NHẬN DẠNG TIẾNG NÓI CHỮ SỐ VIỆT SỬ DỤNG BỘ CÔNG CỤ

Ngô Thị Thùy Vân[1]
Nguyễn Thị Thu Huyền[2]

Tóm tắt: *Nhận dạng tiếng nói của con người đã và đang thu hút sự quan tâm nghiên cứu của nhiều nhà khoa học trong và ngoài nước. Trong những năm gần đây, có nhiều nghiên cứunhận dạng tiếng nói cho tiếng Việt nhưng chủ yếu tập trung vào nhận dạng từ rời rạc, hay hệ thống nhận dạng liên tục với kích thước nhỏ. Bài báo trình bày hệ thống nhận dạng tiếng nói chữ số Việt sử dụng Hidden Markov Model (HMM) Tool Kit (HTK) để thực nghiệm đánh giá. Kết quả được kiểm nghiệm bằng các tiếng nói chữ số rời rạc, liên tục và có độ chính xác tương đối cao.*

Từ khóa: *nhận dạng tiếng nói, mô hình Markov ẩn, bộ công cụ nhận dạng HTK, chữ số Việt, hệ thống nhận dạng.*

1. Mở đầu

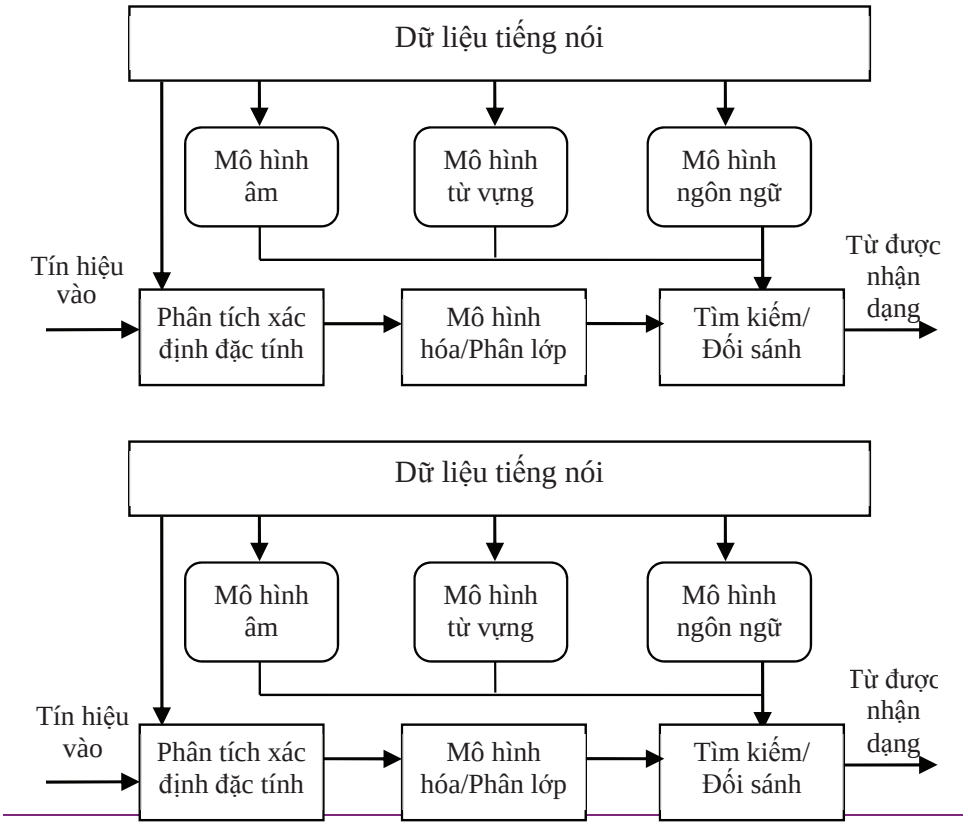
Ngay từ khi máy tính ra đời, con người đã mơ ước máy tính có thể nói chuyện với mình, chính vì vậy việc nghiên cứu các phương pháp và phát triển kỹ thuật nhận dạng tiếng nói đã và đang thu hút rất nhiều sự đầu tư và nghiên cứu của các nhà khoa học trên thế giới. Hiện nay trên thế giới, lĩnh vực nhận dạng tiếng nói (Speech recognition) đã đạt được nhiều tiến bộ vượt bậc, việc ra lệnh, điều khiển các thiết bị điện tử như ti vi, smartphone, máy tính bằng giọng nói không còn quá xa lạ với người dùng. Tuy nhiên nhận dạng ngôn ngữ tiếng Anh đã được nghiên cứu khá hoàn thiện, còn ngôn ngữ tiếng Việt do có tính chất phức tạp về mặt ngữ âm nên cần tập trung nghiên cứu nhiều hơn. Một hệ thống nhận dạng tiếng nói ở nước ta phải được xây dựng trên nền tảng của tiếng nói tiếng Việt.

