NHẬN DẠNG TIẾNG NÓI CHỮ SỐ VIỆT ÁP DỤNG TRONG HỆ THỐNG NHẬP ĐIỂM

ThS. Thái Duy Quý[1]

TÓM TẮT

Nhận dạng tiếng nói của con người đã và đang thu hút sự quan tâm nghiên cứu của nhiều nhà khoa học khi mà công nghệ tự động hóa ngày càng có nhiều ứng dụng trong thực tiễn cuộc sống. Nghiên cứu nhận dạng tiếng nói Việt cũng được quan tâm nghiên cứu nhiều trong những năm gần đây, tuy vậy cho đến nay các kết quả vẫn chưa thỏa mãn những bài toán đặt ra từ thực tế cuộc sống do tính chất phức tạp về ngữ âm của tiếng Việt. Bài báo trình bày bài toán tìm đặc trưng, huấn luyện và nhận dạng tiếng nói Việt, ứng dụng trong hệ thống nhập điểm. Các kết quả được kiểm nghiệm bằng các tiếng nói số rời rạc và tổ hợp ngắn, đồng thời tích hợp trong chương trình nhập điểm cho hệ thống hiện hành.

Từ khóa: *Nhận dạng tiếng nói Việt, nhận dạng chữ số, speech recognition, HMM, MFCC*

1. Đặt vấn đề

1.1. Giới thiệu

Trong giao tiếp giữa người với người, tiếng nói là phương pháp trao đổi thông tin tự nhiên và hiệu quả nhất. Mục tiêu của các kỹ thuật nhận dạng tiếng nói theo nghĩa rộng là tạo ra những máy có khả năng nhận biết được thông tin tiếng nói và hành động theo tiếng nói đó. Nhận dạng tiếng nói là một phần của quá trình tìm kiếm thông tin để máy có thể "nghe", "hiểu" và "hành động" theo thông tin đồng thời "nói lại" để hoàn tất việc trao đổi thông tin.

Cho đến nay, vấn đề giao tiếp giữa con người và máy tính tuy đã được cải thiện nhiều nhưng chủ yếu vẫn còn khá thủ công thông qua các thiết bị nhập, xuất. Giao tiếp với thiết bị máy bằng tiếng nói sẽ là phương thức giao tiếp văn minh và tự nhiên nhất. Dấu ấn giao tiếp người - máy sẽ mất đi mà thay vào đó là cảm nhận của sự giao tiếp giữa người với người, nếu hoàn thiện thì đây sẽ là một phương thức giao tiếp tiện lợi và hiệu quả trong công việc [4]. Mặc dù nhận dạng ngôn ngữ tiếng Anh đã được nghiên cứu khá hoàn thiện nhưng do có sự khác biệt về ngữ âm, ngữ nghĩa với tiếng Việt nên khó có thể áp dụng các chương trình nhận dạng khác hiện hành để nhận dạng tiếng Việt. Một hệ thống nhận dạng tiếng nói ở nước ta phải được xây dựng trên nền tảng của tiếng nói tiếng Việt [5].

1.2. Tổng quan tình hình nghiên cứu

Các kỹ thuật nhận dạng tiếng nói trên thế giới đã có từ thập niên 60 và đã đạt được nhiều thành tựu đáng kể [1]. Các hệ thống nhận dạng giọng nói tiếng Anh đã được áp dụng trong nhiều lĩnh

[1]Trường Đại học Đà Lạt

vực như trong xử lý văn bản bằng tiếng nói, tự động hóa trong phân xưởng, các hệ thống an ninh, dịch thuật, hệ thống trả lời tự động, robot thông minh,…

