Kiến trúc vi mạch cho nhận dạng tiếng nói tiếng Việt thiết kế theo quy trình ASIC, trên nền công nghệ 130 nm TSMC

Hoàng Trang*, Phạm Đăng Lâm, Trần Văn Hoàng

Trường Đại học Bách khoa, Đại học Quốc gia TP Hồ Chí Minh

Ngày nhận bài 19.3.2015, ngày chuyển phản biện 25.3.2015, ngày nhận phản biện 21.4 .2015, ngày chấp nhận đăng 26.4.2015

Nhận dạng tiếng nói đã được nghiên cứu từ hơn 60 năm qua. Những nỗ lực đầu tiên được thực hiện từ những năm 50 đến đầu những năm 70 của thế kỷ trước, hệ thống nhận dạng tiếng nói được thiết kế để nhận dạng phát âm rời rạc trong môi trường nhiễu thấp, chủ yếu là các hệ thống với bộ từ vựng nhỏ (10-100 từ), trong trường hợp người nói cũng là người huấn luyện. Ngày nay, các hệ thống nhận dạng với số từ vựng lớn được xây dựng trên nền tảng hệ thống máy tính với tốc độ xử lý cao. Khi mật độ tích hợp vi mạch tăng, việc tiếp cận ứng dụng nhận dạng trên phần cứng hay các thiết bị cầm tay trở nên khả thi. Trong nghiên cứu này, nhóm tác giả trình bày một kiến trúc vi mạch được thiết kế theo quy trình ASIC, trên nền công nghệ 130 nm TSMC, ứng dụng trong nhận dạng giọng nói tiếng Việt, để đáp ứng yêu cầu khắt khe về hiệu năng nhận dạng và tính thời gian thực trong các ứng dụng thực tế.

***Từ khóa:* dãy cổng lập trình được (FPGA), hàm phân bố xác suất Gauss, hệ thống nhận dạng giọng nói tự động (ASR), mô hình Markov ẩn (HMM), trích đặc trưng thang tần số mel (MFCC).**

Chỉ số phân loại 1.2

AN ASIC BASED ARCHITECTURE FOR VIETNAMESE SPEECH RECOGNITION ON THE BASIC OF 130 NM TSMC TECHNOLOGY

Summary

Speech recognition has been researched over sixty years. The first researches were conducted from the 1950s till the early 1970s, some complete recognition systems were developed to recognise incoherent pronounciation in low noise conditions and only adapted to small word libraries (10-100 words) which belong to the trainers as well as the recognition persons. Today, the recognition systems adapting to the large word library are built based on computer system with high performance. Moreover, when integrated density is enhanced, the access to the hardware or handset applying recognition technology becomes feasible. In this work, an ASIC based architecture for Vietnamese speech recognition on the basis of 130 nm TSMC technology is illustrated to meet the real time requirements as well as confirm the highly effective performance.

***Keywords:* auto speech recognition (ASR), field-programmable gate array (FPGA), Gaussian probability distribution, hidden Markov model (HMM), mel-frequency cepstral coefficient (MFCC).**

Classification number 1.2

Đặt vấn đề

Các nhóm tác giả trong [1, 2] đã trình bày lý thuyết về trích đặc trưng tiếng nói dùng phương pháp MFCC và bộ giải mã tiếng nói dùng HMM với nhiều cải tiến mới dùng cho nhận dạng các số tiếng Anh từ 0 đến 9. Các nhóm tác giả trong [3] đã thực hiện thành công trên FPGA Cyclone II, với việc nhận dạng 1 số từ, thể hiện bằng việc bật sáng các đèn led bằng giọng nói. Trong bộ trích đặc trưng, do tính chất phức tạp khi thực thi phần cứng phép biến đổi FFT, nên nhóm tác giả này sử dụng module FFT 1024 điểm, được hỗ trợ sẵn trong IP Megacore của Altera. Đối với bộ giải mã tiếng nói, nhóm tác giả trong [4] dùng phương pháp lượng tử vectơ (VQ), phương pháp này chỉ thích hợp cho nhận dạng với bộ từ vựng nhỏ (khoảng 10-20 từ). Ngoài ra, các tác giả này cũng đưa

*Tác giả chính: Email: hoangtrang@hcmut.edu.vn

ra hướng phát triển tiếp theo là dùng mô hình HMM, nhưng yêu cầu thực thi phần cứng phức tạp. Tương tự, một số tổ chức nổi tiếng trên thế giới như ATR, AT&T hay IBM... [5-10] cũng đưa ra nhiều giải pháp khác nhau cho ứng dụng nhận dạng giọng nói, chủ yếu mô hình HMM được áp dụng một cách phổ biến. Bên cạnh đó, cũng có các công trình đề nghị cách xây dựng môi trường kiểm tra thiết kế nhận dạng tiếng nói [11], hay các thiết kế nâng cao hơn trong giao tiếp bộ nhớ [12], cũng như thiết kế layout cho chip nhận dạng tiếng nói trên, những công nghệ mới như 40 nm [13]. Bảng 1 thể hiện danh sách và so sánh chip của các hãng trên thế giới.