2. Nội dung

2.1. Nhận dạng tiếng nói

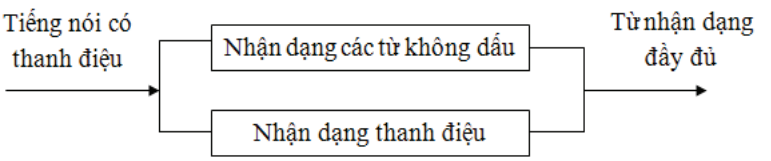
Nhận dạng tiếng nói là quá trình xử lý tiếng nói nhằm biến tín hiệu tiếng nói do người phát ra thành tín hiệu số, sau đó sử dụng một số giải thuật để đối chiếu giữa tín hiệu thu được tương ứng với dữ liệu tham chiếu nào trong bộ tham chiếu (từ điển nhận dạng). Về bản chất, đây là quá trình biến đổi tín hiệu âm thanh thu được của người nói qua Micro, đường dây điện thoại hoặc các thiết bị khác thành một chuỗi các từ một cách chính xác và hiệu quả. Kết quả của việc nhận dạng sau đó có thể được ứng dụng trong học tập, điều khiển, nhập dữ liệu…

1. Khoa Ngoại ngữ, Đại học Thái Nguyên
2. Khoa Ngoại ngữ, Đại học Thái Nguyên



2.2. Hệ thống nhận dạng tiếng nói tiếng Việt và mô hình Markov ẩn

Hệ thống nhận dạng tiếng nói tiếng Việt giống như hệ thống nhận dạng các ngôn ngữ có thanh điệu khác, bao gồm hai quá trình nhận dạng song song đó là: nhận dạng các từ không có thanh điệu và nhận dạng thanh điệu rồi tổng hợp để đưa ra quyết định.



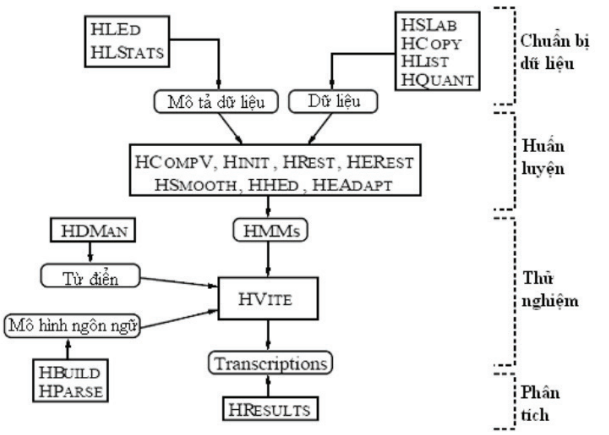
Hình 2: Cấu trúc hệ thống nhận dạng ngôn ngữ có thanh điệu

Mô hình Markov ẩn (Hidden Markov Model – HMM) là mô hình thống kê dùng để mô hình hóa các loại tín hiệu theo thời gian, với các tham số không biết trước. Thực tế nghiên cứu trong và ngoài nước cho thấy, trong lĩnh vực nhận dạng tiếng nói mô hình Markov ẩn cho kết quả nhận dạng tốt hơn các phương pháp khác.

