ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Quang Trung

NHẬN DẠNG VÀ SẢN XUẤT TIẾNG NÓI BẰNG MẠNG NƠRON TỰ TỔ CHỨC

Chuyên ngành:Khoa học máy tính

Mã số: 62.48.01.01

TÓM TẮT LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

Hà Nội - 2017

Công trình được hoàn thành tại:Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

Người hướng dẫn khoa học:PGS. TS. Bùi Thế Duy

Phản biện 1:

.....

Phản biện 2:.....

.....

Phản biện 3:.....

.....

Luận án sẽ được bảo vệ trước Hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ họp tại: Đại học Công nghệ, Đại học Quốc Gia Hà Nội

Vào hồigiờngàythángnăm

Có thể tìm hiểu luận án tại:

- Thư viện Quốc gia Việt Nam
- Trung tâm Thông tin - Thư viện, Đại học Quốc gia Hà Nội

PHẦN MỞ ĐẦU

1. Tính cấp thiết của luận án

Ngày nay, với sự bùng nổ của xã hội thông tin, con người không còn chỉ có nhu cầu giao tiếp với nhau nữa mà còn cần giao tiếp với những thiết bị điện tử. Hình thức giao tiếp người - máy thông qua ngôn ngữ tự nhiên sẽ đem lại nhiều ứng dụng, góp phần giải phóng sức lao động của con người. Chính vì vậy, việc làm cho máy tính có thể nhận thức được tiếng nói (hiểu tiếng nói) có tầm quan trọng đặc biệt liên quan đến quá trình phát triển của văn minh nhân loại. Nhận thức tiếng nói nói riêng đã được nghiên cứu từ đầu những năm 1950 (Sumby & Pollack, 1954) (Cooper, 1952) (Broadbent D. &., 1957). Tuy nhiên, những nghiên cứu về nhận thức tiếng nói ở thời kỳ đầu chỉ tập chung vào một số bài toán cụ thể như bài toán tách nguồn tiếng nói, bài toán nhận dạng tiếng nói, bài toàn nhận dạng hay xác thực người nói.

Gần đây, nghiên cứu về nhận thức tiếng nói đã đạt được nhiều thành tựu to lớn. Tuy nhiên, các nghiên cứu về nhận thức tiếng nói chỉ xây dựng các hệ thống có thể hiểu ở mức độ phân biệt được tiếng nói ở một khía cạnh nào đó. Các nghiên cứu này chỉ tập trung mô phỏng hoạt động nhận thức tiếng nói xảy ra ở vùng vỏ não thính giác đặc biệt là vùng vỏ não thính giác sơ cấpvà vùng vỏ não thính giác thứ cấp. Rất ít nghiên cứu đặt bài toán nhận thức tiếng nói trong mối quan hệ với nhận thức của các hệ giác quan khác là quá trình nhận thức xảy ra ở vùng vỏ não liên kết đa giác quan.

Các nghiên cứu về vai trò của vùng vỏ não liên kết đa giác quan trong nhận thức tiếng nói là ít được nghiên cứu, trong khi đó, quá trình nhận thức tiếng nói ở con người là một quá trình phức tạp, với sự tham gia của tất cả các giác quan, các vùng vỏ não, đặc biệt là

vùng vỏ não liên kết, vùng chiếm tỷ lệ rất cao trong vỏ não con người.

Xuất phát từ những lý do trên, việc lựa chọn đề tài nghiên cứu hướng tiếp cận mới cho bài toán nhận thức tiếng nói trong đó đề xuất mô hình mô phỏng quá trình nhận thức tiếng nói thông qua việc học mối quan hệ hay liên kết giữa vùng vỏ não thính giác với các vùng vỏ não khác đặc biệt là liên kết giữa vùng vỏ não thính giác với vùng vỏ não thị giác.

Kết quả đề tài này có thể ứng dụng trong việc nhận dạng tiếng nói tác từ, các câu rời rạc, nhận dạng mệnh lệnh trong điều khiển học hay trong ứng dụng trong giao tiếp người máy, hay ứng dụng trong tìm kiếm video dựa trên đoạn một hội thoại ngắn.

2. Mục tiêu của luận án

Mục tiêu chính của luận án là xây dựng mô hình nhận thức tiếng nói dựa trên mô phỏng vùng vỏ não liên kết giữa thính giác và thị giác bằng cách xây dựng mô hình học mối quan hệ giữa các đặc trưng thu được từ âm thanh và hình ảnh trên vùng vỏ não liên kết đa giác quan này.

Phạm vi nghiên cứu của đề tài tập trung vào các vấn đề sau: Xử lý với các đoạn tín hiệu âm thanh của tiếng nói, lựa chọn đặc trưng dựa trên đặc trưng về ảnh phổ của tín hiệu tiếng nói, nhận thức tiếng nói ở mức độ liên kết giữa tín hiệu tiếng nói với từ định nghĩa sẵn, nhận thức tiếng nói ở khía cạnh liên kết với tín hiệu hình ảnh.

3. Các đóng góp của luận án

- Đề xuất sử dụng đặc trưng SIFT được trích chọn từ ảnh phổ của tín hiệu tiếng nói.

- Đề xuất sử dụng kết hợp giữa phương pháp phân lớp LNBNN và phương pháp trích chọn đặc trưng SIFT trên ảnh phổ của tiếng nói áp dụng cho bài toán nhận dạng tiếng nói.

- Đề xuất xây dựng mô hình nhận thức tiếng nói mô phỏng việc nhận thức của con người ở vùng não liên kết đa giác quan bằng cách xây dựng mô hình học mối quan hệ giữa tín hiệu tiếng nói với tín hiệu hình ảnh.

- Đề xuất cải tiến hiệu năng của mô hình thông qua việc rút gọn dữ liệu dựa trên trung vị của các thành phần của véc tơ đặc trưng.

- Đề xuất cài đặt phương pháp phân lớp LNBNN trên nền Hadoop, cho phép kết hợp nhiều máy tính có cấu hình thấp hơn để tạo thành một hệ thống xử lý song song, phân tán mạnh hơn.

4. Bố cục của luận án

Chương 1: Giới thiệu sơ lược các bài toán cơ bản của bài toán nhận thức tiếng nói, các bước trong quá trình nhận thức tiếng nói ở con người, trong việc mô phỏng nhận thức tiếng nói của các mô hình học máy. Giới thiệu tổng quan các nghiên cứu về bài toán nhận thức tiếng nói, cũng như các khó khăn trong bài toán này.