Bảng 1: các sản phẩm nhận dạng giọng nói

STT	Tên chip	Hãng thiết kế, sản xuất	Số từ ⇔ đặc tính số giọng người nói	Thời gian nhận dạng	Độ chính xác	Độc lập/ phụ thuộc người nói
1	DVC306 D6106	DSP Communications	16 từ: 8 giọng nói của 8 người 128 từ: 1 giọng duy nhất	Đối với 16 từ: <1 s	92% (theo luật Hyde - Hyde's law)*	Không/có
2	HM2007	Motorola	20 từ: 1 giọng nói duy nhất	1,9 s	Không công bố cụ thể	Không/có
3	MSM 6679	OKI Group	Tối đa 25 từ: 2 giọng người nói	Không công bố cụ thể	97% (theo luật Hyde - Hyde's law)*	
4	RSC-164	SENSORY	20-50 từ: Không đề cập đến bao nhiêu giọng người nói, mà chỉ đề cập là phụ thuộc người nói (cần huấn luyện trước)	Không công bố cụ thể	93% (theo luật Hyde - Hyde's law)*	
5	TC 8860F	TOSHIBA	10 từ: Nhận dạng từ rời rạc, phụ thuộc người nói (cần huấn luyện trước)	Không công bố cụ thể	Không công bố cụ thể	Không/có
6	TC8865F-00	TOSHIBA	20 từ: Nhận dạng từ rời rạc, phụ thuộc người nói (cần huấn luyện trước)	Không công bố cụ thể	Không công bố cụ thể	Không/có
7	5A128	RICOH	10 từ: Phụ thuộc người nói 3 từ: Độc lập người nói	Tối đa: 5 s (đối với nhận dạng 10 từ) Tối đa: 2 s (đối với nhận dạng 3 từ)	95% (theo luật Hyde - Hyde's law)*	Có/có
8	RF5A65	RICOH	60 từ: 1 giọng nói	Tối đa: 2 s	92% (theo luật Hyde - Hyde's law)*	Không/có

(*) Luật Hyde (Hyde's law): luật này được đưa ra bởi R.S. Hyde và được sử dụng rộng rãi trong giới chuyên môn về lĩnh vực nhận dạng tiếng nói như sau: "độ nhận dạng chính xác của bộ nhận dạng tiếng nói là 98%, vì bộ nhận dạng tiếng nói có độ nhận dạng chính xác là 98% thì bộ mẫu kiểm tra phải được sắp xếp để chứng tỏ điều đó"

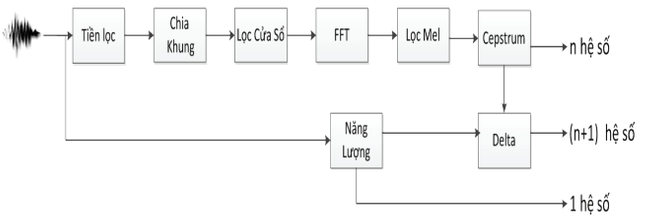
Như vậy, những thành quả cơ bản có được trong nhận dạng tiếng nói là nhờ kết hợp hai hướng tiếp cận phần cứng và phần mềm. Tuy nhiên, nhiều hạn chế còn tồn tại với các cách tiếp cận như: các hướng tiếp cận phương pháp huấn luyện vẫn thông qua hệ thống máy tính nhằm phục vụ những phép toán và giải thuật phức tạp; việc nhận dạng bằng phần cứng dựa trên các sản phẩm cầm tay còn hạn chế bởi khối lượng giải thuật đồ sộ cũng như khả năng mở rộng số từ vựng.

Ở Việt Nam, vấn đề này mới được quan tâm nghiên cứu trong những năm gần đây. Đến nay, đã có một số công trình nghiên cứu về lĩnh vực này, theo nhiều hướng tiếp cận khác nhau, song kết quả đạt được chưa cao, vẫn có khả năng phát triển hơn nữa, vì hầu hết chỉ theo hướng tiếp cận phần mềm là chủ yếu. Để khắc phục những tồn tại của các hệ thống nhận dạng giọng nói trên thế giới nói chung và ở Việt Nam nói riêng, nhóm nghiên cứu trình bày một kiến trúc vi mạch nhận dạng giọng nói tiếng Việt cải tiến ở cấp độ từ đơn. Thiết kế được sản xuất trên nền công nghệ 130 nm TSMC, sau đó được kiểm tra ứng dụng hoàn chỉnh trên bo mạch tự phát triển. Bài viết này là kết quả nghiên cứu của Đề tài "Nghiên cứu thiết kế lõi IP mềm, IP cứng cho IC nhận dạng tiếng nói tiếng Việt và chế tạo thiết bị trợ giúp người khuyết tật bằng tiếng nói", mã số KC01.23/11-15, thuộc Chương trình KH&CN trọng điểm cấp nhà nước KC01/11-15 về nghiên cứu ứng dụng và phát triển công nghệ thông tin và truyền thông.

