PHƯƠNG PHÁP I-VECTOR TRONG NHẬN DẠNG NGƯỜI NÓI

Phùng Thị Thu Hiền*
Trường Đại học Kỹ thuật Công nghiệp – ĐH Thái Nguyên

TÓM TẮT

Nhận dạng người nói ngày càng có nhiều ứng dụng mang ý nghĩa thực tế, đặc biệt là các ứng dụng trong nhận diện người nói và xác thực người nói. Tuy nhiên việc nâng cao chất lượng của các ứng dụng này là điều cần quan tâm nghiên cứu. Bài báo này trình bày tổng quan về phương pháp nhận dạng người nói SR (Speaker recognition), các thuật toán nhận dạng người nói dựa trên mô hình GMM (Gaussian mixture model), mô hình GMM-UBM (Gaussian mixture model - *Universal Background Model*) và Phương pháp phân tích JFA – *Joint factor analysis.* Đặc biệt bài báo trình bày về phương pháp I-vector, phương pháp này sử dụng một tập các nhân tố có tổng số chiều thấp nên giúp làm tăng hiệu quả của phương pháp nhận dạng người nói, sau đó tiến hành kiểm chứng lại phương pháp i-vector trên bộ dữ liệu *NIST 2008 SRE.*

Từ khóa: *Phương pháp i-vector; mô hình GMM; nhận dạng người nói SR; kỹ thuật phân tích FA; thích nghi UBM; Phương pháp trích chọn đặc trưng MFCC.*

ĐẶT VẤN ĐỀ

Con người có khả năng nhận diện ra giọng nói của một người chỉ trong một vài giây, tuy nhiên làm thế nào để máy tính có thể làm được như vậy là một vấn đề cần nghiên cứu.

Hệ thống nhận dạng người nói (SR) gồm có hai lĩnh vực chính là nhận diện người nói (identification) và xác thực người nói (verification) [1]. Trong đó, nhận dạng người nói là nhằm trả lời câu hỏi giọng nó này của ai, còn xác thực người nói nhằm trả lời câu hỏi giọng nói này có phải của người A không? Cả hai hệ thống nhận dạng người nói và xác thực người nói đều sử dụng mô hình GMM.

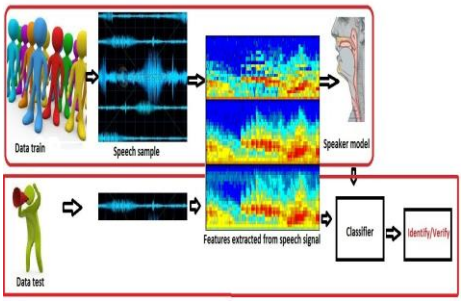
Ứng dụng của hệ thống nhận dạng người nói được áp dụng trong rất nhiều lĩnh vực khác nhau như trong lĩnh vực xác thực bảo mật của ngân hàng, thẻ tín dụng, hệ thống cá nhân,...

Đã có rất nhiều nghiên cứu về nhận dạng tiếng nói đặc biệt là về nhận dạng người nói như trong các nghiên cứu của nhóm tác giả N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, và P. Ouellet trong [2], N. Dehak và các cộng sự trong [3]. Trong các bài báo này, các tác giả đã đưa ra, phân tích và đánh giá về phương pháp i-vector trong bài toán xác thực người nói.

Tuy nhiên, các nghiên cứu về i-vector tại Việt Nam và việc ứng dụng của I-vector vào bài toán nhận dạng người nói tại Việt Nam còn rất hạn chế. Trong bài báo này, sẽ trình bày tổng quan về nhận dạng người nói, đưa ra một số thuật toán nhận dạng người nói trong mô hình GMM và cuối cùng là thực nghiệm lại phương pháp i-vector trong bài toán nhận dạng người nói.