Chương 2: Giới thiệu tổng quan về các lý thuyết, mô hình và một số mô hình học máy cho bài toán nhận thức tiếng nói. Chương này cũng giới thiệu một số phương pháp trích chọn đặc trưng phổ biến được sử dụng trong các mô hình học máy cho bài toán nhận thức tiếng nói.

Chương 3: Giới thiệu tổng quan về ảnh phổ của tín hiệu tiếng nói, đặc trưng SIFT và cách trích chọn đặc trưng SIFT từ ảnh phổ của tín hiệu tiếng nói, giới thiệu hướng tiếp dựa trên ảnh phổ cho bài toán nhận thức tiếng nói kết hợp với việc áp dụng phương pháp phân lớp LNBNN. Mô hình được tiến hành 6 thí nghiệm khác nhau để

đánh giá hiệu quả của mô hình cho bài toán nhận dạng tiếng nói các từ, cụm từ độc lập.

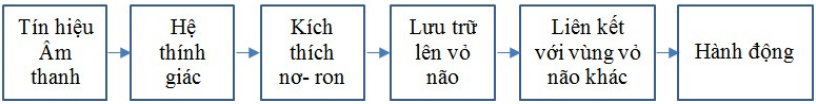
Chương 4: Giới thiệu tổng quan về quá trình nhận thức của con người, đánh giá các vấn đề tồn tại, đề xuất mô hình nhận thức tiếng nói dựa trên việc học mối quan hệ giữa tiếng nói với khái niệm cho trước và tín hiệu hình ảnh thu được biểu diễn cho một sự vật, hiện tượng xảy ra cùng lúc với tín hiệu âm thanh được nghe thấy.

Chương 5: Giới thiệu hai cải tiến cho bài toán nhận thức tiếng nói đó là đề xuất một phương pháp rút gọn đặc trưng bằng lượng tử hóa các thành phần của đặc trưng SIFT thành nhị phân sau đó mã hóa lại thành một đặc trưng mới và đề xuất cài đặt phương pháp phân lớp LNBNN trên nền tảng Hadoop cho bài toán nhận dạng tiếng nói.

Chương 1. TỔNG QUAN VỀ NHẬN THỨC TIẾNG NÓI

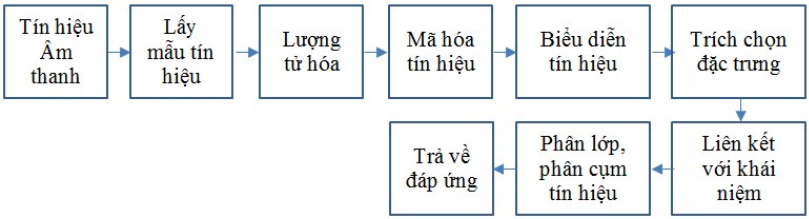
1.1. Giới thiệu

Nhận thức tiếng nói là phân biệt hay hiểu được sự khác nhau giữa các tín hiệu tiếng nói để từ đó có hành động đáp ứng phù hợp. Quá trình nhận thức tiếng nói ở con người gồm các bước sau:



Hình 1.1 Sơ đồ quá trình nhận thức tiếng nói

Các mô hình học máy cho bài toán nhận thức tiếng nóimô phỏngcơ chế hoạt động nhận thức tiếng nói của con người. Quá trình mô phỏng nhận thức tiếng nói trong máy tính cơ bản có những bước sau:



Hình 1. 2 Mô phỏng các bước trong nhận thức tiếng nói của máy tính

1.2. Một số bài toán trong nhận thức tiếng nói

Các nghiên cứu về nhận thức tiếng nói thường tập trung nhiều nhất trong việc giải quyết một số bài toán cụ thể đó là bài toán nhận dạng người nói và bài toán nhận dạng tiếng nói.

1.3. Quá trình nhận thức tiếng nói ở người

Quá trình nhận thức tiếng nói được bắt đầu từ việc thu nhận tín hiệu âm thanh ở người được trải qua một số giai đoạn sau:Thu nhận tín hiệu tiếng nói ở tai ngoài; Thu nhận tiếng nói ở tai giữa; Cơ chế truyền sóng âm ởốc tai đến nhận thức tiếng nói ở não.

1.4. Quá trình mô phỏng nhận thức âm thanh trên máy tính

Tín hiệu tiếng nói là tín hiệu tương tự, do đó để hệ thống máy tính có thể mô phỏng được quá trình nhận thức tiếng nói thì tín hiệu tiếng nói phải được biến đổi, biểu diễn và xử lý một cách phù hợp với máy tính. Các bước trong các mô hình học máy cho bài toán nhận thức tiếng nói gồm các bước sau:Lấy mẫu tín hiệu tiếng nói; Lượng tử hoá các mẫu; Mã hóa các mẫu lượng tử hóa; Biểu diễn tín hiệu tiếng nói;Trích chọn đặc trưng tiếng nói; Liên kết với khái niệm; Phân lớp, phân cụm dữ liệu.

1.5. Tổng quan về nghiên cứu về nhận thức tiếng nói

Những nghiên cứu đầu tiên về nhận thức tiếng nói là nghiên cứu khả năng phân biệt một tín hiệu nhất định từ các âm thanh khác mà chúng xuất hiện đồng thời trong cùng môi trường hay còn được gọi tên là hiệu ứng bữa tiệc hay bài toán nhận thức nhiều người nói(Cherry, 1953),(Broadbent & Ladefoged, 1957).

Nghiên cứu đầu tiên về bài toán nhận dạng tiếng nói được thực hiện trong phòng thí nghiệm Bell vào năm 1952 để nhận dạng các số của một người nói. Sau thành công của thí nghiệm này, nhiều hướng nghiên cứu được đưa ra nhằm nâng cao như: Hướng tiếp cận tích hợp nguồn hay khả năng tích hợp thông tin từ nhiều phương thức khác nhau cho bài toán nhận dạng tiếng nói(Sumby & Pollack, 1954), (Massaro, 1998); Hướng nghiên cứu vai trò của não đối với nhận dạng tiếng;Nghiên cứu về vai trò của bộ nhớ đối với nhận thức tiếng

nói có thể kể đến là Miller như(Miller G. , 1956), (Pisoni, 1973),(Goldinger, 1998),(Allen & Miller, 2004),(Smith, 2004).

Các nghiên cứu về nhận dạng tiếng nói đã được một số tác giả tổng hợp và xây dựng nên các lý thuyết và mô hình cho bài toán nhận thức tiếng nói: mô hình nhận dạng tiếng nói dựa trên phân tích bằng tổng hợp (*analysis-by-synthesis*) (Halle & Stevens, 1962); lý thuyết vận động (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967); Lý thuyết lượng tử hóa (*Quantal Theory*)(Stevens, The quantal nature of speech: Evidence from articulatory-acoustic data, 1972),(Stevens, On the quantal nature of speech, 1989); Mô hình nhận Cohort(Marslen-Wilson, Functional parallelism in spoken word recognition, 1987);Lý thuyết mẫu (Pierrehumbert,2001).