Trích đặc trưng MFCC và mô hình HMM

Trích đặc trưng MFCC

Phương pháp trích đặc trưng MFCC được lựa chọn cơ bản rút trích các đặc trưng tần số của tiếng nói. Những bước chính của trích đặc trưng MFCC được mô tả như hình 1.

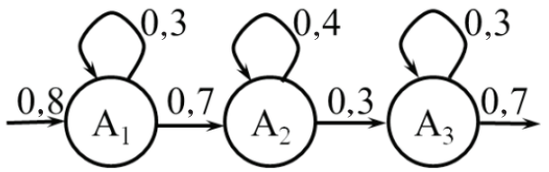


Hình 1: các bước trích đặc trưng MFCC

Chất lượng và xác suất nhận dạng của toàn bộ hệ thống ngoài phụ thuộc vào việc chọn lựa mô hình xác suất tối ưu, còn phụ thuộc nhiều vào chất lượng của các đặc trưng được trích xuất. Mặt khác, trích xuất MFCC gồm nhiều bước nhỏ khác nhau, trong đó một số hệ số nhạy cảm ảnh hưởng đến xác suất nhận dạng của hệ thống tồn tại bên trong những khối này, như hệ số của các bộ lọc và số bộ lọc tương ứng.

Mô hình xác suất HMM

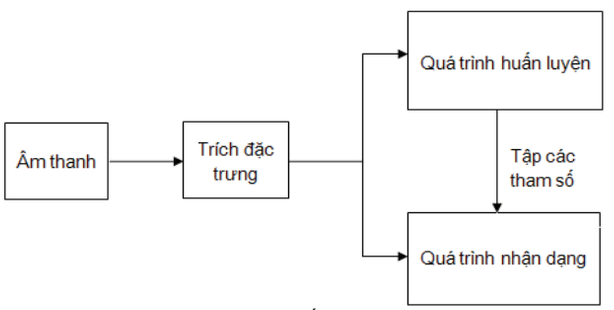
Mô hình HMM là mô hình thống kê quen thuộc, được áp dụng rộng rãi trong nhiều ứng dụng khác nhau. Trong ứng dụng nhận dạng giọng nói, mô hình HMM dạng tiến (hình 2) được sử dụng, mang lại những thành công nhất định. Mô hình này tương tự các mô hình HMM khác, các thông số bao gồm: xác suất chuyển trạng thái, tập các giá trị "mean" và "covarian" cũng như xác suất cho từng bộ trộn bên trong trạng thái.



Hình 2: mô hình HMM dạng tiến

Hệ thống nhận dạng giọng nói

Bất cứ một hệ thống nhận dạng giọng nói nào cũng bao gồm hai quá trình chính: quá trình huấn luyện và quá trình nhận dạng (quá trình giải mã). Trong cả hai quá trình này, không thể thiếu kiến trúc trích đặc trưng giọng nói, hình 3 mô tả các khối cơ bản về một hệ thống nhận dạng giọng nói hoàn chỉnh. Như đã nêu ở trên, quá trình huấn luyện yêu cầu tập mẫu các âm thanh đầu vào lớn, các giải thuật huấn luyện phức tạp, nên cần tiếp cận các hệ thống máy tính cho quá trình này. Sản phẩm của quá trình huấn luyện được gọi là bộ thông số mô hình. Quá trình huấn luyện dựa trên mô hình HMM tiến cơ bản đã được tìm hiểu và ứng dụng rộng rãi nên các công thức phức tạp không được đề cập.



Hình 3: mô hình hệ thống nhận dạng hoàn chỉnh

Tương tự quá trình huấn luyện, mẫu âm thanh được trích đặc trưng trước hết trong quá trình nhận dạng. Giả sử, sau khi trích đặc trưng cho ra 25 vector MFCC 26 chiều, có thể hiểu rằng ngõ vào của quá trình nhận dạng là 25 vector 26 chiều. Kết hợp với tập thông số mô hình từ quá trình huấn luyện gồm 50 mô hình, mỗi mô hình bao gồm tập xác suất chuyển trạng thái (16 giá trị chuyển trạng thái tương ứng 8 trạng thái), tập xác suất bộ trộn (32 giá trị tương ứng mỗi trạng thái có 4 bộ trộn), tập các cặp vector "mean" và "covariance" (32 cặp vector tương ứng mỗi bộ trộn là một cặp). Cụ thể, công thức để tính xác suất cho một vector đặc trưng 26 chiều với một trạng thái bao gồm 4 bộ trộn được thể hiện qua hàm phân bố Gauss như sau:

$$b_j(o)=\sum_{k=1}^{M} c_{jk}\,\mathcal{N}(o,\mu_{jk},U_{jk}),\quad 1\le j\le N$$