Tại Việt Nam, do còn tùy thuộc vào điều kiện nghiên cứu và sự phức tạp của ngữ âm tiếng Việt nên các nghiên cứu về hệ thống nhận dạng giọng nói tiếng Việt vẫn còn nhiều hạn chế và đến nay chưa có hệ thống nào hoàn chỉnh [4]. Mặc dù vậy, hiện nay cũng có nhiều công trình nghiên cứu của các nhà khoa học, có thể kể đến PGS. TS. Lương Chi Mai (Viện Công nghệ Thông tin Hà Nội), PGS.TS. Vũ Hải Quân (Đại học Khoa học Tự nhiên TP. Hồ Chí Minh)… mang lại nhiều những thành công trên lý thuyết và ứng dụng. Trong những sản phẩm nổi bật, có thành tựu của sản phẩm VSpeech của nhóm BK02 [9], tương tác giọng nói với chữ viết để điều kiển một số chức năng cơ bản trên máy tính. Một số sản phẩm của các công ty cũng đã tích hợp các chức năng tìm đường đi, cây xăng, ATM,… trên các hệ thống di động.

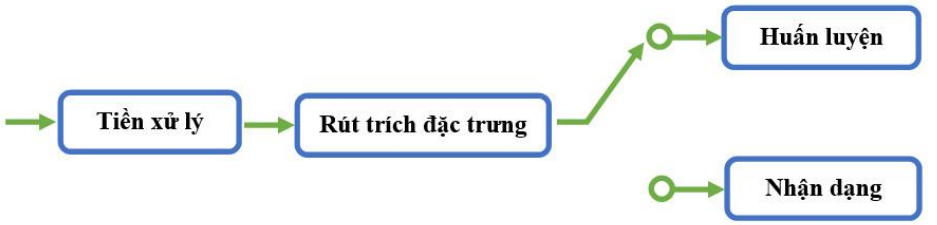
Mặc dù có nhiều nghiên cứu và sản phẩm ứng dụng thực tế nhưng trong các sản phẩm về nhận dạng tiếng nói vẫn chưa có sản phẩm nào đáp ứng cho công việc nhập điểm, một công việc thường xuyên trong nhà trường.

1.3. Mục tiêu của đề tài

Đề tài nghiên cứu thử nghiệm hướng nhận dạng tiếng nói Việt dựa trên việc trích đặc trưng của tiếng nói bằng phương pháp MFCC (Mel Frequency Ceptrums Coefficients), và nhận dạng bằng mô hình HMM (Hidden Markov Models). Đồng thời một chương trình nhận dạng bằng tiếng nói Việt được xây dựng với bộ từ vựng nhỏ là các tiếng nói số, dùng trong hệ thống nhập điểm. Chương trình được xây dựng bằng ngôn ngữ C# trên nền .Net dựa vào một số thư viện. Các bước minh họa sử dụng một số hàm trong ngôn ngữ Matlab.

2. Hệ thống nhận dạng tiếng nói Việt

Về mặt tổng quát, một hệ thống nhận dạng thường bao gồm hai phần chính là *huấn luyện* (training) và *nhận dạng* (recognition) được thể hiện như trong hình 1. Trong đó "*Rút trích đặc trưng*" là quá trình đưa ra được những đặc trưng thích hợp cho nhận dạng. "*Huấn luyện*" là quá trình hệ thống "học" và "lưu trữ" những mẫu chuẩn được cung cấp, từ đó hình thành bộ từ vựng của hệ thống. Và quá trình "*nhận dạng*" là quyết định xem mẫu nào được đưa vào căn cứ vào bộ từ vựng đã được huấn luyện.



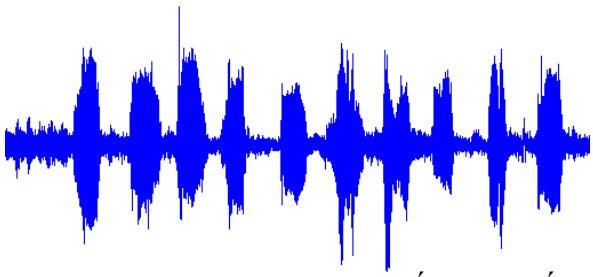
Hình 1: *Tổng quan một hệ thống nhận dạng*

Một hệ thống nhận dạng tiếng nói cũng theo quy tắc các bước của một hệ nhận dạng tổng quát. Tín hiệu thu vào là các âm thanh nói từ micro, đặc trưng của âm thanh thường là tiếng và âm vị của ngôn ngữ và quá trình huấn luyện dựa trên các tập tin âm thanh đã thu vào từ trước.