Trong khoa học máy tính, nhiều mô hình học máy cũng được nghiên cứu và áp dụng cho bài toán nhận thức tiếng nóinhư mô hình Markov ẩn (HMM), mô hình GMM, phương pháp SVM, hay mạng nơ-ron(Sak, 2014)(Soltau, 2014).

1.6. Một số khó khăn trong nhận thức tiếng nói

Tính tuyến tính: trong một phát âm liên tục mỗi âm thường chịu ảnh hưởng rất lớn từ các âm trước và sau nó.

Phân đoạn tiếng nói: là quá trình xác định ranh giới giữa các từ, âm tiết, âm vị trong ngôn ngữ nói.

Vấn đề phụ thuộc người nói: mỗi người nói sẽ có cấu trúc của bộ máy tạo âm khác nhau dẫn đến đặc tính của tiếng nói phát ra chịu ảnh hưởng rất nhiều vào người nói.

Vấn đề nhiễu: tín hiệu tiếng nói thường bị ảnh hưởng bởi các tạp âm từ môi trường ngoài.

Đơn vị nhận thức cơ bản: lựa chọn đơn vị nhỏ nhất để phân tích.

1.7. Hướng tiếp cận mới cho bài toán nhận thức tiếng nói

Từ những phân tích trên có thể thấy bài toán nhận thức là một lĩnh vực rất rộng, từ đó khái niệm nhận thức tiếng nói trong nghiên cứu này được hiểu là "*nhận thức tiếng nói là nhận thức hay hiểu được sự khác nhau giữa các tín hiệu tiếng nóiđể từ đó có hành động đáp ứng phù hợp*".

Trong khuôn khổ của nghiên cứu này chúng tôi chỉ tập trung nghiên cứu tới khía cạnh nhận thức tiếng nói ở khía cạnh liên kết giữa tín hiệu tiếng nói với một khái niệm (bài toán nhận dạng từ, cụm từ độc lập – chương 3) và liên kết giữa tín hiệu tiếng nói với tín hiệu hình ảnh, đề xuất mô hình nhận thức tiếng nói dựa trên mô hình mô phỏng quá trình liên kết thông tin ở vùng vỏ não liên kết đa giác quan (chương 4). Đây là một hướng tiếp cận mới so với các tiếp cận trước đây cho bài toán nhận thức tiếng nói bởi vì các hướng tiếp cận trước đây chủ yếu tập trung mô phỏng quá trình nhận thức tiếng nói ở vùng nhớ sơ cấp và vùng nhớ liên kết của cơ quan thính giác, rất ít nghiên cứu đề cập tới vùng nhớ liên kết đa giác quan này.

Chương 2.Lý thuyết, mô hình và phương pháp cho bài toán nhận thức tiếng nói

2.1.Giới thiệu

Trong phần này sẽ giới thiệu một số lý thuyết và mô hình cho bài toán nhận thức tiếng nói đồng thời giới thiệu một số mô hình học máy và phương pháp trích chọn đặc trưng tiếng nói trong các mô hình học máy cho bài toán nhận thức tiếng nói.

2.2.Một số lý thuyết cho bài toán nhận thức tiếng nói

Lý thuyết vận động: được phát triển bởi Liberman và các đồng nghiệp vào năm 1967. Nguyên lý cơ bản của lý thuyết này là dựa trên việc sản sinh tiếng nói trong đường phát âm của người nói.

Lý thuyết phân tích bằng tổng hợp: nhận thức tiếng nói dựa trên thông tin về quá trình sản xuất tiếng nói.

Lý thuyết mẫu: được giới thiệu lần đầu tiên trong tâm lý học như là một mô hình nhận thức và phân loại, sau đó được Lacerda (1995), Johnson(1997), Pierrehumbert (2001) áp dụng cho bài toán nhận thức tiếng nói [30]. Lý thuyết này dựa trên liên kết giữa bộ nhớ và kinh nghiệm trước với các từ vựng.

2.3.Một số mô hình cho bài toán nhận thức tiếng nói

Mô hình TRACE là một framework lấy tất cả các nguồn thông tin khác nhau trong tiếng nói và tích hợp chúng để nhận dạng các từ.

Mô hình nhận thức tiếng nói Cohort được đề xuất bởi Marslen-Wilson vào năm 1984 để nhận dạng từ vựng bằng cách sử dụng các âm vị ban đầu để kích hoạt tập các từ có cùng âm vị khởi đầu. Khi thu nhận được thêm thông tin tiếp theo, tập từ vựng được thu hẹp.

Mô hình luồng kép của Hickok và Poeppel (2007) chứng minh sự hiện diện của hai mạng nơ-ron riêng biệt trong xử lý tiếng nói. Một mạng nơ-ron chủ yếu xử lý với các giác quan và thông tin âm vị liên quan đến các khái niệm và ngữ nghĩa. Mạng còn lại hoạt động với giác quan và thông tin âm vị liên quan đến hệ thống động cơ và hệ thống cấu âm.

Mô hình tính toán nơ-ron mô phỏng các con đường của nơ-ron thần kinh ở những vùng khác nhau của não bộ có liên quan đến quá trình sản xuất và nhận thức tiếng nói. Các vùng não chứa tri thức tiếng nói thu được bằng cách huấn luyện các mạng nơ-ron để phát hiện tiếng nói trong vùng vỏ não và vỏ não tiểu não.

2.4. Một số mô hình học máy cho bài toán nhận thức tiếng nói

Mô hình Markov ẩn: HMM là mô hình điển hình tiếp cận theo mô hình âm học cho bài toán nhận dạng tiếng nói. HMM là mô hình xác suất dựa trên lý thuyết về chuỗi Markov gồm các thành phần sau:

* $O = \{o_1, o_2, .., o_T\}$ là tập các vector quan sát.

* $S = \{s_1, s_2, .., s_N\}$ là tập hữu hạn các trạng thái s gồm N phần tử

* $A = \{a_{11}, a_{12}, .., a_{MN}\}$ là ma trận hai chiều trong đó a_{ij} thể hiện xác suất để trạng thái $s_i$chuyển sang trạng thái s_j, với $a_{ij} \geq 0$ và $\sum_{j=1}^{k} a_{ij} = 1 \; \forall i$.