Với

$$\sum_{k=1}^{M} c_{jk}=1,\quad 1\le j\le N$$

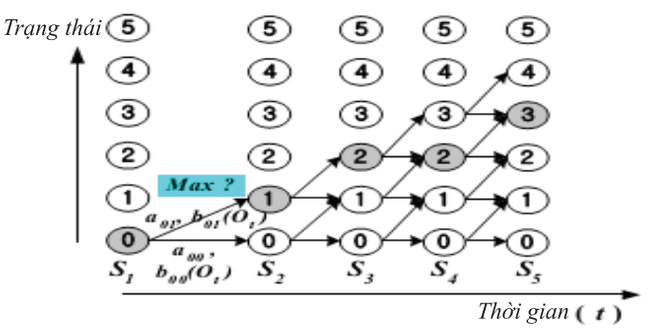
$$c_{jk}\ge 0,\qquad 1\le j\le N,\ 1\le k\le M$$

$$\mathcal{N}(o,\mu,U)=\frac{1}{\sqrt{(2\pi)^n|U|}}e^{-\frac{1}{2}(o-\mu)'U^{-1}(o-\mu)}$$

$$=\frac{1}{\sqrt{(2\pi)^{26}|U|}}e^{-\frac{1}{2}(o-\mu)'U^{-1}(o-\mu)}$$

Trong đó, N là số trạng thái (N = 8); M là số bộ trộn (M = 4); o là vector đặc trưng (hay còn gọi là chuỗi quan sát - 26 phần tử); μ là vector "mean" (26 phần tử); U là vector "covariance" (26 phần tử); c_{jk} là xác suất rơi vào bộ trộn thứ k trong trạng thái thứ j.

Vấn đề ở đây là, trong ví dụ trên ta có 25 vector MFCC trích đặc trưng cùng với 8 trạng thái, vậy thì vector nào tương ứng trạng thái nào sẽ cho kết quả xác suất lớn nhất? Theo logic, cần tính đến là 168 lần hàm phân bố xác suất Gauss, cùng với các thuật toán lựa chọn sẽ cho kết quả chính xác. Phương pháp này có thể xem là phương pháp vét cạn. Để cải thiện điều này, thuật toán Viterbi được mô tả như hình 4, đang được áp dụng rộng rãi.



Hình 4: giải thuật Viterbi

Kiến trúc vi mạch ứng dụng nhận dạng giọng nói

Trích đặc trưng MFCC

Phương pháp: kiến trúc vi mạch tiếp cận việc nhận dạng giọng nói bao gồm hai thành phần trích đặc trưng và nhận dạng giọng nói, như thể hiện ở hình 5.

Hình 5: các khối chính trong kiến trúc vi mạch nhận dạng giọng nói

Tuy nhiên, khi tiếp cận phần cứng, có một số vấn đề nảy sinh, trong đó FFT là vấn đề đáng lưu ý. Để có thể giải quyết những thông số ảnh hưởng xấu đến kết quả trích đặc trưng và nhận dạng, việc kết hợp giữa khảo sát và tính khả thi trong thực nghiệm trên là rất cần thiết. Mặt khác, các ngôn ngữ khác nhau và đặc tính vùng miền tạo nên sự khác biệt về các đặc trưng. Việc áp đặt các cấu hình cứng nhắc sẽ cho những kết quả không như mong đợi.

Thực nghiệm: kiến trúc trích đặc trưng tương tự như lý thuyết tiếp cận bao gồm các thành phần chính với những đặc trưng chi tiết đã được khảo sát trong bảng 2.

Bảng 2: cấu hình trích đặc trưng MFCC

Thông số	Chọn lựa	So sánh	
Có sử dụng bộ lọc	Có	Sử dụng bộ lọc	Xác suất (%)
		Có	92,7
		Không	83,3
Giá trị hệ số bộ tiền lọc	31/32	Hệ số	Xác suất (%)
		31/32	92,2
		15/16	84,3
		7/8	78,6
Số điểm trong một khung	160	Số điểm	Xác suất (%)
		160	92,3
		80	84,4
Tỷ lệ chồng lấp	50%	Tỉ lệ (%)	Xác suất (%)
		50	92,2
		40	90
		30	87
Số điểm FFT	256	Số điểm	Xác suất (%)
		160	92,3
		80	84,4
Cấu hình bộ lọc MEL	Vuông	Kiểu bộ lọc	Xác suất (%)
		Tam giác	94,5
		Chữ nhật	90,1
Số bộ lọc MEL	27	Số bộ lọc	Xác suất (%)
		27	94
		26	90
		25	91
		24	89
		23	90
Số hệ số ceptrum		12	
Số hệ số năng lượng		1	
Số hệ số cho một vector MFCC		26	