3. Tiền xử lý

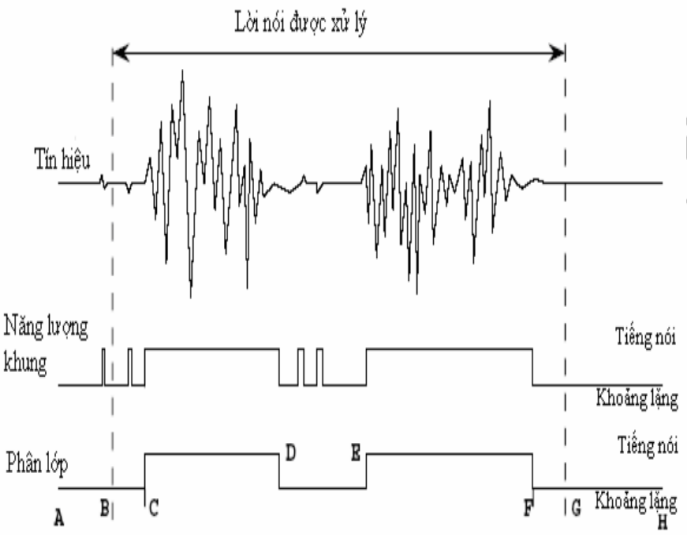
Tiếng nói sau khi được thu từ micro sẽ được lấy mẫu tín hiệu, một mẫu tín hiệu thường được biểu diễn dưới dạng sóng. Hình 2 mô tả sóng âm của các số từ một đến mười. Đối với tín hiệu âm thanh, mẫu sẽ được lấy theo một chu kỳ thời gian, công thức lấy mẫu được xác định bởi công thức 1:

$$X_s(t) = \sum_{n=-\infty}^{\infty} x(t)\delta(t - nT) \quad \textbf{(1)}$$

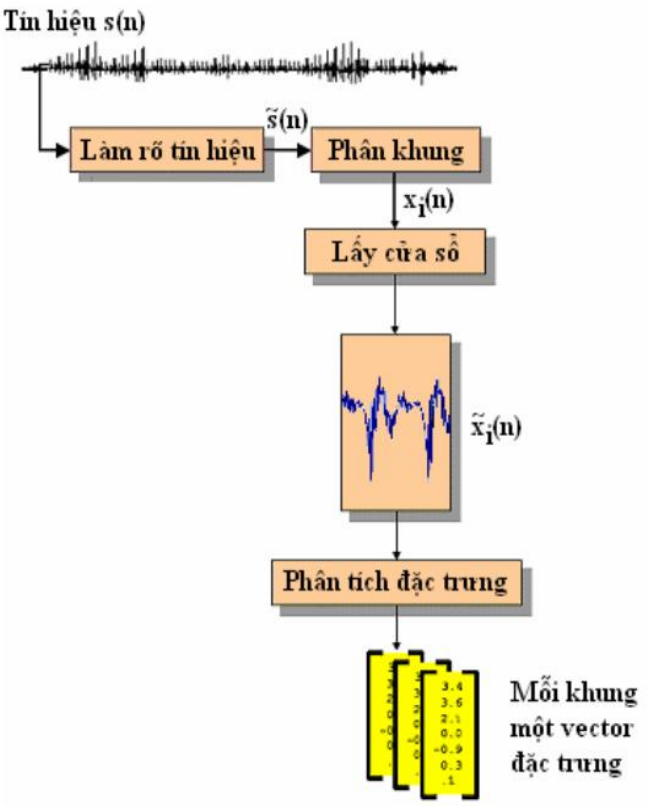


Hình 2: *Mô hình sóng âm các số từ một đến mười*

Tìn hiệu sau khi lấy xong sẽ thông qua một bộ lọc tín hiệu. Bộ lọc tín hiệu có thể bao gồm bộ khử nhiễu, bộ khôi phục tín hiệu biến dạng, bộ dò tìm điểm cuối để xác định đâu là tiếng ồn, đâu là tiếng nói và khoảng lặng giữa hai tiếng nói. Một ví dụ về phương pháp dò tìm điểm cuối được mô tả trong hình 3.