2.3. Bộ công cụ nhận dạng tiếng nói tiếng Việt HMM Tool Kit (HTK)

HTK là một tập công cụ để xây dựng mô hình ngữ âm cho mục đích nhận dạng tiếng nói, được phát triển bởi Steve Young và các đồng nghiệp của ông ở trường Đại học

Cambridge [2] HTK tích hợp hầu hết các kĩ thuật về mô hình Markov ẩn, các kĩ thuật về xử lý và nhận dạng tiếng nói. Ngoài ra, nó còn cho phép ta xây dựng các mô hình ngôn ngữ, cú pháp văn phạm để quá trình nhận dạng tiếng nói đạt hiệu quả cao hơn.



Hình 3: Các công cụ và chức năng trong HTK.

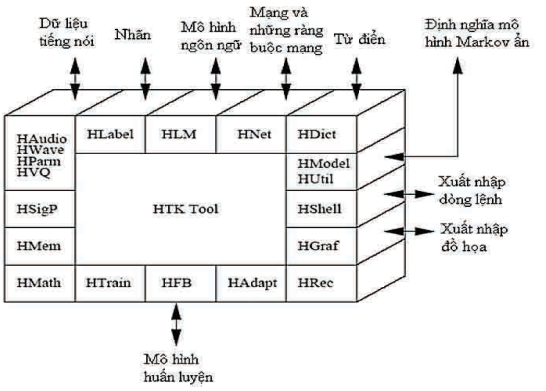
```
~h "hmm1"
<BEGINHMM>
<NUMSTATES> 5
<STATE> 2
<MEAN> 39
 -1.233344e+001 5.331151e+000 3.702762e+000 3.509391e+000 3.326789e+000
<VARIANCE> 39
 2.965368e+001 1.900993e+001 3.466729e+001 3.015772e+001 4.367296e+001
<GCONST> 8.402066e+001
<STATE> 3
<MEAN> 39
 -1.233344e+001 5.331151e+000 3.702762e+000 3.509391e+000 3.326789e+000
<VARIANCE> 39
 2.965368e+001 1.900993e+001 3.466729e+001 3.015772e+001 4.367296e+001
<GCONST> 8.402066e+001
<STATE> 4
<MEAN> 39
 -1.233344e+001 5.331151e+000 3.702762e+000 3.509391e+000 3.326789e+000
<VARIANCE> 39
 2.965368e+001 1.900993e+001 3.466729e+001 3.015772e+001 4.367296e+001
<GCONST> 8.402066e+001
<TRANSP> 5
 0.000000e+000 1.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000
 0.000000e+000 6.000000e-001 4.000000e-001 0.000000e+000 0.000000e+000
 0.000000e+000 0.000000e+000 6.000000e-001 4.000000e-001 0.000000e+000
 0.000000e+000 0.000000e+000 0.000000e+000 7.000000e-001 3.000000e-001
 0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000
<ENDHMM>
```

Hình 4: Cấu trúc tập tin mô hình Markov ẩn (HMM) được tạo bởi HTK

2.4. Một số Modul sử dụng trong quá trình xây dựng hệ thống nhận dạng tiếng nói tiếng việt trong bộ công cụ HTK

Modul này sẽ copy một hay nhiều file dữ liệu vào một file đầu ra được chỉ định trước, nó chuyển đổi dữ liệu sang dạng tham số. Là modul để rút trích đặc trưng trong

tập tin chứa tiếng nói. HCopy được sử dụng theo các bước sau:



Hình 5: Các Module và các chức năng trong HTK.HCopy

Bước 1: Tạo tập tin script dùng để chứa tên các tập tin cần chuyển đổi và tên các tập tin kết quả (chẳng hạn như đặt tên là chuyendoi.scp).