Mô hình HMM

Phương pháp: như đã nêu trên, việc tính theo khuôn mẫu với phương pháp vét cạn cho kết quả tốt nhất, nhưng do khả năng đáp ứng tính thời gian thực nên không khả thi. Nếu dựa trên yêu cầu ứng dụng cao về độ chính xác, phương pháp vét cạn vẫn được chọn lựa, nhưng trên thực tế, những ứng dụng cụ thể luôn đòi hỏi tính đáp ứng thời gian thực ở mức độ cần thiết. Do đó, để đáp ứng được hai yêu cầu khắt khe này, kỹ thuật song song được áp dụng nhằm tăng tốc độ xử lý. Tuy nhiên, kỹ thuật này có điểm yếu là quá trình quản lý, điều khiển dữ liệu phức tạp nên sự hao tốn tài nguyên cần được lưu ý chặt chẽ.

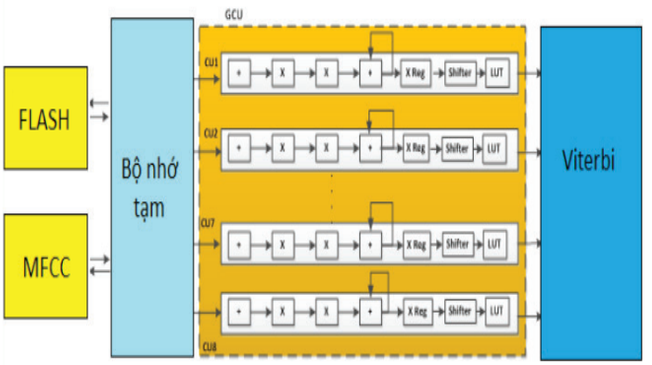
Một vấn đề khác cần được quan tâm khi tiếp cận phần cứng theo hướng ASIC là các kiến trúc và chức năng gần như không thể thay đổi. Đây là một trong những bất lợi khi tiếp cận ASIC thay vì FPGA. Để giải quyết vấn đề này, tiếp cận phần cứng sao cho khả năng cấu hình động là điều cần thiết, dựa trên mô hình MFCC tiếp cận với mỗi vector MFCC 26 chiều. Các thông số cơ bản của mô hình HMM cũng được giới thiệu chi tiết ở bảng 3 với khả năng thay đổi được.

Bảng 3: mô hình HMM tiếp cận phần cứng

Số trạng thái	8 (có thể thay đổi được)
Số bộ trộn	4
Số vector "mean"/"covariance"	26 (có thể thay đổi được)

Thực nghiệm: khi tiếp cận mô hình song song có hai giải pháp như sau: một vector MFCC được tính với nhiều trạng thái khác nhau; một trạng thái được tính với nhiều MFCC khác nhau.

Về khía cạnh phần cứng, số vector MFCC là không ổn định cho các mẫu từ dài ngắn khác nhau, ví dụ với các mẫu từ đơn, thường không dưới 10 vector MFCC. Trong khi đó, các mô hình HMM cho các ngôn ngữ khác nhau có số trạng thái khác nhau, nhưng tương đối nhỏ, số trạng thái thường lớn hơn hoặc nhỏ hơn gần với 10. Chính vì vậy, việc áp dụng cơ chế tính song song cho mỗi MFCC, với nhiều trạng thái khác nhau mang tính khả thi và ổn định cao. Một vấn đề khác cần được quan tâm là khối kiến trúc tính xác suất phân bố Gauss. Khối này giữ vai trò không thể thiếu trong quá trình tính toán các xác suất thành phần tương ứng với mỗi vector MFCC và trạng thái. Để có thể thực hiện được cơ chế song song, kiến trúc chi tiết được áp dụng như hình 6.



Hình 6: tổng quan các module trong quá trình nhận dạng

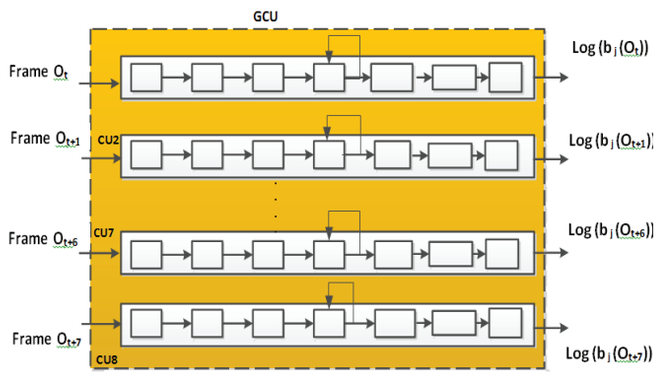
Có thể thấy, 8 luồng song song được áp dụng. Để tạo 8 luồng song song này, bộ nhớ tạm dùng để lưu trữ mô hình thông số cho một trạng thái và 8 MFCC một lúc được sử dụng. Chi tiết kiến trúc được mô tả ở phần sau:

Module Write - Mem: thực hiện việc đọc các giá trị mô hình trạng thái của quá trình huấn luyện từ solfware, được lưu vào trong FLASH (với các cấu hình sử dụng bộ nhớ FLASH được trình bày trong bảng 4), các giá trị đặc trưng của mẫu âm thanh được lưu trong SRAM.

