**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**🙠🕮🙢**

**XỬ LÝ NGÔN NGỮ NÓI**

**BÀI TẬP 3: NHẬN DẠNG TIẾNG NÓI LIÊN TỤC**

GVHD: PGS. TS. Vũ Hải Quân

Nhóm thực hiện:

Ngành Khoa Học Máy Tính – Cao học khóa 23

1. Đỗ Đặng Minh - 13 11 015
2. Huỳnh Công Toàn - 13 11 026
3. Hồ Văn Tấn - 13 11 058

*Tp. Hồ Chí Minh, tháng 05 năm 2014*

# Thông tin nhóm

|  |  |  |  |
| --- | --- | --- | --- |
| **MSHV** | **Họ tên** | **Số điện thoại** | **E-mail** |
| 13 11 015 | Đỗ Đặng Minh | 0168-993-5242 | [masterminh219@gmail.com](mailto:masterminh219@gmail.com) |
| 13 11 026 | Huỳnh Công Toàn | 0121-516-1090 | [alex7huynh@gmail.com](mailto:alex7huynh@gmail.com) |
| 13 11 058 | Hồ Văn Tấn | 090-290-9334 | [tanhv90@gmail.com](mailto:tanhv90@gmail.com) |

Mục lục

[Thông tin nhóm 1](#_Toc387418544)

[I. Cơ sở lý thuyết 2](#_Toc387418545)

[1) Trích chọn đặc trưng 2](#_Toc387418546)

[2) Mô hình Markov ẩn (HMM) 2](#_Toc387418547)

[3) Các bài toán với mô hình Markov ẩn 3](#_Toc387418548)

[II. Thực nghiệm 3](#_Toc387418549)

[1) Dữ liệu thực nghiệm 3](#_Toc387418550)

[2) HTK 3](#_Toc387418551)

[3) Phương pháp thực hiện 4](#_Toc387418552)

[4) Thống kê kết quả 6](#_Toc387418553)

[III. Tài liệu tham khảo 7](#_Toc387418554)

# Cơ sở lý thuyết

1. Trích chọn đặc trưng

Các đặc trưng mang thông tin về tiếng nói là các tần số formant, tần số âm cơ bản và năng lượng.

Các đặc trưng mang thông tin về đường bao phổ:

* Hệ số dự đoán tuyến tính (LPC).
* Các hệ số cepstrum.
* Các hệ số cepstrum dự đoán tuyến tính (LPCC).
* Các hệ số cepstrum tần số mel (MFCC).

Trích chọn đặc trưng trong xử lý ngôn ngữ nói là biến đổi từ giọng nói thô thành những đặc trưng của người nói. Đặc trưng được sử dụng trong bài này là đặc trưng MFCC (Mel Frequency Cepstral Coefficients), đặc trưng MFCC thể hiện hình thái của cuống họng, … khi ta nói.

Các bước trính chọn đặc trưng MFCC:

* Phân khung (Frame blocking): chia tín hiệu đầu vào thành các đoạn nhỏ khoảng 20ms­30ms. Phân khung tín hiệu mỗi khung N mẫu, hai khung kề nhau lệch M mẫu.
* Lấy cửa sổ (Windowing): giảm sự gián đoạn của tín hiệu ở đầu và cuối mỗi khung vừa được chia.
* Biến đổi FFT (Fast Fourier Transform): chuyển đổi mỗi khung với N mẫu từ miền thới gian sang miền tần số.
* Biến đổi sang thang đo mel (Mel-frequency wrapping).
* Hệ số cepstrum (cepstral coefficients).

1. Mô hình Markov ẩn (HMM)

Mô hình Markov ẩn có nền tảng của mô hình thống kê được sử dụng trong hệ thống nhận dạng giọng nói hiện đại. Các biến thể của HMMs được sử dụng rộng rãi và thường được coi là thành công nhất.

Cho *µ=(S, A, Π)* gọi *X = (X1, X2, …, XT)* là một dãy các biến ngẫu nhiên với giá trị nằm trong tập *S={s1, s2, s3, …, sN}* - tập không gian trạng thái, thỏa các tính chất sau:

* Xác xuất chuyển trạng thái độc lập với thời gian.
* Xác suất chuyển trạng thái (lưu vào ma trận A): và .
* Xác suất trạng thái khởi động: và .

µ được mô tả như trên là một mô hình Markov.