* $B = \{b_{2t}, b_{it}, .., b_{(N-1)t}\}$ là tập các hàm xác suất phát tán của các trạng thái từ s_2đến s_{N-1}, trong đó b_{it}thể hiện xác suất để quan sát $o_t$thu được từ trạng thái s_itại thời điểm t.

Mô hình mạng nơ-ron: Mạng nơron MLPlà một cấu trúc mạng gồm có một lớp vào, một lớp ra và một hoặc nhiều lớp ẩn. Vector đầu vào sẽ được đưa qua lớp vào sau đó các tính toán được thực hiện lan truyền tiến từ lớp vào tới các lớp ẩn và kết thúc ở lớp ra. Ngoài mạng MLP, mô hình mạng hồi quy cũng thường được sử dụng cho bài toán nhận thức tiếng nói.

Mô hình ngôn ngữ: Mô hình ngôn ngữ là một tập xác suất phân bố của các đơn vị trên một tập văn bản cụ thể. Một cách tổng quát thông qua mô hình ngôn ngữ cho phép ta xác định xác suất của một cụm từ hoặc một câu trong một ngôn ngữ.

2.5.Một số phương pháp trích chọn đặc trưng tiếng nói

Phương pháp trích đặc trưng MFCC:tính toán các giá trị phổ của tín hiệu cho băng tần trên miền tần số mà tai người dễ cảm thụ nhất.

Phương pháp mã dự đoán tuyến tính LPC: tính các hệ số để xấp xỉ một mẫu bởi tổ hợp tuyến tính của các mẫu trước đó.

Phương pháp trích đặc trưng PLP: dựa trên cơ sở phương pháp mã dự báo tuyến tính LPC. Đặc trưng này được tạo ra dựa trên đặc tính vật lý của tai người khi nghe.

Chương 3. Hướng tiếp cận trích chọn đặc trưng từ ảnh phổ của tín hiệu cho bài toán nhận thức tiếng nói

3.1.Giới thiệu

Các mô hình học máy cho bài toán nhận thức tiếng nói hiện nay hầu hết là sử dụng các đặc trưng dựa MFCC, LPC và PLP. Các đăc trưng này sử dụng các bộ lọc tần số dẫn tới một số thành phần tần số có trong tín hiệu tiếng nói đã bị bỏ qua, làm mất thông tin có trong tín hiệu tiếng nói. Các đặc trưng này rất nhạy cảm với nhiễu và thiếu thông tin về pha.Thêm vào đó, các mô hình học máy thường đòi hỏi dữ liệu đầu vào phải cùng kích thước, do đó các mô hình học máy thường phải biến đổi dữ liệu ban đầu để biểu diễn dữ liệu thành các véc tơ cùng chiều dẫn đếnlàm mất thông tin.

Chương này chúng tôi đề xuất sử dụng trích chọn đặc trưng SIFT trực tiếp từ ảnh phổ của tín hiệu tiếng nói kết hợp phương pháp học máy LNBNN cho bài toán nhận thức tiếng nói.

3.2. Ảnh phổ của tín hiệu tiếng nói

Ảnh phổ của tiếng nói là một phương pháp biểu diễn tín hiệu trên miền kết hợp thời gian và tần số trong đó một chiềubiểu diễn tần số, một chiều biểu diễn thời gian và giá trị mỗi điểm ảnh là độ lớn của các thành phần tần số có trong tín hiệu.

3.3.Đặc trưng bất biến SIFT

SIFT là đặc trưng bất biến đối với phép tịnh tiến, co dãn và phép xoay. Phương pháp trích rút các đặc trưng SIFT được tiếp cận theo

phương pháp thác lọctheo các bước sau: Phát hiện các điểm cực trị Scale-Space; Định vị các điểm hấp dẫn; Xác định hướng cho các điểm hấp dẫn; Mô tả các điểm hấp dẫn.

3.4.Thuật toán phân lớp NBNN

Thuật toán 3.1

Đầu vào:
$C = \{C_1, C_2, .., C_L\}$ là tập nhãn của dữ liệu huấn luyện
$T = \{T_1, T_2, ..., T_L\}$ là tập các đặc trưng của dữ liệu huấn luyện
$Q = \{d_1, d_2, .., d_Q\}$ $with$ $d_i \in R^m$ $\forall i = 1 ... Q$ là một truy vấn

Đầu ra: nhãn của Q

for all $d_i \in Q$ do
    for all classes C do
        $totals[C] \leftarrow totals[C] + \|d_i - NN_C(d_i)\|^2$
    end for
end for
return $\text{argmin}_C\ totals[C]$

3.5.Phương pháp phân lớp LNBNN

Phương pháp phân lớp LNBNN được Sancho đề xuất nhằm cải tiến thuật toán NBNN cho bài toán phân lớp ảnh.

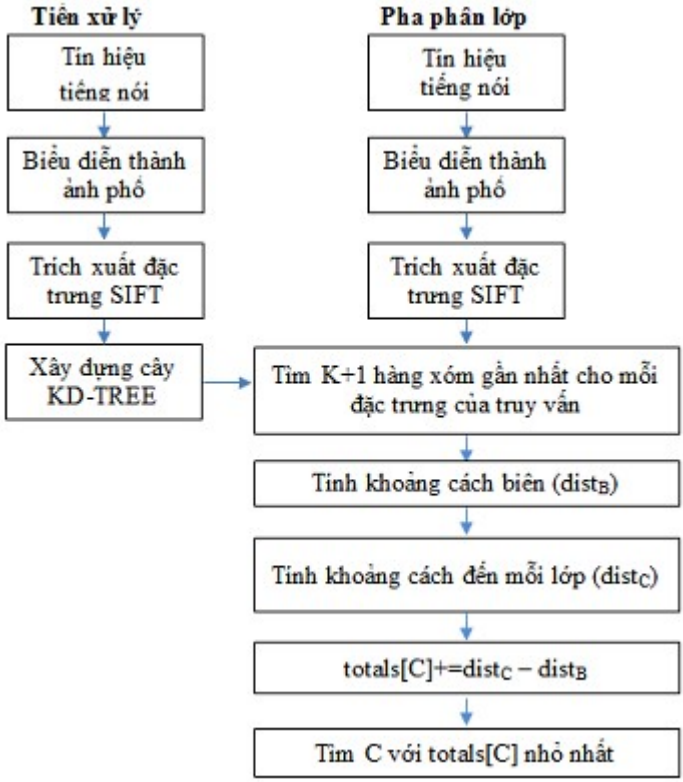
Thuật toán 3.2

Đầu vào:
$T = \{T_1, T_2, \ldots, T_N\}$ là tập mẫu huấn luyện
$T_i = \{d_{i_1}, d_{i_2}, \ldots, d_{i_{N_i}}\}$ and $d_{i_j} \in R^m \; \forall j = 1..N_i$
$C = \{C_1, C_2, .., C_N\}$ là tập nhãn
$Query\ Q = \{d_1, d_2, .., d_Q\}, d_i \in R^m \; \forall i = 1..Q$
Tham số k
Đầu ra: nhãn của Q

