHCNVN	CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM	TOÀN QUYỀN NƯỚC	TẠI THUỴ ĐIỂN TRỊNH

Để chuyển từ viết tắt CHXHCNVN thành "Cộng hòa xã hội chủ nghĩa Việt Nam", đầu tiên bộ dò tìm từ viết tắt phải căn cứ vào ngữ cảnh trái và phải. Nếu câu văn đầu vào đúng trường hợp này thì từ

CHXHCNVN được quy đổi đúng và ngược lại sẽ không tìm thấy. Việc tìm từ dựa vào 1 ngữ cảnh duy nhất sẽ dẫn tới việc đa số trường hợp hệ thống phải tốn thời gian rất lớn để tìm hết tất cả các trường hợp từ viết tắt (3.602 từ) và sót từ vì CHXHCNVN không chỉ nằm trong ngữ cảnh duy nhất này mà còn trong rất nhiều ngữ cảnh đa dạng khác nhau. Hệ thống nhúng sẽ lược bỏ đi ngữ cảnh của từ viết tắt. Việc dò tìm sẽ không dựa vào ngữ cảnh mà căn cứ vào việc từ viết tắt được tìm thấy hay không. Điều này làm tăng tốc độ và xác suất quy đổi từ viết tắt ra tiếng Việt cao hơn rất nhiều.

Đánh giá mức độ tối ưu của giải thuật mới:

Gọi N là số từ, cụm từ cần đọc trong 1 văn bản tiếng Việt.

X là số lần xuất hiện của từ, cụm từ đó trong bộ từ điển tiếng Việt.

Y là số lần xuất hiện của từ, cụm từ đó trong bộ từ điển từ viết tắt.

Z là số lần xuất hiện của từ, cụm từ đó trong bộ từ điển tiếng nước ngoài.

A số lượng file âm thanh trong cùng 1 thư mục được chia nhỏ.

Bảng 2: đánh giá mức độ tối ưu

Bộ dữ liệu	Số lượng từ, file	Số lần truy xuất trên hệ thống máy tính	Số lần truy xuất hệ thống nhúng
Từ điển tiếng Việt	23092	Nx23092	NxX
Từ điển viết tắt	3602	Nx3602	NxY
Từ điển tiếng nước ngoài	7655	Nx7655	NxZ
Âm thanh	14371	Nx14371	NxA

Vì X<<23092, Y<<3602, Z<<7655 và N<<A nên giải thuật truy xuất trên hệ thống nhúng sẽ có tốc độ cao hơn rất nhiều so với trên máy tính. Hay nói cách khác, thời gian truy xuất sẽ được giảm đi rất nhiều.

Phân đoạn để tách từ, cụm từ theo phương pháp so khớp cực đại (longest matching). Thuật toán HASH dùng trong [19], mỗi từ điển sẽ được tải lên RAM trong lúc chạy chương trình. Như vậy sẽ có 3 từ điển, việc này dẫn tới tiêu tốn một dung lượng RAM rất lớn (khoảng 30 MB). Hệ thống nhúng dùng thuật toán SHA1 (Secure Hash Algorithm 1) để tạo ra một bộ từ điển có khả năng truy xuất nhanh và tốn bộ nhớ ít. Bộ từ điển có cấu trúc mỗi từ hoặc cụm từ cùng với ngữ cảnh của nó sẽ là một file text với tên của file chính là mã SHA1 của từ, cụm từ đó. Sau khi tìm ra được phân đoạn dài nhất, phân đoạn này sẽ được đưa qua hàm SHA1 để tạo ra số tương ứng và duy nhất của nó. Như vậy, thay vì dò tìm một cách trực tiếp dựa vào từ, cụm từ dưới dạng text sẽ chuyển sang dò tìm mã SHA1 với chiều dài nhỏ hơn nhiều. Định dạng của các file này như bảng 3 và 4.

