ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐỖ ĐỨC HÀO

PHÂN TÍCH TÍN HIỆU ĐA PHÂN GIẢI VÀ ÁP DỤNG VÀO XỬ LÝ TIẾNG NÓI

Ngành: Khoa học máy tính

Mã số ngành: 9480101

TÓM TẮT LUẬN ÁN TIẾN SĨ

TP. Hồ Chí Minh – Năm 2025

Công trình được hoàn thành tại: Trường Đại Học Khoa Học Tự Nhiên – Đại Học Quốc Gia Thành Phố Hồ Chí Minh

Người hướng dẫn khoa học:
1. HDC: TS. Trần Thái Sơn
2. HDP: TS. Châu Thành Đức

Phản biện 1: PGS. TS. Hồ Văn Khương

Phản biện 2: PGS. TS. Huỳnh Trung Hiếu

Phản biện 3: PGS. TS. Phạm Thế Bảo

Phản biện độc lập 1: Miễn

Phản biện độc lập 2: Miễn

Luận án sẽ được bảo vệ trước Hội đồng chấm luận án cấp cơ sở đào tạo họp tại Trường Đại học Khoa học Tự nhiên, vào hồi … giờ …. ngày …. tháng … năm ….

Có thể tìm hiểu luận án tại thư viện:
- Thư viện Khoa học Tổng hợp TP.HCM
- Thư viện Đại học Quốc gia TP.HCM
- Thư viện Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM

Chương 1

Tổng quan về luận án

1.1 Giới thiệu

1.1.1 Khái niệm đa phân giải và phân tích đa phân giải

Phân tích đa phân giải (Multi-Resolution Analysis - MRA) hay phân tích tín hiệu đa phân giải (Multi-Resolution Signal Analysis - MSA) là một phương pháp trong xử lý tín hiệu số nhằm phân tích các tín hiệu ở nhiều mức độ chi tiết (độ phân giải) khác nhau. Phương pháp này giúp khám phá các đặc điểm tiềm ẩn bên trong tín hiệu có thể đã bị bỏ qua khi sử dụng các phương pháp phân tích truyền thống. MSA được ứng dụng rộng rãi trong nhiều lĩnh vực như xử lý hình ảnh, nén tín hiệu, xử lý âm thanh.

Trong luận án này, quá trình MSA tập trung mô hình hoá hai khía cạnh chính của tín hiệu: sự ổn định và sự biến đổi theo nhiều mức độ khác nhau. Việc mô hình hoá hai khía cạnh này của tín hiệu theo thời gian đều đóng những vai trò quan trọng, đặc biệt là trong các bài toán về xử lý tiếng nói. Khía cạnh ổn định của tín hiệu liên quan đến tính nhất quán và khả năng duy trì xuyên suốt của các đặc tính theo thời gian, đảm bảo thông tin liên tục và chính xác. Ngược lại, sự biến đổi tín hiệu biểu thị những thay đổi dần dần hoặc đột ngột về tần số, biên độ và các thuộc tính khác, phản ánh sự đa dạng và phong phú của thông tin được mã hóa trong tín hiệu.

Luận án đề xuất mô hình hoá cả khía cạnh ổn định lẫn khía cạnh biến đổi của dữ liệu theo tiếp cận MSA. Nói cách khác, sự khác biệt của MSA trong luận án so với các mô hình MSA thông thường có thể được tóm tắt như sau: MSA sẽ xác định giá trị tức thời và xu hướng biến đổi của tín hiệu trên miền TFD theo nhiều mức độ phân giải khác nhau để mô hình hoá dữ liệu tín hiệu.

1.1.2 Phạm vi nghiên cứu của luận án

Về tổng thể, luận án tập trung nghiên cứu và phát triển phương pháp phân tích tín hiệu theo hướng đa phân giải, từ đó sử dụng các phân tích này vào giai đoạn trích xuất và biểu diễn đặc trưng cho một số bài toán trong lĩnh vực xử lý tiếng nói.