Một HMM là một bộ 5 (S, K, Π, A, B):

* Tập trạng thái: *S={s1, s2, s3, …, sN}*
* Tập ký tự xuất: *K={k1, k2, k3, …, kM} = {1, …, M}*
* Xác suất chuyển trạng thái ban đầu:
* Ma trận xác suất chuyển trạng thái:
* Xác suất sinh ký hiệu:
* Dãy trạng thái:
* Dãy xuất:

1. Các bài toán với mô hình Markov ẩn

Bài toán đánh giá: cho trước chuỗi quan sát tương ứng với mô hình , xác định xác suất tương ứng có điều kiện .

Bài toán so khớp: cho trước chuỗi quan sát tương ứng với mô hình , xác định chuỗi trạng thái nghiên cứu tương ứng sao cho xác suất tương ứng đạt cực đại.

Bài toán huấn luyện: cho trước chuỗi quan sát , xác định các tham số phù hợp cho mô hình sao cho xác suất tương ứng co điều kiện là cực đại.

# Thực nghiệm

1. Dữ liệu thực nghiệm

Dữ liệu huấn luyện là 270 tập tin wav (mỗi tập tin có độ dài 2-4 giây). Mỗi tập tin wav đi kèm với một tập tin txt ghi nhận lại câu đã nói.

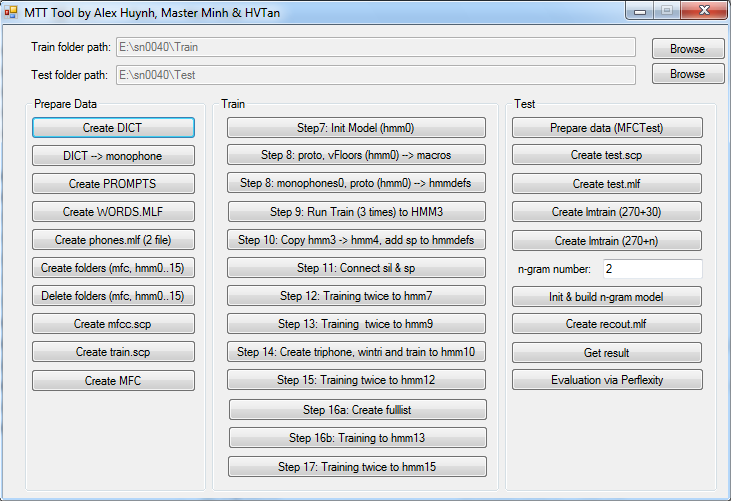
Dữ liệu để kiểm thử là 30 tập tin (mỗi tập tin có độ dài 2-4 giây) còn lại.

Mục tiêu nhận dạng là từng câu và từng từ xem các từ các tập tin kiểm thử có được nhận dạng chính xác hay không.

Nhóm sẽ thực hiện hai thực nghiệm: thực nghiệm 1 với text là 300 dòng của bộ train và test; thực nghiệm 2 với text là 270 dòng của bộ train và 140 dòng thu thập thêm từ các bài báo. Với mỗi thực nghiệm nhóm sẽ tiến hành kiểm tra bằng mô hình bigram và trigram, sau đó sẽ so sánh kết quả.

Công cụ sử dụng chính là HTK (Hidden Markov Model Toolkit).

Ngoài ra để thuận tiện cho việc tạo lập các tập tin chính, nhóm đã viết một công cụ hỗ trợ MTT Tool bằng ngôn ngữ C# để thao tác nhanh hơn. Công cụ này có thể thực hiện tự động tạo ra các tập tin cần thiết, chạy tất cả các lệnh HTK để cuối cùng hiển thị ra kết quả nhận dạng.



1. HTK

HTK cung cấp các công cụ cho việc xây dựng một mô hình Markov ẩn (HMMs). HMMs có thể dùng cho các bài toán với dữ liệu là một chuỗi theo thời gian, và HTK cũng có thể sử dụng cho các bài toán dạng này. Tuy nhiên, HTK được thiết kế chính là phục vụ cho các bài toán về tiếng nói. Dưới đây là một vài công cụ cơ bản thường được dùng:

* HCopy: copy và thức hiện việc rút trích đặc trưng tùy thuộc vào file cấu hình.
* HCompV: công cụ khởi tạo tham số cho các hàm phân bố xác suất có trong từng trạng thái của mô hình Markov ẩn tương ứng với dữ liệu huấn luyện.
* HRest: công cụ huấn luyện mô hình bằng việc tính toán, ước lượng lại các tham số của mô hình.
* HParse: công cụ chuyển đổi từ dữ liệu đầu vào thành dữ liệu theo format HTK.
* HVite: công cụ so khớp/nhận dạng với đầu vào là dữ liệu cần nhận dạng, và trả về mô hình tốt nhất/khớp nhất mà nó thuộc vào.
* HResults: công cụ thống kê, đánh giá kết quả nhận dạng

1. Phương pháp thực hiện
2. tạo từ điển phiên âm DICT

* Từ điển này giúp máy biết cách phát âm từng từ.
* Sử dụng công cụ MTT Tool để tạo.