Mỗi dòng trong tập tin script chứa 2 đường dẫn:

Tên_tập_tin_cần_xử_lý Tên_tập_tin_kết_quả_tương_ứng

Ví dụ:

c:/YOU_2/wave/1.mfc

c:/YOU_2/wave/10.mfc

c:/YOU_2/wave/11.mfc

c:/YOU_2/wave/12.mfc

c:/YOU_2/wave/13.mfc

c:/YOU_2/wave/14.mfc

c:/YOU_2/wave/15.mfc

Bước 2: Tạo một tập tin cấu hình có tên HCopy.cfg chứa các thông tin như kiểu tập tin nguồn, kiểu tập tin đích, kích thước cửa sổ…

Ví dụ:

```
SOURCEKIND = WAVEFORM
# dạng hình sóng
#SOURCEFORMAT = WAV
#kiểu tập tin .wav
TARGETRATE = 100000.0
#tần số tập tin kết quả 100Hz
WINDOWSIZE = 250000.0
#kích thước cửa sổ 25ms
USEHAMMING = T
#dùng cửa sổ hamming = true
```

Bước 3: Thực thi lệnh để tạo ra tập tin đích, chẳng hạn dòng lệnh như sau:

HCopy –C HCopy.cfg –S chuyendoi.scp

HParse

Modul này dùng để tạo tập tin mô hình ngôn ngữ từ tập tin văn phạm. có thể sử dụng bằng cách sau:

Bước 1: Tạo tập tin văn phạm phù hợp với ngữ cảnh (chẳng hạn đặt tên là gram), Ví dụ tập tin đó như sau:

$digit= moojt | hai | ba | boosn | nawm | sasu | bary | tasm | chisn | khoong;

(<$digit>)

Bước 2: Thực thi lệnh HParse:

HParse gram wdnet

Kết thúc quá trình này ta thu được tập tin wdnet. Tập tin này được dùng để gán nhãn trong modul HVite.

HVite

HVite là modul được dùng để nhận dạng trong hệ thống nhận dạng tiếng nói bằng mô hình Markov ẩn, được sử dụng qua các bước như sau:

Bước 1: Tạo tập tin script chứa tất cả các tập tin cần nhận dạng ví dụ đặt tên là test.scp.

Bước 2: Chuẩn bị các tập tin như: từ điển *dict*, mạng ngôn ngữ *wdnet,* các mô hình HMM *hmmlist,* tập các mô hình HMM đã huấn luyện *hmmset.*

Bước 3: Thực thi lệnh HVite với các dòng lệnh về các tham số:

HVite –w wdnet –I recout.mlf –S test.scp –H hmmset dict hmmlist

Kết thúc lệnh tệp tin Master lable recout.mlf chứa mô tả các dữ liệu cần nhận dạng được tạo ra.

HCompV

HCompV đùng để khởi tạo mô hình Markov ẩn khi tập tin huấn luyện chưa được đánh nhãn. Các bước sử dụng HCompV như sau:

Bước 1: Tạo tập tin script chứa tất cả tập tin dùng huấn luyện (chẳng hạn đặt tên là train.scp).

Bước 2: Tạo mô hình HMM khởi đầu giả sử tên là proto.

Bước 3: Thực thi HCompV với lệnh sau:

HCompV –S train.scp proto

Kết thúc lệnh ta thu được mô hình HMM với tham số của tập tin dữ liệu.

HRest

Dùng để huấn luyện mô hình HMM, được thực hiện theo các bước sau:

Bước 1: Tạo tập tin script chứa tất cả các tập tin dùng để huấn luyện (chẳng hạn có tên là Train.scp).

Bước 2: Khởi tạo tập tin mô hình Hmm bằng HCompV như đã nói ở trên.

Bước 3: Thực thi lệnh HRest với dòng lệnh và tham số như:

HRest –S train.scp vidu

Kết thúc lệnh trên ta thu được mô hình HMM đã huấn luyện trong tập tin vidu.