1.2 Động lực nghiên cứu

Việc nâng cao chất lượng cho đặc trưng trong học máy nói chung và xử lý tiếng nói nói riêng mang lại ý nghĩa quan trọng cả về mặt công nghệ và kinh tế. Về công nghệ, các đặc trưng chất lượng cao giúp tăng tốc độ tính toán và giảm thiểu bộ nhớ lưu trữ cần thiết, cho phép các mô hình học máy hoạt động hiệu quả hơn. Về mặt kinh tế, việc sử dụng đặc trưng chất lượng cao giúp giảm giá thành phát triển và vận hành các ứng dụng.

1.3 Mục tiêu của luận án

Luận án này hướng đến việc xây dựng một phương pháp đa phân giải theo định hướng MCT, nhưng áp dụng vào lĩnh vực xử lý tiếng

nói thay vì áp dụng vào cơ khí hay xử lý ảnh như các nghiên cứu trước đây. Từ đó, các mục tiêu cụ thể của luận án bao gồm:

Mục tiêu 1: Xây dựng mô hình làm giàu thông tin cho vector đặc trưng tiếng nói bằng phương pháp phân tích tín hiệu đa phân giải. Luận án sẽ phát triển các kỹ thuật và thuật toán mới để phân tích và xử lý tín hiệu tiếng nói ở nhiều cấp độ phân giải khác nhau, qua đó khám phá các đặc trưng tiềm ẩn trong tín hiệu mà các phương pháp truyền thống có thể bỏ sót.

Mục tiêu 2: Nâng cao khả năng tự điều chỉnh cho vector đặc trưng. Luận án đăt mục tiêu xây dựng các giải pháp có khả năng cải thiện khả năng tự điều chỉnh của quá trình trích xuất đặc trưng theo dữ liệu và bài toán cụ thể.

1.4 Đóng góp khoa học của luận án

Về tổng thể, luận án này tập trung nghiên cứu ở những giai đoạn đầu của quá trình giải quyết các bài toán nhận dạng trong lĩnh vực xử lý tiếng nói. Các công bố khoa học liên quan đến nội dung luận án được liệt kê trong Bảng 1.1. Các đóng góp về mặt khoa học của luận án bao gồm các khía cạnh như sau:

Đóng góp 1: Đề xuất thuật toán trích xuất đặc trưng tiếng nói dựa trên phép biến đổi LCT. Luận án đã sử dụng đặc trưng LCT với hệ số chirp âm vào các bài toán nhận dạng đơn giản như nhận dạng giới tính, nhận dạng vùng miền. Các công trình liên quan bao gồm CT01, CT02.

Đóng góp 2: Nâng cao khả năng chống nhiễu cho phép biến đổi Chirplet tuyến tính. Cụ thể, luận án đã kết hợp phép biến đổi LCT với bộ tự mã hoá biến phân và bộ lọc Chebyshev để lọc

Bảng 1.1: Các công bố khoa học liên quan trực tiếp đến luận án

STT	Nơi - Năm	Chỉ mục	Trạng thái
CT01	ICCCI - 2022	CORE-Rank B	Đã công bố
CT02	JIT - 2023	ISI, Scopus Q2	Đã công bố
CT03	ICCCI - 2020	CORE-Rank B	Đã công bố
CT04	JoC - 2020	Scopus Q3	Đã công bố
CT05	IEEE Access - 2020	ISI, Scopus Q1	Đã công bố
CT06	TALLIP - 2023	ISI, Scopus Q2	Đã công bố
CT07	CSSP - 2024	ISI, Scopus Q2	Đã công bố
CT08	ACIIDS - 2025	CORE-Rank B	Đã chấp nhận

nhiễu và tăng cường tiếng nói, qua đó nâng cao hiệu quả cho mô hình nhận dạng với dữ liệu đầu vào bị nhiễu. Các công trình liên quan bao gồm CT03, CT04, CT05.