Hình 3: *Một ví dụ về dò tìm điểm cuối trong sóng âm*



Hình 4: *Các quy trình trong rút trích đặc trưng MFCC*

4. Rút trích đặc trưng

Sau quá trình tiền xử lý đã có được các mẫu tiếng nói khử nhiễu. Phần trích đặc trưng sẽ đưa ra được vector đặc trưng cho mô hình cần nhận dạng. Có nhiều phương pháp trích đặc

trưng khác nhau như Wavelets, LPC, MFCC… Chúng tôi chọn phương pháp trích đặc trưng MFCC (Thang tần số Mel) do tốc độ tính toán cao, độ tin cậy lớn và đã được sử dụng rất hiệu quả trong các chương trình nhận dạng tiếng nói trên thế giới [4].

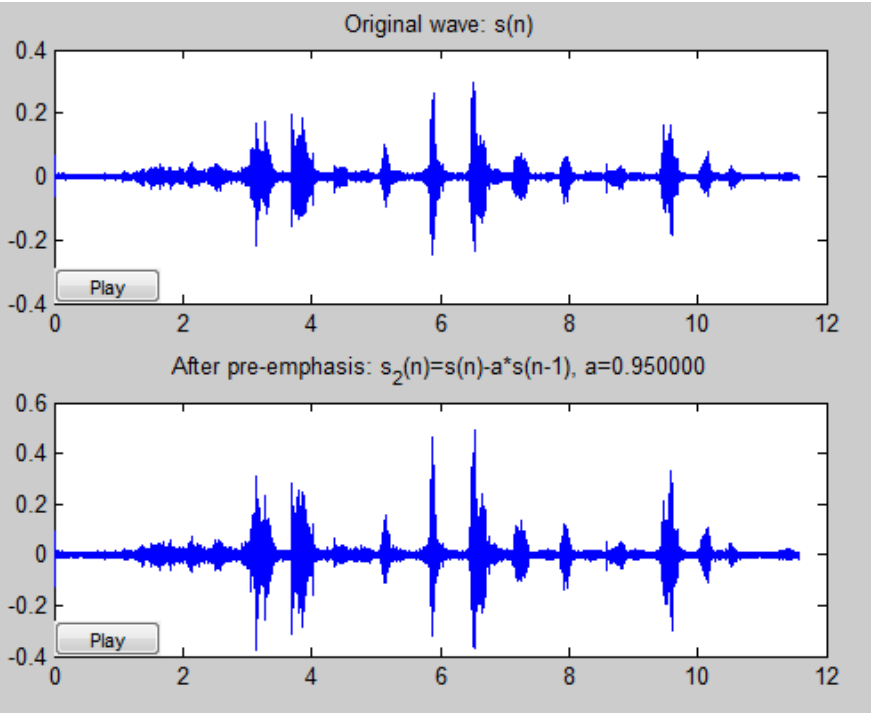
Phương pháp rút trích đặc trưng MFCC được mô tả như trong hình 4. Trong mô hình này ta có bốn bước để rút trích đặc trưng như: làm rõ tín hiệu, phân khung, lấy cửa sổ và phân tích đặc trưng. Chi tiết các bước được trình bày theo các mục sau đây.

4.1. Làm rõ tín hiệu

Bước này mục đích chính là làm tăng tín hiệu và nổi rõ các đặc trưng của tín hiệu giúp nâng cao mức độ nhạy cảm trong các bước sau [3].