for all $d_i \in Q$ **do**
  find $\{p_1, p_2, .., p_{k+1}\}$ là $k+1$ hàng xóm gần nhất của d_i
  $dist_B = \|d_i - p_{k+1}\|^2$
  for all classes C in the k nearest neighbors **do**
    $dist_C = min_{\{p_j | Class(p_j) = C\}} \|d_i - p_j\|^2$
    $totals[C] \leftarrow totals[C] + dist_C - dist_B$
  end for
end for
return $\text{argmin}_C\ totals[C]$

3.6.Hướng tiếp cận ảnh phổ cho bài toàn nhận dạng tiếng nói

Trong nghiên cứu này, chúng tôi đề xuất mô hình phân lớp tiếng nói dựa trên ảnh phổ của tín hiệu tiếng nói bằng cách áp dụng phương pháp phân lớp LNBNN kết hợp với phương pháp trích chọn đặc trưng bất biến SIFT trên ảnh phổ của tín hiệu tiếng nói (Hình 3.8).



Hình 3. 1Mô hình phân lớp tiếng nói bằng LNBNN kết hợp với đặc trưng SIFT trên ảnh phổ của tiếng nói

3.7. Thí nghiệm và kết quả

3.7.1. Dữ liệu thí nghiệm: thí nghiệm được tiến hành trên 06 bộ dữ liệu là: ISOLET, English Digits, Vietnamese Places, Vietnamese Digits, TMW, JVPD.

3.7.2. Thí nghiệm so sánh độ chính xác phân lớp của đặc trưng SIFT với đặc trưng MFCC khi sử dụng LNBNN

Bảng 3. 1 So sánh độ chính xác phân lớp của LNBNN với SIFT và MFCC

Bộ dữ liệu	SIFT	MFCC

ISOLET	0.73	0.34
English Digits	0.96	0.94
Vietnamese Places	0.95	0.39
Vietnamese Digits	0.97	0.72
TMW	1.00	0.39
JVPD	0.97	0.53

3.7.3. Thí nghiệm với dữ liệu co dãn theo thời gian

Bảng 3. 1 So sánh kết quả đối với dữ liệu bị co dãn một chiều

Database	Origin	Scale 10%	Scale 20%	Scale 30%
ISOLET	0.734	0.731	0.729	0.724
English Digits	0.962	0.962	0.959	0.958
Vietnamese Places	0.953	0.951	0.948	0.941
VietnameseDigits	0.972	0.971	0.969	0.965
TMW	1.000	1.000	0.991	0.985
JVPD	0.973	0.972	0.967	0.963

3.7.4. Thí nghiệm so sánh LNBNN và các phân loại khác

Bảng 3.3 So sánh độ chính xác của các phương pháp phân lớp với đặc trưng MFCC

Method	ISOLET	EN Digits	VN Places	VN Digits	TMW	JVPD
LNBNN	34.0	94.1	38.5	72.0	39.0	87.1
Naïve Bayes	64.2	98.6	67.6	42.4	44.6	44.5
Bayes Net	57.0	**99.5**	70.2	47.5	21.3	21.3
SVM	61.6	**99.5**	**78.0**	62.8	40.7	96.5
RandomForest	**64.4**	98.4	71.8	**73.5**	**56.7**	**97.2**
TreeJ48	38.1	90.2	53.8	42.4	15.2	82.7

Bảng 3.4 So sánh độ chính xác của các phương pháp phân lớp với đặc trưng SIFT

Method	ISOLET	EN Digits	VN Places	VN Digits	TMW	JVPD
LNBNN	**72.8**	**96.2**	**95.0**	**96.9**	**100.0**	**96.9**
Naïve Bayes	32.8	50.4	58.5	53.1	34.1	55.8
Bayes Net	20.6	57.2	70.5	47.7	33.1	60.8
SVM	3.8	11.3	12.5	14.6	8.5	35.2
RandomForest	37.7	70.7	78.5	55.2	69.0	62.4
Tree J48	18.3	47.3	60.3	34.6	17.4	46.8

3.7.5. Thí nghiệm khả năng học tăng cường của LNBNN

Bảng 3.5So sánh độ chính xác phân lớp khi bổ sung thêm dữ liệu

Database	20%	40%	60%	80%	100%

	training samples	training samples	training samples	training samples	training samples
ISOLET	0.46	0.56	0.60	0.68	0.73
English Digits	0.90	0.92	0.94	0.95	0.96
VN Places	0.91	0.92	0.93	0.94	0.95
VN Digits	0.27	0.72	0.71	0.82	0.97
TMW	0.92	0.93	0.98	0.99	1.00
JVPD	0.94	0.96	0.96	0.95	0.97

Bảng 3.6 So sánh độ chính xác phân lớp khi bổ sung thêm lớp (tri thức)

Database	20% classes	40% classes	60% classes	80% classes	100% classes
ISOLET	0.55	0.64	0.60	0.60	0.73
English Digits	1.00	0.98	0.98	0.97	0.96
VN Places	1.00	0.97	0.95	0.94	0.95
VN Digits	1.00	0.97	0.98	0.96	0.97
TMW	1.00	1.00	1.00	1.00	1.00
JVPD	1.00	1.00	0.97	0.97	0.97

3.6.Kết luận

Trong chương này, chúng tôi đã đề xuất một phương pháp trích chọn đặc trưng tiếng nói ở mức độ thính giác dựa trên ảnh phổ của tín hiệu tiếng nói đồng thời kết hợp với phương pháp phân lớp LNBNN phương pháp phân lớp phi tham số có ưu điểm là cho phép mô hình có thể học thêm mẫu dữ liệu huấn luyện, học thêm tri thức mà không phải huấn luyện lại.

Chương 4.Mô hình nhận thức tiếng nói thông qua học mối quan hệ giữa tín hiệu tiếng nói và hình ảnh

4.1. Giới thiệu

Trong chương này, chúng tôi xây dựng mô hình nhận thức tiếng nói thông qua việc học mối quan hệ giữa các đặc trưng từ một cặp dữ liệu tiếng nói và hình ảnh xảy ra đồng thời mà người học thu nhận được thông qua hai cơ quan cảm giác chính đó là thính giác và thị giác.