Đóng góp 3: Đề xuất hai thuật toán mở rộng khả năng của LCT bằng cách sử dụng nhiều hàm tuyến tính bao gồm GLCT và CGLCT. Thứ nhất, luận án kết hợp nhiều phép biến đổi LCT độc lập để tạo thành phép biến đổi GLCT. Sau đó, luận án đã giảm chiều cho không gian đặc trưng thu được từ phép biến đổi GLCT bằng phương pháp phân tích giá trị đơn rút gọn (Truncated Singular Value Decomposition - tSVD). Công trình liên quan là CT06.

Đóng góp 4: Đề xuất hai thuật toán nâng cao khả năng biểu diễn quy luật biến đổi của tín hiệu gồm PCT và MPCT. Từ đường tần số tức thời dạng tuyến tính, luận án đề xuất sử dụng đường tần số tức thời dạng đa thức để có thể biểu diễn được sự thay đổi phức tạp của tiếng nói. Sau đó, luận án đã sử dụng mô hình gom cụm dựa trên mật độ để lựa chọn ra các hàm đa thức tốt nhất cho phép biến đổi PCT. Các công trình liên quan bao gồm CT07, CT08.

Chương 2

Một số kiến thức nền tảng

Chương 2 cung cấp những kiến thức nền tảng về xử lý tiếng nói và phân tích tín hiệu, qua đó giúp người đọc dễ nắm bắt hơn những nội dung chính ở các chương sau. Chương này có hai nội dung chính bao gồm: (1) Khung lời giải để giải quyết một bài toán trong xử lý tiếng nói và (2) Một số phương pháp phân tích tín hiệu số.

2.1 Khung lời giải cho bài toán xử lý tiếng nói

2.1.1 Khung lời giải tổng quát

Quá trình này bao gồm các giai đoạn chính: tiền xử lý dữ liệu, biểu diễn đặc trưng (bao gồm cả hai pha là trích xuất đặc trưng và biến đổi đặc trưng), nhận dạng (huấn luyện mô hình và đánh giá kết quả).

2.1.2 Giai đoạn tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu tiếng nói là bước quan trọng nhằm chuẩn bị dữ liệu cho các hệ thống nhận dạng và xử lý tiếng nói phía sau. Quá trình này gồm hai bước chính như sau: chuẩn hoá cường độ âm thanh và loại bỏ nhiễu.

2.1.3 Giai đoạn trích xuất và biểu diễn đặc trưng

Trích xuất và biểu diễn đặc trưng là một bước quan trọng trong nhiều ứng dụng xử lý tiếng nói như nhận dạng giọng nói, nhận diện cảm xúc và phân loại giọng nói. Một số phương pháp phổ biến để trích xuất và biểu diễn đặc trưng tiếng nói bao gồm ảnh phổ (Spectrogram), ảnh phổ Mel (Mel-Frequency Cepstrum - MFC), và Mel-Frequency Cepstral Coefficients (MFCC).

2.1.4 Giai đoạn huấn luyện mô hình và đánh giá hiệu quả

Trong quá trình huấn luyện, mô hình đã học từ dữ liệu thông qua quá trình tối ưu hóa hàm mất mát, nhằm giảm thiểu sai số giữa giá trị dự đoán của mô hình và giá trị thực tế. Các kỹ thuật như lan truyền ngược và giảm gradient được sử dụng để cập nhật trọng số của mô hình. Sau khi huấn luyện, mô hình được đánh giá bằng cách sử dụng tập dữ liệu kiểm tra độc lập. Các chỉ số đánh giá như độ chính xác, độ nhạy, và điểm F1 được tính toán để đo lường hiệu suất của mô hình.

2.2 Phân tích tín hiệu và phương pháp đa phân giải

Việc phân tích tín hiệu cung cấp một cái nhìn chi tiết về sự biến đổi của tần số theo thời gian, giúp hiểu rõ hơn về cấu trúc và nội dung của tín hiệu trên miền Thời gian - Tần số. Phép biến đổi Fourier là một công cụ phân tích và tổng hợp tín hiệu một cách hiệu quả giữa miền thời gian và miền tần số. Có hai dạng chính của phép biến đổi Fourier: phép biến đổi Fourier liên tục và phép biến đổi Fourier rời rạc.