Bộ làm rõ tín hiệu có phương trình sai phân như sau:

$$\tilde{s} = s(n) - as(n-1) \quad \textbf{(2)}$$

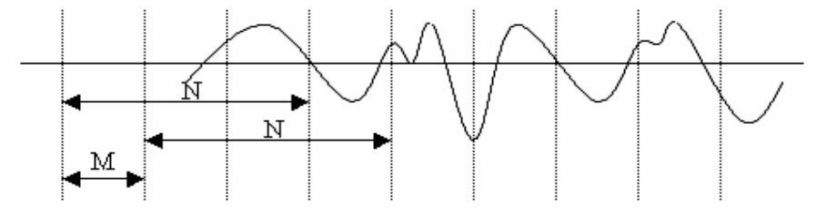


Hình 5*. Mô hình bước sóng trước và sau khi làm rõ tín hiệu*

4.2. Phân khung

Trong bước này, $\tilde{s}$ được chia thành các khung, mỗi khung gồm N mẫu, khoảng cách giữa các khung là *M* mẫu. Hình 5 minh họa cách phân thành các khung với $M = \frac{1}{3}N$. Nếu ta ký hiệu khung thứ *i* là $x_i(n)$ và có tất cả *L* khung trong tín hiệu tiếng nói thì:

$$x_i(n) = s(M.i + n) \text{ với n =0,1,…,N-1; i=0,1,…,L-1} \quad \textbf{(3)}$$



Hình 6: *Âm tiếng nói được phân đoạn thành các khung*

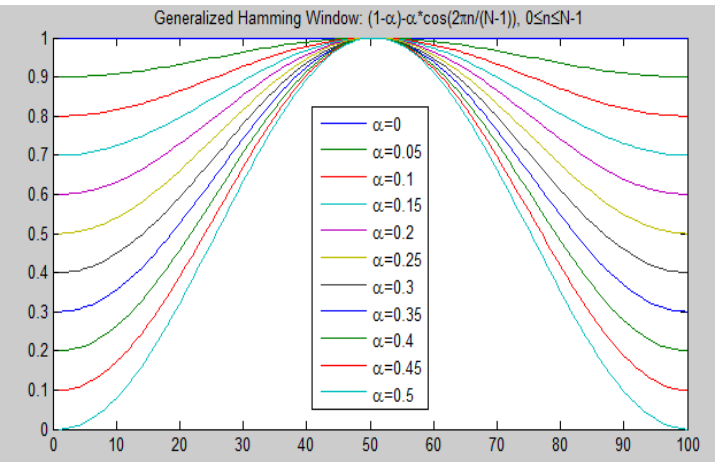
4.3. Lấy cửa sổ

Bước tiếp theo trong xử lý là lấy cửa sổ tín hiệu ứng với mỗi khung để giảm thiểu gián đoạn tín hiệu ở đầu và cuối mỗi khung. Dãy tín hiệu con được lấy ra từ một tín hiệu dài hơn hoặc dài vô hạn x(n) gọi là một cửa sổ tín hiệu. Quá trình quan sát tín hiệu x(n) bằng một đoạn x(N(n)) trong khoảng $n_0 \ldots (n_0 + N - 1)$ tương đương với việc nhân x(n) với một hàm cửa sổ $w(n-n_0)$ như sau:

$$x_N(n) = x(n).w(n-n_0) = \begin{cases} x(n) & n_0 \le n \le n_0+N-1 \\ 0 & (n<n_0) \vee (n>n_0+N-1) \end{cases} \quad \textbf{(4)}$$

Trong nhận dạng tiếng nói, hàm cửa sổ thường hay được dùng nhất là Hamming, có dạng như công thức (5). Tín hiệu của cửa sổ Hamming được biểu diễn trong hình 7.