4.2. Các phương pháp học mối quan hệ

Học mối quan hệ bằng mạng nơ-ron: thường được dùng để học mối quan hệ giữa các dữ liệu trong cùng một miền. Mối quan hệ được thể hiện ở trọng số của mạng.

Học mối quan hệ bằng HMM: học mối quan hệ giữa dữ liệu trong cùng một miền có tính liên kết theo thời gian, dạng chuỗi. Mối quan hệ được thể hiện ở ma trận chuyển trạng thái.

Học mối quan hệ dựa trên luật: thường học mối quan hệ trong văn bản. Quan hệ thể hiện ở dạng luật.

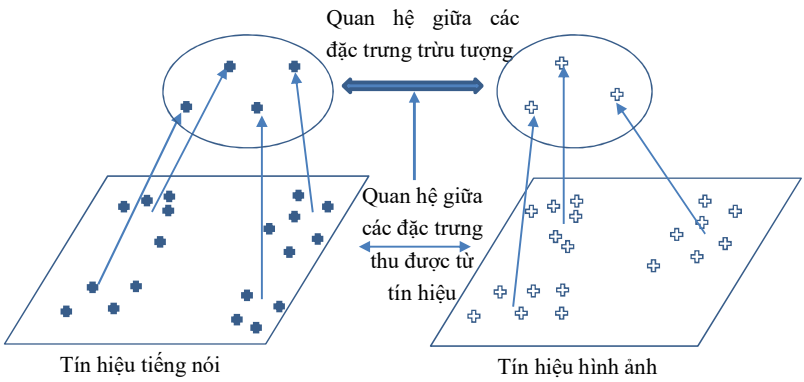
4.3.Đề xuất mô hình nhận thức tiếng nói

Cơ sở đề xuất mô hình

Vỏ não là lớp vỏ ngoài của chất xám trên bán cầu. Một số vùng vỏ não có chức năng đơn giản hơn, gọi là vỏ não sơ cấp (Wanda, 2017). Vỏ não gồm các khu vực trực tiếp tiếp nhận thông tin từ các cơ quan giác quan như thị giác, thính giác, xúc giác, vị giác và vùng vỏ não liên kết có các chức năng phức tạp hơn vùng vỏ não sơ cấp. Vùng vỏ não liên kết được chia làm hai loại là vùng vỏ não liên kết của các cơ quan cảm giác và vùng vỏ não liên kết đa giác quan.

Vùng vỏ não liên kết của mỗi giác quan có vai trò trong việc lưu trữ mối quan hệ giữa các tín hiệu của giác quan đó, trong khi đó, vùng vỏ não liên kết đa giác quan có vai trò trong việc liên kết thông tin của các giác quan khác nhau để nhận thức.

Theo hướng tiếp cận này, để máy tính nhận thức được tiếng nói thực chất là xây dựng được mạng quan hệ giữa tín hiệu tiếng nói với thông tin về các sự vật hiện tượng thu được từ các giác quan khác. Các tín hiệu âm thanh của một đối tượng (khái niệm về lớp trừu tượng) nào đó sẽ được nhận thức bởi một số bởi một số đặc trưng nhất định được gọi là đặc điểm chung của đối tượng đó. Tương tự vậy, các tín hiệu hình ảnh của cùng một đối tượng, một khái niệm cũng sẽ được nhận thức bởi một số đặc trưng hình ảnh chung nhất của đối tượng đó. Khi đó, nhận thức tiếng nói là quá trình xây dựng mạng quan hệ giữa các tập đặc trưng này.



Định nghĩa 1: Quan hệ giữa một mẫu tiếng nói và một mẫu hình ảnh:Một mẫu tiếng nói thu được từ hệ thính giác đồng thời với một hình ảnh của sự vật, hiện tượng từ môi trường xung quanh tại cùng một thời điểm thì được gọi là có quan hệ.

Định nghĩa 2. Quan hệ một đặc trưng tiếng nói với một đặc trưng hình ảnh.

Giả sử có một mẫu tiếng nói S được biểu diễn bằng một tập các đặc trưng $\{f_1 , f_2 , ...\}$, và một mẫu hình ảnh được biểu diễn bởi tập đặc trưng $\{g_1 , g_2 , ...\}$. Khi đó đặc trưng f_i và đặc trưng g_j được gọi là có quan hệ nếu S có quan hệ với I.

Mô hình nhận thức tiếng nói bằng học mối quan hệ giữa tín hiệu âm thanh và hình ảnh

Bài toán được mô hình hóa như sau: Cho một tập dữ liệu huấn luyện là một tập các cặp mẫu gồm một tín hiệu tiếng nói và một hình ảnh mà hai giác quan thu được tại cùng một thời điểm. Như vậy mỗi mẫu huấn luyện là một cặp $<S_i, I_i>$. Như vậy, khi cho một mẫu mới là một cặp <S, I> bất kỳ, hỏi cặp mẫu <S, I> này là có quan hệ với nhau hay không?

Chúng tôi đề xuất cải tiến LNBNN để phân lớp các cặp dữ liệu thành 2 lớp là có quan hệ và không có quan hệ như sau:

Cách 1: Sử dụng pha phân lớp của LNBNN: cải tiến cách lưu trữ và tìm kiếm K hàng xóm gần nhất.

Cách 2: Sử dụng phân lớp LNBNN với ước lượng xác suất KNN: cải tiến ước lượng xác suất bằng KNN.

Cách 3. Sử dụng LNBNN một lớp

Thực chất là bài toán chỉ có một tập nhỏ các cặp dữ có quan hệ được sử dụng làm tập huấn luyện chứ không có cặp dữ liệu không có quan hệ trong tập huấn luyện. Vì vậy bài toán phải coi là bài toán phân lớp quan hệ chỉ có một lớp (one class classification). Từ đó, chúng tôi đề xuất phân lớp theo thuật toán 4.2.