Bên cạnh đó, một công cụ khác là phép biến đổi Wavelet cũng có khả năng phân tích tín hiệu rất tốt. Điểm khác biệt ở đây là WT xác

định kích thước của các miền con theo giá trị của thời gian và tần số tương ứng trong khi FT chia miền TFD thành các miền con bằng nhau. Từ đó, WT có thể phân tích các tín hiệu ở nhiều mức độ phân giải khác nhau, ứng với nhiều mức độ chi tiết khác nhau trên cả miền thời gian lẫn miền tần số.

2.3 Phép biến đổi Chirplet

2.3.1 Ý tưởng về phép biến đổi Chirplet

Phép biến đổi Wavelet tuy có khả năng biểu diễn được cả các thông tin cục bộ, ứng với độ phân giải cao, và cả các thông tin toàn cục, ứng với độ phân giải thấp, tuy nhiên các thông tin này đều chỉ thể hiện trên một thông tin thời gian và tần số xác định. Từ đó phát sinh nhu cầu về một phép biến đổi khác, có khả năng biểu được sự thay đổi của tín hiệu trên miền TFD. Điều này đã dẫn đến sự ra đời của phép biến đổi Chirplet.

2.3.2 Một số ứng dụng của phép biến đổi Chirplet

Phép biến đổi Chirplet và các biến thể của nó đã chứng tỏ được hiệu quả cao trong việc phân tích các loại tín hiệu có cấu trúc phức tạp, đặc biệt là trong các ứng dụng yêu cầu phân giải cao về cả tần số và thời gian. Do đó, phép biến đổi này ngày càng được sử dụng phổ biến, từ các âm thanh trong tự nhiên, cơ khí, cho đến tiếng nói.

Chương tiếp theo của luận án sẽ tập trung phân tích các đặc tính của phép biến đổi Chirplet và bước đầu sử dụng phép biến đổi này như một bộ trích xuất đặc trưng cho tín hiệu tiếng nói và ứng dụng vào một số bài toán nhận dạng đơn giản.

Chương 3

Phép biến đổi Chirplet tuyến tính và ứng dụng

Như đã giới thiệu sơ lược ở cuối chương 2, trong họ các phép biến đổi đa phân giải, phép biến đổi Chirplet là một công cụ chuyên dùng để mô tả sự biến đổi của dữ liệu. Chương này sẽ đi vào phân tích bản chất của phép biến đổi này, từ đó xây dựng thuật toán trích xuất đặc trưng cho tín hiệu tiếng nói. Phần lớn nội dung của chương 2 được công bố ở các công trình: CT01 và CT02.

3.1 Phép biến đổi Chirplet tuyến tính

Công thức tổng quát của LCT được biểu diễn như trong phương trình (3.1):

$$LCT_x(a, t_0, \omega_0, \alpha) = \int_{-\infty}^{\infty} x(t)\psi^*_{a,t_0,\omega_0,\alpha}(t)\, dt, \tag{3.1}$$

trong đó: $x(t)$ là tín hiệu cần phân tích, $\psi_{a,t_0,\omega_0,\alpha}(t)$ là hàm Chirplet cở sở, và $\psi^*_{a,t_0,\omega_0,\alpha}(t)$ là liên hợp phức của $\psi_{a,t_0,\omega_0,\alpha}(t)$. Hàm Chirplet cơ sở này được xác định như sau:

$$\psi_{a,t_0,\omega_0,\alpha}(t) = \frac{1}{\sqrt{a}} \cdot e^{j\omega_0(t-t_0)} \cdot e^{j\frac{\alpha}{2}(t-t_0)^2} \cdot e^{-\frac{(t-t_0)^2}{2a^2}}. \tag{3.2}$$

Trong phương trình (3.2), a là hệ số co giãn, t_0 là điểm thời gian cụ thể, ω_0 là điểm tần số cụ thể, và α là hệ số chirp.