Bảng 3: cấu trúc file từ điển gốc

Tên từ, cụm từ	Tên file âm thanh	Mã SHA1 ngữ cảnh trái	Mã SHA1 ngữ cảnh phải	Vị trí đầu tiên	Vị trí kết thúc

Bảng 4: cấu trúc file từ điển sau khi định dạng lại

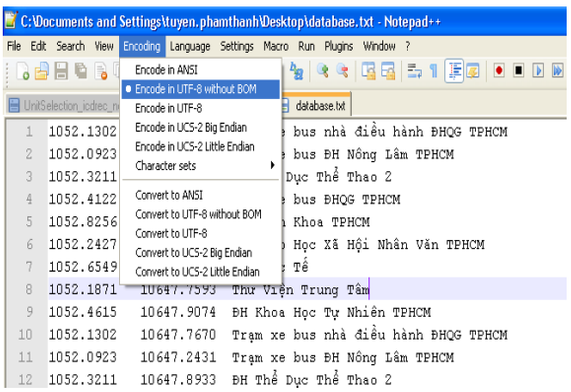
Tên file âm thanh	Mã SHA1 ngữ cảnh trái	Mã SHA1 ngữ cảnh phải	Vị trí đầu tiên	Vị trí kết thúc

Với cách định dạng mới, tên của từ trùng với tên file nên sẽ bỏ được cột Tên từ, cụm từ làm cho kích thước của bộ từ điển giảm từ 48,4 MB xuống còn 38,6 MB.

Các chức năng và định dạng hỗ trợ

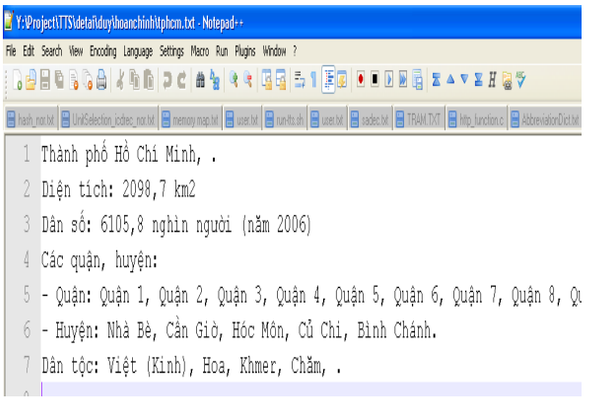
Card tổng hợp tiếng nói hỗ trợ: 23.092 từ, cụm từ tiếng Việt, 7.655 từ nước ngoài, 3.602 từ viết tắt. Cơ sở dữ liệu âm thanh bao gồm 14.370 tập tin âm thanh. File text có định dạng Encode in UTF-8 without BOM, được soạn thảo dễ dàng bằng phần mềm Notepad++.

Cấu trúc file lưu trữ tọa độ và tên địa danh bao gồm: 9 byte vĩ độ, 10 byte kinh độ và tên của địa danh. Mỗi trường này cách nhau một tab. Định dạng của file như hình 5.



Hình 5: file text lưu trữ tọa độ và tên địa danh

File lưu trữ thông tin giới thiệu về địa danh, địa điểm được thu thập từ các báo, tạp chí, các trang web về du lịch. Định dạng của file như hình 6.



Hình 6: file text lưu trữ thông tin địa danh

Việc tính toán khoảng cách giữa thiết bị và địa điểm trong cơ sở dữ liệu theo công thức "haversine" [12]:

$$a = \sin^2(\Delta\varphi/2) + \cos(\varphi_1).\cos(\varphi_2).\sin^2(\Delta\lambda/2)$$

$$c = 2.\mathrm{atan}^{\uparrow}2\ (\sqrt{a}, \sqrt{1-a})$$

Khoảng cách: d = R.c

Trong đó:

φ là vĩ độ, λ là kinh độ, R là bán kính của trái đất (6.371km)

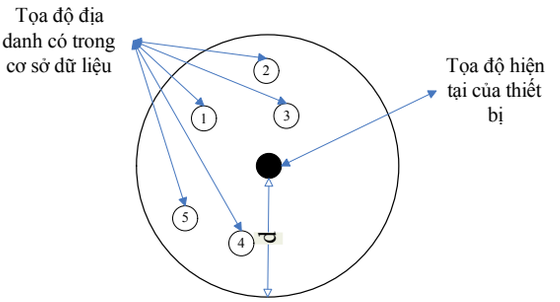
$\Delta\varphi$ = vĩ độ 2 - vĩ độ 1 (đơn vị radian)