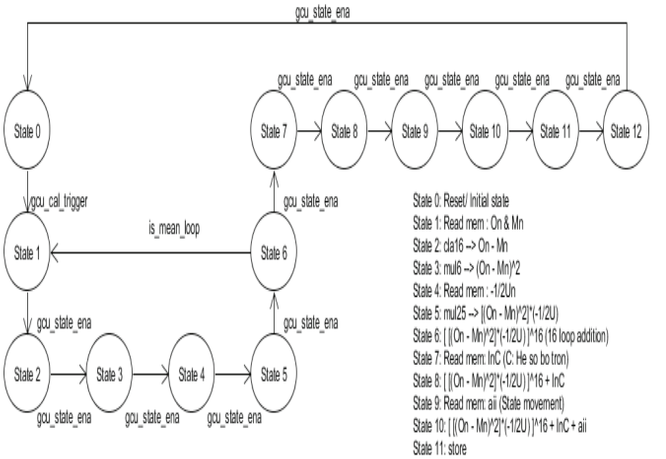
Bảng 4: cấu hình sử dụng bộ nhớ FLASH

[22:18] = 0	Information [17] '0': Các thông số mô hình '1': Các giá trị mô hình	Word_num [16:11]	State_num [10:8]	Mixture [7:6]	[5:0] Mean & Cov & lnC & aii & aij - aii

Module GCU: tính xác suất tại 1 bộ trộn của mỗi trạng thái cho từng MFCC khác nhau, theo 8 luồng dữ liệu. Thực hiện tính song song (theo sơ đồ tính toán trong hình 7) giúp giảm thời gian của quá trình tính giá trị xác suất cho mỗi bộ trộn. Sau khi tính được xác suất của 4 bộ trộn của 1 trạng thái, ta có thể thực hiện giải mã. Máy trạng thái (State machine) cho bộ GCU được thiết kế như trong hình 8.

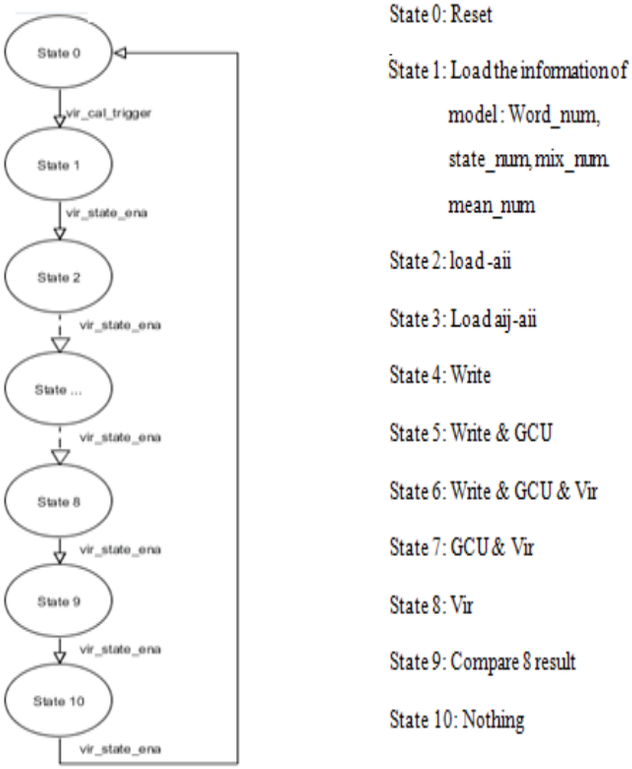


Hình 7: cơ chế tính song song 8 luồng dữ liệu



Hình 8: các trạng thái bộ GCU tính giá trị biot

Module Viterbi: thuật toán Viterbi vét cạn được thực hiện cứng hóa theo đúng như giải thuật giải mã, sau khi nhận giá trị từ bộ GCU (8 luồng), ta có được 8 giá trị xác suất, các giá trị này sẽ được tính toán với các giá trị xác suất chuyển trạng thái, chọn lọc giá trị lớn nhất sau cùng được giữ lại. Như vậy, quá trình được lặp lại nhiều lần mỗi khi có 8 giá trị được hoàn tất từ khối tính GCU, việc lặp lại này được mô tả bởi hình 9, đặc trưng cho quá trình chuyển trạng thái của toàn bộ hệ thống.