1. tạo PROMPTS

* Lưu đường dẫn và câu tương ứng.
* Sử dụng công cụ MTT Tool để tạo.

1. tạo WORDS.MLF

* Danh sách các từ trong một câu và nhãn của chúng.
* Sử dụng công cụ MTT Tool để tạo.

1. tạo monophones0 và monophones1

* Thống kê các từ.
* Sử dụng công cụ MTT Tool để tạo.
* Ở hai tập tin đều có sil. monophones1 khác với monophones0 là có thêm sp ở cuối tập tin. sp là khoảng ngưng giữa các từ, sil là khoảng lặng ở đầu/cuối tập tin ghi âm.

1. tạo phones0.mlf và phones1.mlf

* Thống kê các từ.
* Dùng lệnh HTK:

*HLEd -l \* -d dict -i phones0.mlf mkphone0.led words.mlf*

*HLEd -l \* -d dict -i phones1.mlf mkphone1.led words.mlf*

* mkphones0.led và mkphones1.led là hai tập tin cho trước dùng để cấu hình, có các từ khóa EX (expand), IS (thêm khoảng lặng đầu cuối), DE (xóa tất cả khoảng nghỉ).

1. tạo mfcc.scp, rút trích đặt trưng

* Danh sách tập tin wav và tên mfc tương ứng.
* Sử dụng công cụ MTT Tool để tạo.
* Rút trích đặc trưng bằng lệnh HTK:

*HCopy -T 1 -C config.hcopy -S mfcc.scp*

* HTK sẽ tạo ra các tập tin \*.mfc trong thư mục MFC.

1. khởi tạo mô hình

* Khởi tạo mô hình bằng lệnh HTK:

*HCompV -C config -f 0.01 -m -S train.scp -M hmm0 proto*

* proto là tập tin được cung cấp trước cho biết rằng có 39 đặc trưng MFCC\_0\_D\_A.
  + 0: 12 MFCC features và (1) total energy in the frame
  + D: 13 “Delta coefficients” (đạo hàm bậc 1 của MFCC)
  + A: 13 “Acceleration coefficients” (đạo hàm bậc 2 của MFCC)
  + Mỗi một âm tiết là một mô hình Markov ẩn.
  + Số trạng thái: 5.

1. Từ hmm0, tạo mô hình chuẩn

* Tạo tập tin macros: kết hợp đoạn ~o của tập tin hmm0\proto và ~v của tập tin hmm0\ vFloors.
* Tạo tập tin hmmdefs: với mỗi dòng trong monophones0, kết hợp thêm đoạn <BEGINHMM> và <ENDHMM> của hmm\proto.
* macros và hmmdefs là mô hình chuẩn.

1. Training 3 vòng để ra hmm3

* Huấn luyện bằng lệnh HTK:

*HERest -C config -I phones0.mlf -S train.scp -H hmm0/macros -H hmm0/hmmdefs -M hmm1 monophones0*

*HERest -C config -I phones0.mlf -S train.scp -H hmm1/macros -H hmm1/hmmdefs -M hmm2 monophones0*

*HERest -C config -I phones0.mlf -S train.scp -H hmm2/macros -H hmm2/hmmdefs -M hmm3 monophones0*

* macros và hmmdefs đã được sửa lại, lưu vào hmm3.

1. Thêm sp vào danh sách mô hình

* Tạo hmm4, copy mô hình của hmm3 vào hmm4.
* Sửa hmmdefs trong hmm4: bổ sung sp, chỉnh sửa NUMSTATES thành 3, chỉ giữ lại state 2, TRANSP xóa bỏ 2 cột và 2 dòng cuối.