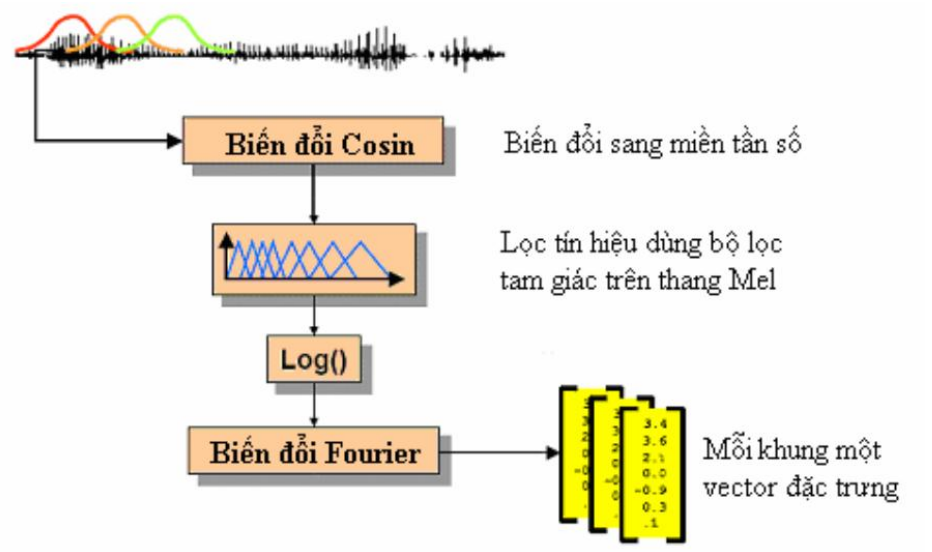
$$x_N(n) = x(n).w(n-n_0) = \begin{cases} 0.54+0.46\cos(2\pi n/N) & |n| \le N/2 \\ 0 & |n| > N/2 \end{cases} \quad \textbf{(5)}$$



Hình 7: *Mô hình sóng cửa sổ Hamming*

4.4. Trích chọn đặc trưng

Bước cuối cùng trong trích chọn đặc trưng MFCC bao gồm thực hiện biến đổi Furier ngược dựa trên độ lớn logarit của ngõ ra của bộ lọc.



Hình 8: *Các bước trích chọn đặc trưng MFCC*

Sau khi tín hiệu tiếng nói được trích đặc trưng thì mỗi từ được đặc trưng bởi một ma trận hệ số thực. Dựa theo [4], chúng tôi định nghĩa một vector đặc trưng bao gồm 10 thành phần như sau:

$$y_t = [f_t, f_t\text{-}f_{t+1}, e_t\text{-}e_{max}, d, f_{max}, f_{min}, f_{max}\text{-}f_{st}, f_{min}\text{-}f_{ed}, f_{min}\text{-}f_{st}, f_{max}\text{-}f_{ed}] \quad \textbf{(6)}$$

trong đó:

- f_t là tần số cơ bản tại khung tín hiệu t
- f_{t+1} là tần số cơ bản tại khung tín hiệu $t+1$
- e_t là năng lượng tại khung tín hiệu t
- e_{max} là năng lượng cực đại trong phần hữu thanh (không phải nhiễu)
- d là số khung phần hữu thanh
- f_{max} là tần số cơ bản cực đại trong vùng hữu thanh
- f_{min} là tần số cơ bản cực tiểu trong vùng hữu thanh
- f_{st} là tần số cơ bản ở khung đầu tiên trong vùng hữu thanh
- f_{ed} là tần số cơ bản ở khung cuối cùng trong vùng hữu thanh

Do mô hình HMM rời rạc được ứng dụng để nhận dạng nên những vector đặc trưng này phải được ước lượng vector thành một chỉ số codebook rời rạc. Phương pháp được sử dụng để ước lượng vector là phương pháp K-means.

5. Huấn luyện cho mô hình

Sau khi thực hiện xong phần rút trích đặc trưng, kết quả là có một cơ sở dữ liệu các vector đặc trưng tương ứng với từng từ. Phần huấn luyện sử dụng mô hình Markov ẩn với dữ liệu huấn luyện là các vector đặc trưng có được từ phần trước. Ứng với mỗi từ cần nhận dạng thì một cơ sở dữ liệu các đặc trưng

từ các lần đọc khác nhau. Sau đó sẽ ước lượng các thông số của mô hình $\lambda = (A, B, \pi)$ để xác suất $P(O|\lambda)$ đạt cực đại, tương ứng với mỗi từ là một λ xác định. Để nhận dạng một từ thì chỉ việc tính xác suất chuỗi quan sát của từ đó ứng với các λ đã được huấn luyện và chọn mẫu nào có xác suất lớn nhất.