Thuật toán 4. 1. Thuật toán học mối quan hệ - Pha phân lớp

Đầu vào:
SF: cây đặc trưng của dữ liệu huấn luyện tiếng nói
IF: cây đặc trưng của dữ liệu huấn luyện hình ảnh
W: Ma trận trọng số quan hệ
{sp, im}: một cặp mẫu truy vấn {speech, image}
Threshold: tham số ngưỡng

Đầu ra: cặp mẫu truy vấn {sp, im} có quan hệ hay không

```
1: TotalWeight = 0;
2:  Tìm tập SP_index là K+1 hàng xóm gần nhất của các đặc trưng của mẫu tiếng nói trong cây SF
3:  Tìm tập IM_index là chỉ số của K+1 hàng xóm gần nhất của các đặc trưng trong mẫu hình ảnh trong cây IM
4:  For each i in SP_index
5:      For each j in IM_index
6.          TínhdistB khoảng cách tới cặp biên được tạo từ phần tử K+1
7:          Tính khoảng cách ngắn nhất distC của cặp dữ liệu
8:          TotalWeight = TotalWeight + w(i,j)*(distC - distB)/(N*M)
9:      End for
10: End for
11: If TotalWeight < Threshold Then
12:    return true
13: Else if
14:     return false
15: End if
```

4.4.Thí nghiệm và kết quả

4.4.1Xây dựng tập dữ liệu thí nghiệm

Bộ dữ liệu thí nghiệm thứ nhất được xây dựng từ bộ dữ liệu DIGITS,và bộ dữ liệu ảnh MNIST. Từ hai bộ dữ liệu này chúng tôi chọn ngẫu nhiên 454 mẫu huấn luyện và chia thành hai tập, tập huấn luyện gồm 266 mẫu và tập kiểm tra là 188 mẫu.

Bộ dữ liệu thứ hai được xây dựng từ bộ dữ liệu tiếng nói là tên gọi của 3 đối tượng (Bút, Quả bóng và Điện thoại) và một bộ dữ liệu ảnh chụp ba đối tượng đó ở khoảng cách và góc chụp khác nhau. Bộ dữ liệu gồm 100 mẫu huấn luyện và 40 mẫu kiểm tra mỗi lớp.

4.4.2 Thí nghiệm học mối quan hệ dựa trên LNBNN

Bảng 4. 1 Kết quả phân lớp mối quan hệ bằng LNBNN trên dữ liệu DIGITS

K	TP	FP	TN	FN	Accuracy
2	1249	633	821	1061	0.614
4	1204	678	771	1111	0.615
6	1206	676	776	1106	0.614
8	1206	676	792	1090	0.610
10	1211	671	792	1090	0.611
12	1212	670	792	1090	0.612
14	1212	670	791	1091	0.612
16	1213	669	790	1092	0.612
18	1213	669	787	1095	0.613
20	1210	672	750	1132	0.622

Bảng 4. 2 Kết quả phân lớp quan hệ với LNBNN trên dữ liệu OBJECTS

K	TP	FP	TN	FN	Accuracy
2	22	18	32	8	0.375
4	28	12	32	8	0.450
6	32	8	32	8	0.500
8	33	6	33	7	0.506
10	34	5	35	5	0.494
12	37	3	37	3	0.500
14	39	1	38	2	0.513
16	40	0	40	0	0.500
18	40	0	40	0	0.500
20	40	0	40	0	0.500

4.4.3 Thí nghiệm học mối quan hệ dựa trên LNBNN với KNN

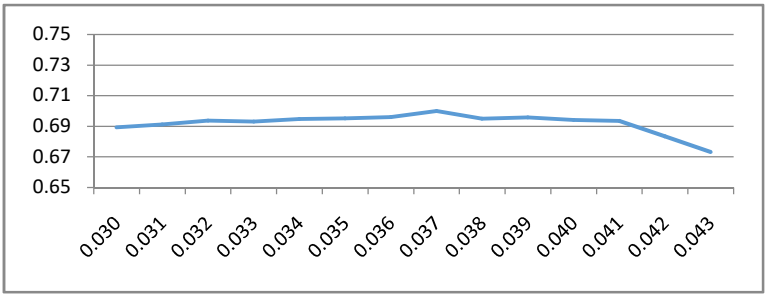
Bảng 4. 3 Kết quả phân lớp quan hệ áp dụng KNN trên dữ liệu DIGITS

K	TP	FP	TN	FN	Accuracy
2	1448	434	924	958	0.639
4	1627	255	1031	851	0.658
6	1696	186	1166	716	0.641
8	1734	148	1340	542	0.605
10	1756	126	1465	417	0.577
12	1790	92	1550	332	0.564
14	1815	67	1688	194	0.534
16	1832	50	1787	95	0.512
18	1850	32	1837	45	0.503
20	1882	0	1882	0	0.500

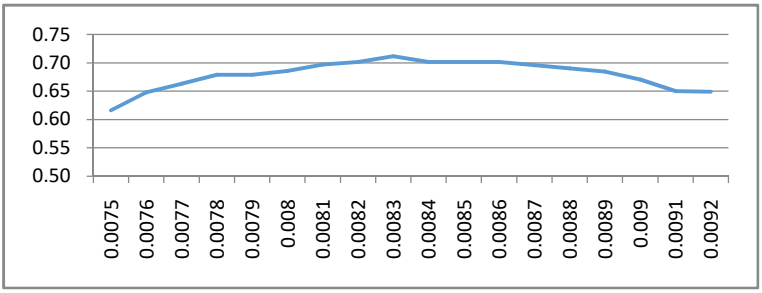
Bảng 4. 4 Kết quả phân lớp quan hệ áp dụng KNN trên dữ liệu OBJECTS

K	TP	FP	TN	FN	Accuracy
2	4	36	0	40	0.550
4	6	34	0	40	0.575
6	9	31	0	40	0.613
8	12	28	1	39	0.638
10	14	26	2	38	0.650
12	16	24	4	36	0.650
14	18	22	6	34	0.650
16	19	21	8	32	0.638
18	21	19	10	30	0.638
20	22	18	12	28	0.625

4.4.4 LNBNN một lớp cho bài toán phân lớp quan hệ



Hình 4. 1 Kết quả phẩn lớp one-class LNBNN trên bộ dữ liệu DIGITS



Hình 4. 2 Kết quả phân lớp one-class LNBNN trên bộ dữ liệu OBJECTS

5.7. Kết luận

Chương này chúng tôi đề xuất một hướng tiếp cận cho bài toán

nhận thức tiếng nói dựa trên mô hình học mối quan hệ giữa các đặc trưng của tiếng nói với các đặc trưng thu được của hình ảnh bằng cách áp dụng phương pháp phân lớp đồng thời đề xuất ba cách cải tiến đối với phương pháp phân lớp LNBNN để áp dụng cho bài toán này. Kết quả thực nghiệm cũng chứng tỏ mô hình này là phù hợp và có thể cải tiến áp dụng cho việc huấn luyện người máy trong việc nhận thức tiếng nói.