2.5. Kết quả thử nghiệm xây dựng hệ thống nhận dạng tiếng nói chữ số tiếng Việt

Xây dựng cơ sở dữ liệu chữ số tiếng việt

Cơ sở dữ liệu trong thực nghiệm là cơ sở dữ liệu tự xây dựng với 1000 mẫu trong tập huấn luyện và 100 mẫu trong tập test. Để thuận tiện cho việc gán nhãn, dữ liệu được thu theo các câu đã được phát sinh ngẫu nhiên (dạng văn bản) nhờ công cụ HTK.

Các bước xây dựng dữ liệu như sau:

- Sinh tổ hợp ngẫu nhiên 1100 câu văn bản có kích thước từ 1 đến 10 từ bộ 10 chữ số từ 0 đến 9.

- Tách 1100 câu thành 22 bộ câu nhỏ, mỗi bộ 50 câu

- Tiến hành thu âm với 22 người nói tương ứng với 22 bộ câu (11 nữ, 11 nam, độ tuổi từ 20 đến 24)

- Lấy ngẫu nghiên 100 câu trong 1100 câu đã thu âm làm bộ test, còn lại 1000 câu làm bộ traing.

Bảng phiên âm 10 chữ số tiếng Việt

Cách phiên âm có vài trò rất quan trọng đảm bảo chất lượng của hệ thống nhận dạng. Hệ thống sử dụng bảng phiên âm âm vị cho hệ thống nhận dạng 10 chữ số tiếng Việt như sau:

Bảng 1: Bảng phiên âm 10 chữ số tiếng Việt

Chữ số	**Phiên âm chính tả**	**Phiên âm âm vị**
0	Khoong	/Kh/ /oo/ /ng/
1	Moojt	/m/ /ooj/ / t/
2	Hai	/h/ / a // i/
3	Ba	/b/ / a /
4	Boosn	/b//oos//n/
5	Nawm	/n//aw//m
6	Sasu	/s//as//u/
7	Bary	/b//ar//y/
8	Tasm	/t//as//m/
9	Chisn	/ch//is//n/

Phương pháp xây dựng hệ thống nhận dạng chữ số tiếng việt

Phương pháp xây dựng hệ thống nhận dạng 10 chữ số phát âm tiếng Việt được tiến hành theo các bước:

- Từ điển: được xây dựng dựa trên bảng phiên âm âm vị bao gồm 2 loại từ điển cho 2 thực nghiệm khác nhau để đánh giá độ chính xác và chọn ra bộ từ điển thích hợp. Một từ điển không chèn các sp (short pause) và một từ điển có chèn thêm các sp.
- Sử dụng bộ công cụ HTK để xử lý rút trích đặc trưng của dữ liệu huấn luyện và dữ liệu Test.
- Xây dựng mô hình Markov ẩn với hàm phát xạ quan sát là hàm mật độ Gauss.
- Số lượng trạng thái trong mô hình Markov ẩn là 5 trạng thái, trong đó có 1 trạng thái khởi đầu và 1 trạng thái kết thúc không có phát xạ quan sát.
- Sử dụng vector đặc tính phổ gồm hệ số MFCC, giá trị năng lượng cùng các delta, delta- delta của các giá trị này tạo thành tập 39 đặc tính phổ tương ứng với mỗi khung tín hiệu 10ms.
- Tiến hành buộc các âm vị không có đủ bộ dữ liệu huấn luyện theo phương pháp dùng cây (tree- based). Các âm vị trong tập dữ liệu kiểm tra mà không có mặt trong dữ liệu huấn luyện sẽ được tổng hợp từ các âm vị đã được huấn luyện giống nhất.
- Thử nghiệm trộn nhiều hàm Gauss và mix các trạng thái.