$\Delta\lambda$ = kinh độ 2 - kinh độ 1 (đơn vị radian)

Vĩ độ 1, kinh độ 1 là tọa độ của địa danh trong file text hình 7

Vĩ độ 2, kinh độ 2 là tọa độ hiện tại thu được từ GPS khi thiết bị di chuyển trên mặt đất.

Như vậy, với bán kính d sẽ có thể xảy ra trường hợp nhiều địa danh nằm trong cùng một bán kính này. Lúc đó thiết bị sẽ đọc liệt kê thứ tự các địa danh này. Người sử dụng sẽ lựa chọn địa danh cần nghe thông tin bằng bàn phím số.



Hình 7: khoanh vùng địa điểm trong bán kính d

Độ sai số khi tính khoảng cách giữa tọa độ hiện tại của thiết bị và tọa độ có trong cơ sở dữ liệu sẽ phụ thuộc vào sai số GPS. Sai số này từ 5-15 m ngoài trời.

Ngoài ra, card còn hỗ trợ chế độ tiết kiệm năng lượng khi không sử dụng. Trong chế độ hoạt động bình thường card tiêu thụ dòng trung bình 300 mA. Khi không sử dụng, có thể đưa card vào chế độ tiết kiệm năng lượng bằng 1 xung vào chân WAKEUP/SLEEP, lúc đó dòng tiêu thụ sẽ chỉ còn 100 mA. Khi card ở chế độ này có thể đánh thức bằng 1 xung tiếp theo.

Phần cứng giao tiếp giữa card TTS và thiết bị khác thông qua chuẩn USART, đây là giao thức đơn giản và phổ biến ở các loại MCU hiện nay.

Tinh chỉnh và tối ưu từ điển của hệ thống

Trong thuật toán "longest matching" [19], khi tìm một từ hoặc một cụm từ trong bộ từ điển dựa vào ngữ cảnh trái, phải. Đầu tiên, tìm từ có cả ngữ cảnh trái và phải. Nếu không thấy sẽ tìm từ có ngữ cảnh trái, nếu không tìm từ có ngữ cảnh phải. Nếu tất cả các trường hợp trên đều không thõa mãn (từ, cụm từ không có ngữ cảnh) thì sẽ dùng từ có ngữ cảnh bất kỳ trong bộ từ điển. Hệ thống sẽ lấy ngay trường hợp đầu tiên để sử dụng. Tuy nhiên, không phải lúc nào ngữ cảnh đầu tiên này cũng có âm điệu tốt. Điều này dẫn đến việc câu nói sẽ bị quá bổng hay

quá trầm, gây khó chịu cho người nghe. Phương án đặt ra là phải sắp xếp lại những trường hợp này cho rơi vào những ngữ cảnh có âm trầm (không quá cao hoặc quá thấp), mặc dù không đúng ngữ điệu nhưng cũng đủ để người nghe có thể nhận diện được từ gì.

Xử lý từ không có trong từ điển hoặc những từ viết sai chính tả: nhìn tổng quan hai loại này giống nhau và khó có thể phân biệt được ra dạng nào. Thông thường có hai cách để xử lý. Thứ nhất là tách từ ra thành từng chữ để đọc riêng lẻ. Thứ hai là xem nó như là một âm câm. Hệ thống này chọn cách thứ hai.

Mặt khác, do đây là bộ từ điển cắt bằng phương pháp nhận dạng tiếng nói nên cũng có những từ được cắt ra không chính xác vị trí hoặc không trùng khớp giữa bộ từ điển từ và bộ âm thanh. Do đó cần đòi hỏi quá trình nghe và tinh chỉnh lại cho phù hợp với từng lĩnh vực cụ thể. Bộ dữ liệu của thiết bị đã được tinh chỉnh nhằm hỗ trợ tốt nhất lĩnh vực du lịch và giao thông công cộng.