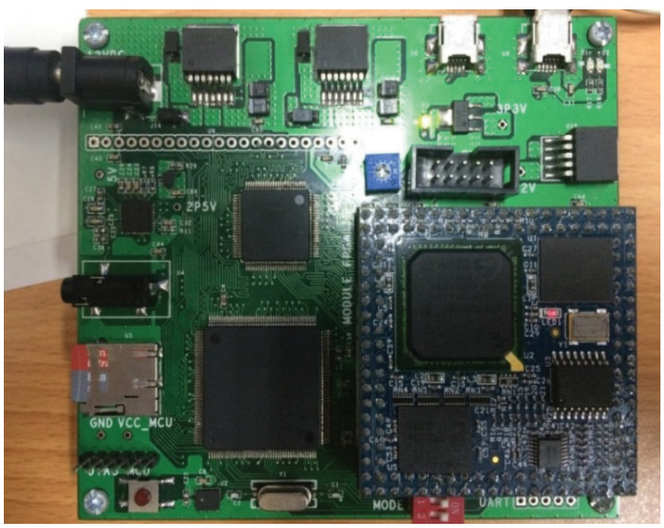


Hình 9: sơ đồ máy trạng thái bộ giải mã VIR

Xác suất lớn nhất cuối cùng tương ứng với chỉ số từ được lưu trong thanh ghi điểm số cuối cùng. Khi kết thúc, hệ thống xuất ra tín hiệu "Result_ack", cùng với chỉ số của từ trong thanh ghi điểm số cuối cùng, đó là kết quả nhận dạng.

Kết quả và thảo luận

Việc thực nghiệm trước hết được mô phỏng với các giản đồ xung clock, đảm bảo chức năng được thực hiện một cách chính xác. Sau khi xác nhận các kết quả mô phỏng là khả quan, một lần nữa toàn bộ kiến trúc được kiểm tra và chạy thử trên FPGA nhằm đảm bảo chính xác, tính hiện thực. Đây là bước không thể thiếu trong quy trình thiết kế chip ASIC, đòi hỏi khắt khe trong các khâu kiểm tra và hiện thực. Để có thể kiểm tra tính chính xác của thiết kế một cách thực tế, một mô hình trên kit thực tế đã được thực hiện. Trong mô hình này, toàn bộ thiết kế HMM được kiểm tra bằnh cách hiện thực trên kit FPGA tự xây dựng của nhóm nghiên cứu iHearTech (hình 10) [14]. Các kết quả về trích đặc trưng, thông số mô hình không còn xử lý dưới dạng file mà được nạp lên các kiến trúc phần cứng được hỗ trợ trên kit như SRAM, ROM…



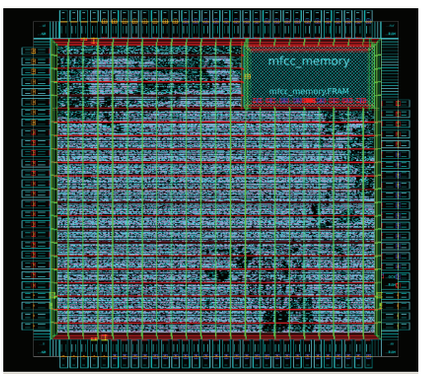
Hình 10: mô hình kiểm nghiệm trên kit FPGA của nhóm tác giả [14]

Lấy ngẫu nhiên một số mẫu đúng và sai (50 mẫu bất kỳ) đã mô phỏng của thiết kế phần cứng lẫn phần mềm, đem kiểm tra lại trên FPGA với mô hình kiểm tra trên, cho các kết quả tương đồng. Điều này đảm bảo tính chính xác, khả thi trong thực tế của hệ thống. Cụ thể, kết quả tiêu biểu thu được từ tập mẫu của nhiều người, với mô hình HMM cho hệ thống nhận dạng đã được hệ thống tiến hành thử nghiệm với bộ từ vựng gồm 50 từ (hình 11), được đánh giá cho thấy kết quả khả thi của thiết kế.

"KHÔNG","MỘT","HAI","BA","BỐN","NĂM","SÁU","BẢY","TÁM"
"CHÍN","LỊCH","SỬ","VĂN","HÓA","GIÁO","DỤC","KHOA","HỌC"
"NÔNG","NGHIỆP","CÁ","HEO","GÀ","VỊT","SỨC","KHỎE","CÂY"
"HOA","BẬT","TẮT","MỞ","ĐÓNG","ĐÈN","QUẠT","CỬA ","PHÒNG"
"KHÁCH","NGỦ","BẾP","DỪNG","BỎ","QUA","TIẾP","TỤC","TÔI"
"NGHE","MUỐN","TIN","CHÀO","BẠN"

Hình 11: bộ từ vựng cần nhận dạng

Kết quả nhận dạng phần cứng được so sánh với các kết quả phần mềm cũng như mô phỏng cho thấy, độ tương đồng lớn (92%). Điều này cho thấy tính khả thi và độ tin cậy của việc thực thi phần cứng. Sau khi các phương pháp kiểm tra được thực hiện chính xác, thiết kế được chuyển xuống tổng hợp và thực hiện ở các bước vật lý thấp hơn. Hình 12 cho thấy, sản phẩm cuối cùng hoàn thành ở cấp độ vật lý, trên công nghệ 130 nm TSMC. Hình 13 là chip iHearTech được chế tạo hoàn chỉnh của nhóm nghiên cứu iHearTech chúng tôi [14].