6. Thực nghiệm hệ thống nhập điểm dựa vào tiếng nói

Để thử nghiệm hệ thống nhận dạng, chúng tôi sử dụng bộ công cụ Sphinx [7]. Đây là bộ công cụ mã nguồn mở, tích hợp cả chức năng huấn luyện và nhận dạng trên hai mô hình là ngôn ngữ và mô hình ngữ âm. Bộ công cụ này cũng tiến hành nhận dạng tiếng nói dựa theo các bước như đã nêu ở trên.

Bộ dữ liệu dùng cho nhận dạng và huấn luyện là các tập tin dạng *.wav*, được thu âm từ 100 người. Do ứng cụng của chúng tôi là nhận dạng dựa trên chữ số nên chỉ xây dựng mô hình từ vựng với các chữ số như: *không, một, hai, ba, bốn, năm, sáu, bảy, tám, chín, mười, phẩy, lên, xuống*.

Mô hình ngôn ngữ được sử dụng bảng mã VIQR minh họa như trong bảng 1:

Bảng 1: *Một số từ vựng, chữ số dùng trong huấn luyện*

Mô hình từ vựng	Ý nghĩa	Ký tự cần nhận dạng	Mô hình từ vựng	Ý nghĩa	Ký tự cần nhận dạng
KHO^NG	Không	0	TA'M	Tám	8
MO^T	Một	1	CHI'N	Chín	9
HAI	Hai	2	MU+O+`I	Mười	10
BA	Ba	3	PHA^?Y	Phẩy	,
BO^'N	Bốn	4	LE^N	Lên	Up
NA(M	Năm	5	XUO^'NG	Xuống	Down
SA'U	Sáu	6	VA(NG THI	Vắng thi	VT
BA?Y	Bảy	7			

Kết quả thử nghiệm được thể hiện trong bảng 2. Bảng này cho thấy kết quả nhận dạng: Có 12/15 chữ số được nhận dạng đúng (86%), có 3 chữ số bị nhận dạng nhầm lẫn, trung bình kết quả nhận dạng là 93.3%.

Bảng 2: *Kết quả thực nghiệm*

Số	Kết quả nhận dạng		
0	100%	8	100%
1	70%	9	100%
2	100%	10	100%
3	50%	Phẩy	100%

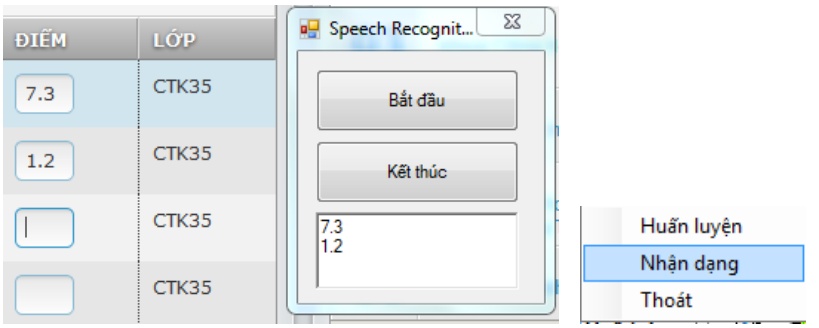
4	100%	Lên	100%
5	100%	Xuống	100%
6	80%	Vắng thi	100%
7	100%		
Trung bình: 93.3%			

Bảng 3 mô tả kết quả nhầm lẫn của các cặp chữ số: *Một - mười, ba - bảy, sáu - bốn.*