TỔNG QUAN VỀ NHẬN DẠNG NGƯỜI NÓI VÀ CÁC THUẬT TOÁN NHẬN DẠNG NGƯỜI NÓI

Tổng quan về nhận dạng người nói



Hình 1. *Mô hình người nói*

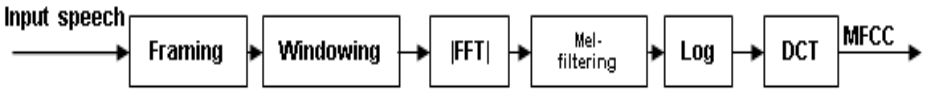
Trong mô hình người nói (speaker model), GMM được tạo ra bằng cách thích nghi UBM, sau đó kết hợp các thành phần trung bình của GMM để tạo ra các supervector và cuối cùng sử dụng kỹ thuật FA (Factor analysis) để tạo ra các i-vector có số chiều thấp.

* *Email: phungthuhien@tnut.edu.vn*

Phương pháp trích chọn đặc trưng MFCC – Mel Frequency Cepstrum Coefficients

MFCC là đặc trưng thường được dùng để diễn tả âm thanh tiếng nói. Theo Beth Logan [4], MFCC gồm 5 bước (Xem minh họa trên hình 2):

1. Chia tín hiệu thành các khung
2. Với mỗi khung, ta thu được biên độ phổ.
3. Lấy log của biên độ
4. Chuyển đổi sang thang Mel
5. Thực hiện biến đổi Cosine rời rạc.



Hình 2. *Quá trình tạo các đặc tính MFCC*

Quan sát quá trình trên ta thấy, âm thanh được chia thành những khung có độ dài cố định. Mục đích là để lấy mẫu những đoạn tín hiệu nhỏ. Trong việc lấy mẫu dữ liệu, chúng ta xem xét đến tín hiệu âm thanh đã được số hóa bằng việc rời rạc hóa các giá trị trên những khoảng đều nhau vì vậy cần phải chắc chắn rằng tốc độ lấy mẫu là đủ lớn để mô tả tín hiệu dạng sóng. Tần số lấy mẫu nên ít nhất gấp đôi tần số dạng sóng như trong định lý của Nyquist.

Phân khung là quá trình chia mẫu tín hiệu thành một số khung chồng lấp lên nhau hoặc không. Mục đích của phân khung là để lấy mẫu các đoạn tính hiệu nhỏ. Vấn đề là bản chất của âm thanh là không ổn định. Vì vậy, biến đổi Fourier sẽ thể hiện tần số xảy ra trên toàn miền thời gian thay vì thời gian cụ thể. Bởi thế khi tín hiệu là không ổn định, tín hiệu đó nên được chia nhỏ thành các cửa sổ rời rạc nhờ đó mỗi tín hiệu trong một cửa sổ trở nên tĩnh và phép biến đổi Fourier có thể thực hiện trên mỗi khung.

Hàm cửa sổ bỏ đi những hiệu ứng phụ và vector đặc trưng cepstral được thực hiện trên mỗi khung cửa sổ. Ý tưởng ở đây là giảm bớt sự méo phổ bằng việc sử dụng các cửa sổ để giảm tín hiệu về không tại điểm bắt đầu và kết thúc mỗi khung.

Biến đổi Fourier rời rạc của mỗi khung được tính toán và lấy logarithm biên độ phổ. Thông tin về pha bị bỏ qua do biên độ phổ là quan trọng hơn pha. Thực hiện lấy logarithm biên độ phổ do âm lượng của tín hiệu là xấp xỉ logarith. Tiếp theo biến đổi phổ theo thang Mel. Từ kết quả này, trong vector Mel – spectral của các thành phần tương quan cao, bước cuối cùng là thực hiện biến đổi cosine rời rạc để tổng hợp vector phổ Mel để tương quan lại các thành phần này

Các thuật toán được sử dụng trong phương pháp nhận dạng người nói dựa trên GMM.