Thiết bị đã được thiết kế và chế tạo với kích thước nhỏ gọn 7,7x6 cm^2. Tốc độ tổng hợp tiếng nói là 10 ms/1 từ đơn. Bộ từ điển hỗ trợ: 23.092 từ, cụm từ tiếng việt. 7.655 từ nước ngoài; 3.602 từ viết tắt. Dễ dàng tích hợp lên các hệ thống có sẵn khác. Thử nghiệm trên bộ dữ liệu gồm 50 địa danh mẫu.

Hình 8: thiết bị điện tử chú thích địa danh, địa điểm bằng giọng nói tiếng việt

Kết luận

Bài báo đã đề xuất "Phương pháp tổng hợp ghép nối chuyên dụng", bao gồm các thuật toán, sắp xếp, tinh chỉnh và tối ưu hóa giải thuật để có thể áp dụng trên hệ thống nhúng có tài nguyên nhỏ; thiết kế, chế tạo card tổng hợp tiếng nói dành riêng cho tiếng Việt. Ứng dụng cụ thể vào thiết bị điện tử tự động chú thích địa danh, địa điểm bằng giọng nói. Ngoài ra, thiết bị còn có khả năng ứng dụng vào lĩnh vực giáo dục như dạy tiếng Việt cho trẻ em, người nước ngoài…; giới thiệu thông tin các hiện vật trong bảo tàng, di tích nhằm phục vụ cho khách du lịch ■

Tài liệu tham khảo

[1] Aylett M.P. *Synthesising hyper articulation in unit selection TTS*. In Proceedings of Eurospeech 2005 (2005).

[2] Colotte V. and Beaufort. *R. Linguistic features weighting for a text-to-speech system without prosody model*. In Proceedings of Eurospeech 2005 (2005).

[3] Hirai T. and Tenpaku S. *Using 5 ms segments in concatenative speech synthesis*. In 5th ISCA Workshop on Speech Synthesis (2005).

[4] Lambert T. and Breen A. *A database design for a TTS synthesis system usinglexical diphones*. In Proceedings of the International Conference on Speech and LaguageProcessing 2004 (2004).

[5] Pollet V. and Coorman G. *Statistical corpus-based speech segmentation*. In Proceedingsof the Interspeech 2004 (2004).

[6] http://www.textspeak.com/oemtts.htm

[7] http://www.rcsys.com/chips.htm

[8] http://www.ispeech.org/

[9] http://www.sparkfun.com/products/9811

[10] http://www.rubidium.com

[11] http://www.speechchips.com/shop/

[12] http://www.movable-type.co.uk/scripts/latlong.html

[13] http://www.alcatel-lucent.com

[14] http://www.99er.net/spkspell.html

[15] http://www.loopycellphones.com/motorola-h17txt-with-motospeak-text-to-speech-bluetooth-headset

[16] http://www.eetimes.com/electronics-news/4085809/Intel-debuts-portable-text-to-speech-reader)

[17] http://garmin.com

[18] Yu Z.L. Wang, K.Z. Zu, Y.Q. Yue, D.J. and Chen G.L. *Data pruning approach to unit selection for inventory generation of concatenative embeddable chinese tts systems*. In Proceedings of the Interspeech 2004 (2004).

[19] Cao Xuân Nam, Vũ Hải Quân, Đinh Điền, Đậu Ngọc Hà Dương, Châu Thành Đức. *VOS report* (2009).

[20] Festival, http://www.cstr.ed.ac.uk/projects/festival/

[21] Lê Tang Hồ. Nhu liệu đọc tiếng Việt, http://noitiengviet.ca/

[22] Lê Hồng Minh. vnspeech, http:www.freewebs.com/vnspeech/

[23] Pham Thanh Nam. Tiếng nói Việt Nam, http://sourceforge.net/projects/vietnamesevoice/

[24] Lương Chi Mai. VnVoice, http://www.vndocr.com/