1. Liên kết sil và sp

* Tạo hmm5, thực hiện lệnh HTK:

*HHEd -H hmm4/macros -H hmm4/hmmdefs -M hmm5 sil.hed monophones1*

1. Training 2 vòng để ra hmm7

* Huấn luyện bằng lệnh HTK:

*HERest -C config -I phones0.mlf -S train.scp -H hmm5/macros -H hmm5/hmmdefs -M hmm6 monophones1*

*HERest -C config -I phones0.mlf -S train.scp -H hmm6/macros -H hmm6/hmmdefs -M hmm7 monophones1*

1. Training 2 vòng để ra hmm9

* Sao chép phones1.mlf thành aligned.mlf để xác định các đọc các từ trong tập tin wav.
* Huấn luyện bằng lệnh HTK:

*HERest -C config -I aligned.mlf -S train.scp -H hmm7/macros -H hmm7/hmmdefs -M hmm8 monophones1*

*HERest -C config -I aligned.mlf -S train.scp -H hmm8/macros -H hmm8/hmmdefs -M hmm9 monophones1*

1. Cải tiến mô hình

* Tạo triphones1 và wintri.mlf bằng lệnh HTK:

*HLEd -n triphones1 -i wintri.mlf mktri.led aligned.mlf*

* Tạo hmm10 và huấn luyện bằng HTK:

*HHEd -B -H hmm9/macros -H hmm9/hmmdefs -M hmm10 mktri.hed monophones1*

1. Training 2 vòng để ra hmm12

* Sửa file wintri.mlf: Thêm \*/ trước tất cả các đoạn text <file\_name>.lab. Bước này nhằm tránh lỗi khi chạy lệnh HERest ở bước tiếp theo.
* Huấn luyện bằng lệnh HTK:

*HERest -B -C config -I wintri.mlf -s stats -S train.scp -H hmm10/macros -H hmm10/hmmdefs -M hmm11 triphones1*

*HERest -B -C config -I wintri.mlf -s stats -S train.scp -H hmm11/macros -H hmm11/hmmdefs -M hmm12 triphones1*

1. Liên phần chung để tạo ra hmm13

* Chép file *tree.hed* vào thư mục thực thi.
* Tạo tập tin fulllist từ monophones1. Đây là tập tin chứa crossword triphone.
* Huấn luyện bằng lệnh HTK

*HHEd -B -H hmm12/macros -H hmm12/hmmdefs -M hmm13 config/tree.hed triphones1 > log*

1. Training 2 vòng để ra hmm15

* Huấn luyện bằng lệnh HTK

*HERest -B -C config/config -I wintri.mlf -s stats -S train.scp -H hmm13/macros -H hmm13/hmmdefs -M hmm14 tiedlist*

*HERest -B -C config/config -I wintri.mlf -s stats -S train.scp -H hmm14/macros -H hmm14/hmmdefs -M hmm15 tiedlist*

* Tạo thư mục LM, file lmtrain.txt chứa tất cả câu (train + test).
* Chạy lệnh HTK để khởi tạo mô hình ngôn ngữ n-gram:

*LNewMap –f WFC LMName empty.wmap*

* Chạy lệnh HTK để khởi tạo số liệu cho mô hình vừa tạo với tham số n là 3 (trigram):

*LGPrep –T 1 –a 100000 -b 200000 -n 3 -d lm empty.wmap lmtrain.txt*

* Huấn luyện bằng lệnh HTK với tham số n là 3 (trigram), c là 0 (lấy hết tất cả các bộ):

*LBuild –T 1 -c 2 0 -c 3 0 -n 3 lm/wmap lmModel lm/gram.0*

1. Chuẩn bị dữ liệu test

* Tạo tập tin mfcc-test.scp và test.scp.
* Rút trích đặc trưng bằng lệnh HTK:

*HCopy -T 1 -C config/config.hcopy -S mfcc-test.scp*

* Nhận dạng bằng lệnh HTK:

*HDecode -H hmm15/macros -H hmm15/hmmdefs -S test.scp -t 220.0 220.0 -C config/config.hdecode -i recout.mlf -w lmModel -p 0.0 -s 5.0 dict tiedlist*

* Tạo tập tin test.mlf để so sánh với kết quả nhận dạng từ recout.mlf. test.mlf là danh sách các tập tin test và từ chính xác được phát âm.
* Hiển thị kết quả so sánh bằng lệnh HTK:

*HResults -I test.mlf tiedlist recout.mlf*

* Tạo tập tin lmtest.txt gồm các câu test.
* Chạy lệnh HTK để kiểm tra perplexity:

*Lplex -n 3 -t lmModel lmtest.txt*

1. Thống kê kết quả thực nghiệm 1 (text 270+30)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kết quả xét trên câu (bigram, n=2) | | | | |
| Mẫu | H | S | N | Correct |
| sn0001 | 1 | 29 | 30 | 3.33% |
| sn0040 | 0 | 30 | 30 | 0% |
| sn0041 |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kết quả xét trên câu (trigram, n=3) | | | | |
| Mẫu | H | S | N | Correct |
| sn0001 | 1 | 29 | 30 | 3.33% |
| sn0040 | 0 | 30 | 30 | 0% |
| sn0041 |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Kết quả xét trên từ (bigram, n=2) | | | | | | | |
| Mẫu | H | D | S | I | N | Correct | Accuracy |
| sn0001 | 52 | 68 | 71 | 0 | 191 | 27.23% | 27.23% |
| sn0040 | 81 | 8 | 170 | 0 | 259 | 31.27% | 31.27% |
| sn0041 |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Kết quả xét trên từ (trigram, n=3) | | | | | | | |
| Mẫu | H | D | S | I | N | Correct | Accuracy |
| sn0001 | 54 | 69 | 68 | 0 | 191 | 28.27% | 28.27% |
| sn0040 | 79 | 7 | 173 | 0 | 259 | 30.50% | 30.50% |
| sn0041 |  |  |  |  |  |  |  |

Correct là độ chính xác của câu, Corr là độ chính xác từ.

H: số mẫu nhận dạng chính xác.

S: số mẫu nhận dạng sai.

N: tổng số mẫu nhận dạng.

D: các lỗi xóa.

I: các lỗi chèn.

Acc là độ chính xác, không xét các từ dư ra trong câu. Nếu không có lỗi chèn (I=0) thì Acc và Corr bằng nhau.

|  |  |  |
| --- | --- | --- |
| Perplexity | | |
| n | bigram | trigram |
| sn0001 | 45.1310 | 41.9588 |
| sn0040 | 36.5275 | 38.2572 |
| sn0041 |  |  |

1. Thống kê kết quả thực nghiệm 2 (text 270+n)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kết quả xét trên câu (bigram, n=2) | | | | |
| Mẫu | H | S | N | Correct |
| sn0001 | 1 | 29 | 30 | 3.33% |
| sn0040 | 0 | 30 | 30 | 0% |
| sn0041 |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kết quả xét trên câu (trigram, n=3) | | | | |
| Mẫu | H | S | N | Correct |
| sn0001 | 1 | 29 | 30 | 3.33% |
| sn0040 | 0 | 30 | 30 | 0% |
| sn0041 |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Kết quả xét trên từ (bigram, n=2) | | | | | | | |
| Mẫu | H | D | S | I | N | Correct | Accuracy |
| sn0001 | 48 | 60 | 83 | 0 | 191 | 25.13% | 25.13% |
| sn0040 | 62 | 9 | 188 | 0 | 259 | 23.94% | 23.94% |
| sn0041 |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Kết quả xét trên từ (trigram, n=3) | | | | | | | |
| Mẫu | H | D | S | I | N | Correct | Accuracy |
| sn0001 | 47 | 61 | 83 | 0 | 191 | 24.61% | 24.61 |
| sn0040 | 60 | 10 | 189 | 0 | 259 | 23.17% | 23.17% |
| sn0041 |  |  |  |  |  |  |  |

|  |  |  |
| --- | --- | --- |
| Perplexity | | |
| n | bigram | trigram |
| sn0001 | 347.1678 | 342.3504 |
| sn0040 | 367.3414 | 339.6240 |
| sn0041 |  |  |

Nhận xét:

* Mô hình bigram cho kết quả cao hơn so với trigram.
* Kết quả nhận dạng trên một câu rất thấp, gần như 0%. Một phần lí do do giọng phát âm chưa rõ ràng. Mô hình cần cải tiến để có kết quả cao hơn.
* Mẫu sn0041 có kết quả nhận dạng thấp nhất trong 3 mẫu do perplexity cao nhất. Perplexity càng thấp thì kết quả nhận dạng càng cao.

# Tài liệu tham khảo

Tiếng Anh:

1. Daniel Jurafsky & James H. Martin, *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*, 2006, pp 192-236.
2. Huỳnh Diệp Tân, et al., *Mô hình Markov*, 2010
3. Nguyễn Bá Công, *Nhận dạng tiếng nói rời rạc*, 2009
4. Steve Young et al., *The HTK Book*, 2009

Website:

1. <http://en.wikipedia.org/wiki/Hidden_Markov_model>
2. <http://htk.eng.cam.ac.uk/>