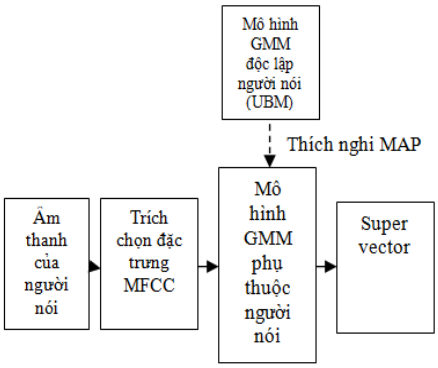
Mô hình GMM

Mô hình GMM biểu diễn mật độ xác suất của một biến ngẫu nhiên theo tổng trọng số của các thành phần, được mô tả bởi hàm mật độ Gaussian. Mô hình GMM cho phép biểu diễn được số lượng rất lớn những mô hình phân phối khác nhau tương ứng với những người nói khác nhau, ta sử dụng GMM để mô hình hóa nhiều người nói khác nhau.

$$p(x \mid s_i) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \sum_k) \quad (1)$$

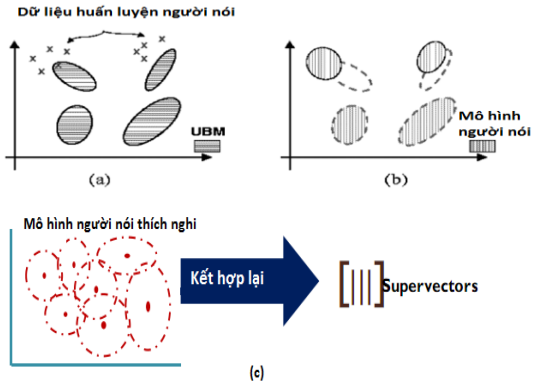
Với μ_k là vector trung bình, $\sum_k$ là ma trận hiệp phương sai, π_k là trọng số hỗn hợp.

Mô hình GMM-UBM



Hình 3. *Mô hình GMM – UBM*

Khâu huấn luyện của mô hình UBM (Universal Background Model) hay mô hình người nói trung bình được thực hiện bằng cách sử dụng thuật toán EM (Expectation Maximization). Mô hình người nói học theo bằng cách sử dụng thuật toán MAP (Maximum A Posterior). Thích nghi MAP là một nội suy tuyến tính của tất cả các thành phần hỗn hợp của UBM để tạo ra tiếng nói gần đúng từ mỗi người nói riêng. Các supervector bao gồm các thành phần trung bình GMM phụ thuộc người nói.



a. Các vector huấn luyện (x) được ánh xạ xác suất trong UBM
b. Chia các tham số hỗn hợp thích nghi bằng cách thống kê dữ liệu mới và các tham số hỗn hợp UBM
c. Tạo ra các supervector từ mô hình UBM

Hình 4. *Tạo các supervector từ mô hình UBM*

Từ hình vẽ 4, giả sử k=3, các suppervector sẽ có dạng như sau:

$$\alpha 1, \mu 1 = \begin{bmatrix} \mu_{11} \\ \mu_{12} \end{bmatrix}, \Sigma_1$$

$$\alpha_2, \mu_2 = \begin{bmatrix} \mu_{21} \\ \mu_{22} \end{bmatrix}, \Sigma_2 \longrightarrow \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \\ \mu_{31} \\ \mu_{32} \end{bmatrix}$$

$$\alpha_3, \mu_3 = \begin{bmatrix} \mu_{31} \\ \mu_{32} \end{bmatrix}, \Sigma_3$$

Supervector

Thuật toán EM – Expectation Maximization

Thuật toán EM giúp tìm ra các tham số của mô hình bằng cách tối đa hóa log-likelihood trong tập dữ liệu không đầy đủ và bằng cách tối đa lại kỳ vọng của log-likelihood từ tập dữ liệu đầy đủ.

Thuật toán EM được chia làm hai bước E_step và M_step, như sau:

Bước E_step: Tính toán hàm phụ

$$\gamma_t(c) = p(c \mid x_t, s) = \frac{\pi_c N(x \mid \mu_c, \sum_c)}{\sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)} \quad (2)$$

Bước M_step: Tối đa hóa hàm phụ với các thông số được xác định theo công thức sau:

- Trọng số hỗn hợp:

$$\pi_c = \frac{1}{T} \sum_{t=1}^{T} \gamma_t(c) \quad (3)$$

- Các vector trung bình:

$$\mu_c = \frac{\sum_{t=1}^{T} \gamma_t(c) x_t}{\sum_{t=1}^{T} \gamma_t(c)} \quad (4)$$

- Hiệp phương sai:

$$\sigma_c = \frac{\sum_{t=1}^{T} \gamma_t(c) x_t^2}{\sum_{t=1}^{T} \gamma_t(c)} - \mu_c^2 \quad (5)$$

Lặp lại bước E_step cho đến khi hội tụ.