Bảng 3: *Mức độ nhầm lẫn ngữ âm của một số từ vựng*

Từ vựng	Từ bị nhận dạng nhầm	Tỉ lệ
Một	Mười	30%
Ba	Bảy	50%
Sáu	Bốn	20%

Trong ứng dụng nhập điểm, chúng tôi xây dựng chương trình dựa trên một số bộ thư viện như Sphinx [7], VSpeech.dll [9] và System.speech [10]. Dữ liệu đưa vào là các số từ 1 đến 9 và các yêu cầu như phẩy, lên, xuống, vắng thi. Để thuận tiện cho việc nhận dạng các điểm lẻ, chúng tôi cũng đưa vào các bộ số lẻ như: một phẩy một, một phẩy hai… Chương trình được viết bằng ngôn ngữ C# trên nền .Net (hình 9), kết quả nhập điểm với độ chính xác 93.3%. Do dữ liệu huấn luyện còn ít, khi nhận dạng, chúng tôi cũng thiết lập thêm thêm những gợi ý để nâng cao mức độ nhận dạng cho hệ thống nhập điểm.



Hình 9: *Chương trình nhập điểm bằng giọng nói*

7. Kết luận

Mô hình thử nghiệm nhận dạng tiếng nói chữ số trong tiếng Việt theo hướng kết hợp MFCC và HMM tuy còn nhiều hạn chế nhưng đã đáp ứng được mục tiêu của đề tài. Chương trình thử nghiệm được sử dụng để nhập các hệ thống điểm lẻ với bộ từ vựng nhỏ cho độ chính xác có thể chấp nhận được (trên 90%). Nếu điều kiện cho phép, nhóm tác giả sẽ tối ưu hóa chương trình nhận dạng, đưa them nhiều bộ dữ liệu huấn luyện để đạt được kết quả cao hơn và tăng tốc độ xử lý.

TÀI LIỆU THAM KHẢO

1. Thái Hùng Văn, Đỗ Xuân Đạt, Võ Văn Tuấn, (2003), *Nghiên cứu các đặc trưng của tiếng Việt áp dụng vào nhận dạng tiếng nói tiếng Việt* (Luận văn Đại học), Đại học Khoa học Tự nhiên TP. Hồ Chí Minh

2. Nguyễn Văn Giáp, Trần Việt Hùng (2006), *Kỹ thuật nhận dạng tiếng nói ứng dụng trong điều khiển*

3. Nguyễn Hồng Quang (2004), *Nhận dạng tiếng nói Việt, tìm hiểu và ứng dụng*, Trường Đại học Khoa học Tự nhiên

4. http://bk02.sourceforge.net/vspeechsdk/vietnamese/

5. Phan Nguyễn Phục Quốc, Hà Thúc Phùng (2009), *Hệ thống nhận dạng tiếng nói* (Luận văn Đại học), Đại học Bách khoa TP. Hồ Chí Minh

6. CMUSphinx Wiki: http://cmusphinx.sourceforge.net/wiki/

7. http://msdn.microsoft.com

8. Cao Xuân Hạo (1998), *Tiếng Việt - mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa*, Nhà xuất bản Giáo dục

9. Xuedong Huang, Alex Acero, Hsiao-wuen Hon (2001), *Spoken language Processing*, Carnegie Mellon University

10. Mikael Nilson, Marcus Ejnarson (2002), *Speech recognition using Hidden Markov Model performance evaluation in noisy enviroment*, ebook

SPEECH RECOGNITION VIETNAMESE IN APPLYING TO INPUTTING SCORES

ABSTRACT

Speech recognition of the human voice has attracted the attention of many scientists while automation technology has been more and more applied to real life. Researching Vietnamese speech recognitions has also been concerned in recent years, but so far the results have not yet satisfied the problems posed by real life complex because of the nature of phonological Vietnamese. This paper presents the problem of finding features, training and applying Vietnamese speech recognition to inputting score. The results are tested by the discrete and short voice digital while the application was built for the current system.

Keywords**:** *Speech recognition, HMM, MFCC*