Chương 5.Một số cải tiến cho bài toán nhận thức tiếng nói

5.1.Giới thiệu

Trong phần này, chúng tôi đề xuất một phương pháp rút gọn dữ liệu cho đặc trưng SIFT và đề xuất cài đặt phương pháp phân lớp LNBNN trên nền Hadoop cho bài toán phân lớp tiếng nói với dữ liệu lớn.

5.2.Rút gọn dữ liệu

Bảng 5. 1 So sánh độ chính xác phân lớp trên các bộ dữ liệu

Database	*Origin SIFT KD-TREE*	*Binary SIFT Linear Brute Force*	*Binary SIFT Hierarchical Clustering*	*Binary SIFT MIH*
ISOLET	56.3	56.3	56.3	56.3
EN DIGITS	95.4	95.8	95.3	**96.2**
VN PLACES	**91.2**	90.5	89.8	90.8
JVPD	**95.1**	94.6	93.7	95.0
TMW	**83.1**	89.9	89.9	89.9

Bảng 5.2 So sánh thời gian chạy trên các dữ liệu khác nhau (tính bằng giây)

Databases	*Num descriptor*	*Origin SIFT KD-TREE*	*Binary SIFT Linear Brute Force*	*Binary SIFT Hierarchical Clustering*	*Binary SIFT MIH*
ISOLET	327,396	657	654	124	473
EN.DIGITS	581,134	1,584	3,848	643	2,331
VN PLACES	856,121	725	13,359	307	1,919
JVPD	489,998	11,144	1,613	228	901
TMW	3,605,234	25,364	73,595	1,892	43,295

Chúng tôi đề xuất một phương pháp rút gọn dữ liệu bằng cách lượng tử hóa các thành phần của đặc trưng SIFT dựa trên trung vị của chúng. Như vậy, sau khi lượng tử hóa với các giá trị trung vịmỗi

điểm đặc trưng SIFTsẽ trở thành một véc tơ 128 bit, sau đó chúng được mã hóa thành véc tơ 16 bytes giảm kích thước 8 lần.

5.3.Cài đặt phương pháp phân lớp LNBNN trên nền Hadoop

Việc cài đặt thuật toán LNBNN được tiến hành ở các thủ tục Setup, Map, Reduce và Cleanup. Hai thủ tục chính là Map và Reduce được trình bày ở thuật toán 5.1 và 5.2.

Thuật toán 5. 1 Thuật toán LNBNN Hadoop – thủ tục Map

Input:
Value là dòng dữ liệu trong tập huấn luyện bao gồm cả dữ liệu và nhãn
Out put:
A list of <***KeyOut, ValueOut***>*pair.*
1. Convert Value (current line in training) to a vector **curVec**
2. **For** each ***test_vector*** **in** ***testList*** **do**
3. Calculate distance from curVec to test_vector
4. Create KeyOut =<feature_id, distance > is a pair of feature point id in query (***test_vector***) and its distance to the current feature point in training set (***curVec***)
5. Create ValueOut =<label, distance> is a pair of class label and its distance from a feature point id in query (***test_vector***) to the current feature point in training set (***curVec***)
6. Context.write(KeyOut,ValueOut)
7. **End for**

Trong thí nghiệm này chúng tôi thiết kế một hệ thống phân tán bao gồm 03 node được kết nối thông qua mạng cục bộ được tiến hành trên 04 cơ sở dữ liệu là DIGITS, VN PLACES, TMW, JVPD. Kết quả so sánh thời gian chạy được trình bày ở bảng 5.5.

Bảng 5.5 So sánh thời gian truy vấn trung bình một đặc trưng(tính bằng giây)

Database	Number feature	Single node	2 nodes	3 nodes
JVPD	489,998	295	302	201
English Digits	581,134	363	245	261
VN Places	3,190,303	1,902	1,858	1,927
TMW	3,605,234	2,253	1,606	1,471
VN Places + TMW	6,795,537	4,281	4,088	4,253
JVPD + English Digits + VN Places + TMW	7,866,669	4,806	4,700	4,938

Thuật toán 5. 2 Thuật toán LNBNN Hadoop – thủ tục Reduce

Input:
- ***K*** là số hàng xóm gần nhất cần tìm
- ***Key*** là một cặp gồm chỉ số của điểm đặc trưng và khoảng cách (Feature point Id of query, distance),
- ***Value*** là tập các cặp (class label, distance)

Output:
Totals : tổng khoảng cách từ truy vấn tới tất cả các lớp

```
1. Count =0;
2. For each RecordKey  in Value do
3.       If Count = K then
4.            BG_distance = recordKey.getDistance()
5.                 break;
6.       Else
7.              Count  = Count +1;
8.       End if
9.    If recordKey not in NeighborList then
10.    Add recordKey to NeighborList
11.  End if
12. End for
13. For each neighbor   in NeighborList do
14.         Totals[neighbor] +=   neighbor.Distance() – BG_distance;
15. End For
```

5.4. Kết luận

Trong chương này chúng tôi đề xuất hai cải tiến cho phương pháp phân lớp LNBNN cho bài toán nhận dạng tiếng nói dựa trên đặc trưng SIFT trích chọn từ ảnh phổ của tín hiệu tiếng nói. Một là, chúng tôi đề xuất phương pháp rút gọn đặc trưng bằng việc biến đổi đặc trưng SIFT từ 128 chiều, với mỗi chiều là một byte thành đặc trưng SIFT nhị phân, sau đó mã hóa lại thành một véc tơ 16 chiều để giảm kích thước lưu trữ và tăng tốc độ tính toán. Hai là, chúng tôi đề xuất cài đặt phương pháp phân lớp LNBNN song song, phân tán trên nền tảng Hadoop, một framework phổ biến cho bài toán xử lý dữ liệu lớn.

DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ

[1] Quang Trung, Nguyễn; Thế Duy, Bùi; Thị Châu, Ma; 2015, An Image based approach for speech perception, 2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science, Springer, 208 – 213.

[2]Quang Trung, Nguyen; The Duy, Bui;(2016) Speech classification using SIFT features on spectrogram images, Vietnam Journal of Computer Science, 3(4), 247-257.

[3] The Duy, Bui; Quang Trung, Nguyen; Speech classification by using binary quantized SIFT features of signal spectrogram images, 2016,3rd National Foundation for Science and Technology Development Conference on Information and Computer Science, IEEE.

[4] Quang Trung, Nguyen; The Duy, Bui; 2016, MapReduce based for speech classification , SoICT '16: Proceedings of the Seventh Symposium on Information and Communication Technology, ACM.

[5] The Duy, Bui; Quang Trung, Nguyen; (2016), Learning relationshipbetween speech and image, The Eighth International Conference on Knowledge and Systems Engineering (KSE) 2016, IEEE, 103-108.