Phương pháp phân tích JFA – Joint factor analysis

Trong JFA, một câu nói của người nói được mô tả bởi một supervector, bao gồm các thành phần độc lập người nói, các thành phần phụ thuộc người nói, các kênh phụ thuộc và các yếu tố còn lại. Mỗi thành phần được đại diện bởi một tập các nhân tố có số chiều thấp. Một vector GMM được mô tả như sau:

$$s = m + V_y + U_x + D_z \quad (6)$$

Với: - s: supervector của người nói

- m: supervector độc lập người nói

- V_y: thành phần phụ thuộc người nói, V: Ma trận eigenvoice, y: các nhân tố người nói

- U_x : thành phần phụ thuộc vào channel, U: ma trận eigenchannel, x: các nhân tố channel

- D_z: Thành phần còn lại phụ thuộc người nói, z: các nhân tố còn lại đặc trưng cho người nói

Các bước huấn luyện JFA như sau:

- Bước 1: Huấn luyện ma trận eigenvoice V, với giả sử U và D bằng 0

- Bước 2: Huấn luyện ma trận eigenchannel U từ V, giả sử D bằng 0

- Bước 3: Huấn luyện ma trận D từ U và V

Sử dụng các ma trận này để tính y (người nói), x (kênh), z (các yếu tố còn lại)

Phương pháp I – vector

Từ công thức (6), ta có công thức tính supervector:

$$s = m + T_w \quad (7)$$

Trong đó:

- s: supervector của người nói

- m: supervector độc lập người nói

- T_w: Ma trận tổng biến thiên, T_w được gọi là i-vector

Hệ thống i-vector sử dụng một tập các nhân tố biến thiên có tổng số chiều thấp (w) để mô tả cho một lời nói. Phương pháp tính i-vector [2]: Tính toán dựa trên thống kê của Baum-welch cho mỗi người nói với các đặc trưng âm thanh $x_1, x_2, \ldots, x_T$ cho mỗi thành phần c.

- Bước 0: $$N_c(s) = \sum_{t=1}^{T} \gamma_t(c) \quad (8)$$

- Bước 1: $$F_c(s) = \sum_{t=1}^{T} \gamma_t(c) x_t \quad (9)$$

- Bước 2:

$$S_c(s) = diag(\sum_{t=1}^{T} \gamma_t(c) x_t^*) \quad N_c(s) = \sum_{t=1}^{T} \gamma_t(c) \quad (10)$$

Từ (8), (9), (10) có:

$$\tilde{F}_c(s) = F_c(s) - N_c(s) m_c \quad (11)$$

Với m_c là giá trị trung bình UBM cho thành phần hỗn hợp c.

$$\tilde{S}_c(s) = S_c(s)(diag(F_c(s) m_c^* + m_c F_c(s)^* - N_c(s) m_c m_c^*)) \quad (12)$$

Giá trị i-vector cho mỗi người nói được tính theo công thức sau:

$$w(s) = l_T^{-1}(s) T^* \Sigma^{-1} \tilde{F}(s) \quad (13)$$

Với $\sum^{-1} \tilde{F}(s)$ là nghịch đảo của ma trận hiệp phương sai và $l_T(s)$ được xác định bởi công thức sau:

$$l_T(s) = I + T^* \Sigma^{-1} NN(s) T \quad (14)$$

MÔ PHỎNG, TÍNH TOÁN, THẢO LUẬN

Số liệu đầu vào

Nghiên cứu này sử dụng cơ sở dữ liệu NIST 2008 SRE. Cơ sở dữ liệu NIST 2008 SRE gồm các utterances 2s, 4s, 8s, 10s, 20s, 50s và 2.5 phút(full).