Kết quả thực nghiệm

Thử nghiệm với từ điển có chèn short pause và không chèn short pause

Trong khi nói, giữa những câu những từ sẽ có khoảng ngừng nghỉ khác nhau, do vậy để máy có thể phân biệt được điều này là rất khó khăn. Để kiểm tra sự ảnh hưởng của yếu tố ngừng nghỉ giữa các câu, các từ tới độ chính xác của hệ thống, nhóm tác giả đã tiến hành thử nghiệm trên 2 bộ từ điển phiên âm 10 chữ số tiếng Việt. Một bộ từ điển phiên âm không chèn thêm các âm quy định là khoảng nghỉ và một bộ từ điển có chèn thêm các sp quy định là những khoảng nghỉ giữa các từ.

ba	**b a**
bay	**b ar y**
bon	**b oos n**
chin	**ch is n**
hai	**h a i**
khong	**kh oo ng**
mot	**m ooj t**
nam	**n aw m**
sau	**s as u**
tam	**t as m**
silence	**sil**

+Từ điển không chèn thêm các sp.

```
#dict
ba b a
ba b a sp
bary b ar y
bary b ar y sp
boosn b oos n
boosn b oos n sp
chisn ch is n
chisn ch is n sp
hai h a i
hai h a i sp
khoong kh oo ng
khoong kh oo ng sp
moojt m ooj t
moojt m ooj t sp
nawm n aw m
nawm n aw m sp
sasu s as u
sasu s as u sp
tasm t as m
tasm t as m sp
silence sil
```

+Từ điển có chèn thêm các sp.

Kết quả của thử nghiệm độ chính xác của hệ thống nhận dạng theo 2 bộ từ điển ở trên được cho trong bảng sau:

Bảng 2. Kết quả thử nghiệm hệ thống nhận dạng với bộ từ điển có chèn sp và không chèn sp

Hệ thống nhận dạng	Mức câu	Mức từ
Bộ từ điển không chèn sp	56%	90%
Bộ từ điển có chèn sp	70%	90%

Như vậy, với bộ từ điển có chèn thêm các sp, độ chính xác mức câu sẽ tăng lên.

3. Kết luận

Bài báo đã trình bày hệ thống nhận dạng tiếng nói dựa trên bộ công cụ HTK. Mô hình thử nghiệm nhận dạng tiếng nói chữ số Việt được xây dựng dựa trên bộ công cụ HTK đã đáp ứng được mục tiêu của nhóm tác giả. Chúng tôi đã thử nghiệm với 1000 câu làm dữ liệu huấn luyện và 100 câu làm dữ liệu test, kết quả cho độ chính xác có thể chấp nhận được (70% đối với mức câu) và (90% đối với mức từ).

TÀI LIỆU THAM KHẢO

[1] Nguyễn Văn Giáp, Trần Việt Hồng (2013), "Kỹ thuật nhận dạng tiếng nói và ứng dụng trong điều khiển", *Tạp chí phát triển khoa học công nghệ* - ĐHQG Hồ Chí Minh.

[2] Nguyễn Thị Thu Huyền (2018), "Mô hình Markov ẩn và ứng dụng xây dựng hệ thống nhận dạng tiếng nói*", Luận văn Thạc sĩ chuyên ngành CNTT trường Đại học Công nghệ TT&TT*- Đại học Thái Nguyên, 44-56.

[3] Nguyễn Duy Phương (2007), "Mô hình Markov ẩn và ứng dụng trong nhận dạng tiếng nói", *Luận văn Thạc sĩ chuyên ngành CNTT trường Đại học Công nghệ* - ĐHQG Hà Nội, 17-22.

[4] Nguyen Hong Quang, Trinh Van Loan, Le The Dat (2010), "Automatic Speech Recognition for Vietnamese using HTK system", *IEEE-RiVF 2010*, Hanoi, November, 103-106.

Title: VIETNAMESE NUMERAL RECOGNITION BY USING HTK

NGO THI THUY VAN
NGUYEN THI THU HUYEN
School of Foreign Languages – TNU

Abstract*: Human speech recognition has been being studied both at home and abroad. Several studies on Vietnamese speech recognition have recently been carried out but they mainly focus on discrete word recognition or small-scale uninterrupted recognition systems. The paper will present a system of Vietnamese numeral recognition using Hidden Markov Model (HMM) Toolkit (HTK) for empirical assessment. The results are tested via discrete or continuous numerals with relatively high accuracy.*

Keywords*: Speech recognition, Hidden Markov Model, HTK, Vietnamese numerals, recognition system.*