Hình 12: kiến trúc ở cấp độ vật lý công nghệ 130 nm TSMC

Hình 13: chip iHearTech được chế tạo hoàn chỉnh [14]

Kết luận

Nhóm nghiên cứu đã trình bày kiến trúc vi mạch được thiết kế theo quy trình ASIC, trên nền công nghệ 130 nm TSMC, ứng dụng trong nhận dạng giọng nói tiếng Việt. Các kết quả kiểm tra theo quy trình ASIC cũng như trên FPGA cho xác suất nhận dạng cao, thời gian dưới 1 giây, đáp ứng yêu cầu khắt khe về thời gian thực trong các ứng dụng cụ thể.

Các phát triển ứng dụng cụ thể, tích hợp sản phẩm vi mạch, sẽ được tiếp tục để hướng đến những giá trị thiết thực trong cuộc sống. Bên cạnh đó, kiến trúc vi mạch đề xuất cũng sẽ được nghiên cứu và phát triển với các định hướng nghiên cứu tối ưu về công suất, tốc độ cũng như các ứng dụng câu phức tạp.

Tài liệu tham khảo

[1] Wei HAN, Cheong - Fat CHAN, Chiu-Sing CHOY and Kong - Pang PUN (2006), "An efficient MFCC extraction method in speech recognition", *Circuits and systems, ISCAS 2006, IEEE international symposium*, pp.145-148.

[2] Wei HAN, Cheong - Fat CHAN, Chiu-Sing CHOY, Kong - Pang PUN (2005), "A speech recognizer with selectable model parameters", *Circuits and systems, ISCAS 2005, IEEE international symposium*, pp.5842-5845.

[3] Joe Tebelskis (1995), "Speech recognition using neural networks", *PhD thesis,* Carnegie Mellon University.

[4] Carlos Asmat, David López Sanzo, Kanwen Wu (2007), "Speech recognition using FPGA technology", *Master thesis,* University of McGill.

[5] Waibel A, Hanazawa H, Hintom G, Shikano K, Lang K.J (1989), "Phoneme recognition using time - delay neural networks", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37(3)**, pp.328-339.

[6] Kita K, Kawabata T, Saito H (1989), "HMM continuous speech recognition using predictive LR parsing", *in Proceedings IEEE international conference on acoustics, speech, and signal processing*, pp.703-706.

[7] Wilpon J.G, Lee C.H, Rabiner L.R (1991), "Improvements in connected digit recognition using higher order spectral and energy features", *in proceedings IEEE international conference on acoustics, speech and signal processing*, pp.349-352.

[8] Ney H (1990), "Experiments on mixture - density phoneme modelling for the speaker - independent 1000 word speech recognition DARPA task", *in proceedings IEEE international conference on acoustics, speech, and signal processing*, pp.713-716, Albuquerque, New Mexico, USA.

[9] Chow Y.L, Dunham M.O, Kimball O.A, Krasner M.A, Kubala G.F, Makhoul J, Price P.J, Roucos S, Schwartz R.M (1987), "BYBLOS: The BBN continuous speech recognition system", *in proceedings IEEE international conference on acoustics, speech, and signal processing,* pp.89-92, Dallas, Texas, USA.

[10] Kubala F, Feng M, Makhoul J, Schwartz R (1989), "Speaker adaptation from limited training in the BBN BYBLOS speech recognition system", *in proceedings of the DARPA speech and natural language workshop, morgan kaufmann publishers,* Inc., Palo Alto, California, USA, pp.100-105.

[11] Shingo Yoshizawa, Naoya Wada, Noboru Hayasaka, Yoshikazu Miyanaga (2004), "VLSI architecture for HMM - based speech recognition system and its verification platform", *Proceeding of International Symposium on Communications and Information Technologies ISCIT*, pp.700-703.

[12] Choi Y, You K, Choi J, Sung W (2010), "A real - time FPGA - based speech recognizer with optimized DRAM access", *IEEE Transaction on Circuits System I, Regular papers,* **57(8)**, pp.2119-2131.

[13] Guangji He, Takanobu Sugahara, Yuki Miyamoto, Tsuyoshi Fujinaga, Hiroki Noguchi, Shintaro Izumi, Hiroshi Kawaguchi, Masahiko Yoshimoto (2012), "A 40 nm 144 mW VLSI processor for real-time 60 k word continuous speech recognition", *IEEE Transactions on circuits and systems - I, Regular papers*, **59(8)**, pp.1656-1666.

[14] Online: http://iheartech.hcmut.edu.vn/.