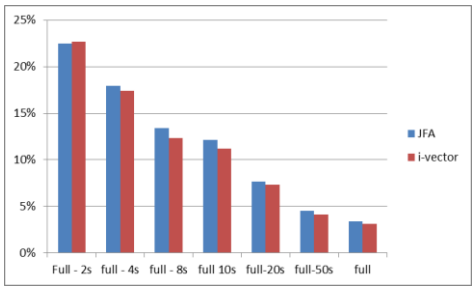
Phương pháp, công cụ mô phỏng

Với cơ sở dữ liệu trên, lựa chọn các tham số để tính toán MFCC gồm: độ dài của khung là 25ms, frame shift là 10ms, số lượng các hệ số cepstral là 19, sử dụng log energy, hệ số delta và delta-delta. Tổng số chiều sử dụng là 60.

Tiếp theo, tác giả mô hình hóa bằng mô hình GMM-UBM, áp dụng phương pháp JFA và cuối cùng là áp dụng phương pháp i-vector để thu được các i-vector trên bộ dữ liệu huấn luyện và bộ dữ liệu kiểm thử.

Kết quả mô phỏng và bình luận

Kết quả sau khi thực nghiệm thu được:



Hình 5. *Biểu đồ tỷ lệ lỗi EER của phương pháp chỉ sử dụng JFA và phương pháp sử dụng i-vector*

Từ hình 5 ta thấy, phương pháp i-vector có tỷ lệ lỗi EER (equal errol rate) thấp hơn phương pháp chỉ sử dụng JFA đơn thuần. Do đó, mặc dù số chiều của i-vector giảm xuống nhưng vẫn đáp ứng rất tốt cho hệ thống nhận dạng người nói.

KẾT LUẬN

Phương pháp i-vector giúp làm giảm số chiều của vector đặc trưng. Tuy nhiên trong phạm vi bài báo này tác giả mới dừng ở mức dựa trên lý thuyết để thực nghiệm tính toán ra được giá trị của i-vector.

Trong các nghiên cứu tiếp theo, tác giả sẽ tiếp tục phát triển các ứng dụng cụ thể dựa trên giá trị của i-vector đã tính toán được. Từ đó, sẽ có những đánh giá đầy đủ, đúng đắn về phương pháp này.

***Lời cảm ơn**: Tác giả xin chân thành cảm ơn sự hỗ trợ, giúp đỡ của GS. Tomoki Toda của viện nghiên cứu Naist Nhật bản đã hướng dẫn tác giả thực hiện bài báo này đồng thời cấp bản quyền cho tác giả được dùng cơ sở dữ liệu NIST 2008 SRE.*

TÀI LIỆU THAM KHẢO

1. Douglas A. Reynolds (1995), "Speaker identification and verification using Gaussian mixture speaker model", *Speech Communication* 17, 91 – 108.

2. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet (2010), "Front-End Factor Analysis For Speaker Verification" *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788 – 798.

3. N. Dehak et al, (2009) *"Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification,"* in Interspeech 2009, Brighton, UK.

4. Beth Logan and Ariel Salomon, 2001, "A Music Similarity Function Based on Signal Analysis", *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference Japan*, 10.1109/ICME.2001.1237829

SUMMARY

I-VECTOR METHOD IN SPEAKER RECOGNITION

Phung Thi Thu Hien*

University of Technology - TNU

Speaker recognition is becoming more and more practical, especially in the areas of speaker identification and speaker verification. However, improving the quality of these applications need to study. In this paper, I present an overview of Speaker recognition (SR), some algorithms for Gaussian mixture model (GMM), Gaussian mixture model - Universal Background Model (GMM-UBM), Joint factor analysis (JFA). Expencially, I present about I-vector approach that uses a set of low-dimentional total variability factors to represent each conversation side, and then I verify the method I -vector on the NIST 2008 SRE datasets.

Keywords: *i-vector, gaussian mixture model GMM, speaker recognition SR, factor analysis - FA, universal Background Model UBM, mel frequency cepstrum coefficients MFCC.*

Ngày nhận bài: 01/9/2017; Ngày phản biện: 14/9/2017; Ngày duyệt đăng: 16/10/2017

* *Email: phungthuhien@tnut.edu.vn*