Công thức dạng rời rạc của LCT được biểu diễn như trong phương trình (3.3):

$$LCT_x[a, n, p, \alpha] = \sum_{k=0}^{N-1} x[k]\psi^*_{a,n,p,\alpha}[k], \tag{3.3}$$

trong đó: $x[k]$ là các mẫu rời rạc của tín hiệu cần phân tích, $\psi_{a,n,p,\alpha}[k]$ là các hàm Chirplet cơ sở rời rạc, và $\psi^*_{a,n,p,\alpha}[k]$ là liên hợp phức của $\psi_{a,n,p,\alpha}[k]$. Hàm Chirplet rời rạc được định nghĩa ở phương trình (3.4):

$$\psi_{a,n,p,\alpha}[k] = \frac{1}{\sqrt{a}} \cdot e^{jv_p(k-u_n)} \cdot e^{j\frac{\alpha}{2}(k-u_n)^2} \cdot e^{-\frac{(k-u_n)^2}{2a^2}}. \tag{3.4}$$

với: a, u_n, v_p, α là các giá trị biểu diễn cho hệ số co giãn, điểm thời gian, điểm tần số, và hệ số chirp.

3.1.1 Xây dựng thuật toán trích xuất đặc trưng tiếng nói

Từ các phương trình (3.2) và (3.4), đường tần số tức thời được biểu diễn bởi một hàm tuyến tính có hệ số góc là α. Trong thuật toán 1, tham số này quyết định tổng năng lượng thu được của phép biến đổi LCT. Kết quả trả về của thuật toán được tổ chức thành một ma trận hai chiều được biểu diễn như trong phương trình (3.5) sau đây:

$$Ft \in R^{n_t \times n_f}, \tag{3.5}$$

với n_t, n_f lần lượt là số điểm thời gian và tần số thực hiện phép biến đổi và Ft là đặc trưng tiếng nói cần trích xuất.

Algorithm 1 Linear Chirplet Transform (LCT)

Require: Tín hiệu tiếng nói rời rạc $x[k]$ với $k \in 0, 1, \ldots, N-1$, a (hệ số co dãn), T (Tập điểm thời gian), Ω (Tập điểm tần số), α (hệ số chirp)

Ensure: Ma trận kết quả $C[T, \Omega]$

Khởi tạo ma trận: $C[T, \Omega] \leftarrow 0$
for $t_n \in T$ **do**
  for $\omega_p \in \Omega$ **do**
    Khởi tạo giá trị biến đổi: $C_{\text{value}} \leftarrow 0$
    for $k \in [0, 1, ...N-1]$ **do**
      $\Delta t \leftarrow k - t_n$
      $E_{\text{real}} \leftarrow -\frac{\pi}{a^2} \cdot (\Delta t)^2$
      $E_{\text{imag}} \leftarrow \omega_0 \cdot \Delta t + \frac{\alpha}{2} \cdot (\Delta t)^2$
      Exponent $\leftarrow E_{\text{real}} + j \cdot E_{\text{imag}}$
      $\psi[k] \leftarrow \frac{1}{\sqrt{a}} \cdot e^{\text{Exponent}}$
      $\psi^*[k] \leftarrow \overline{\psi[k]}$
      $C_{\text{value}} \leftarrow C_{\text{value}} + x[k] \cdot \psi^*[k]$
    Lưu kết quả vào ma trận: $C[n, p] \leftarrow C_{\text{value}}$
return Ma trận $C[T, \Omega]$

3.2 Một số kết quả thực nghiệm

Các kết quả thực nghiệm trên bài toán nhận dạng giới tính và nhận dạng vùng miền người nói cho thấy trong những bài toán mà giá trị của tần số tức thời quan trọng hơn so với quy luật biến đổi của tần số, thì phép biến đổi LCT không có ưu thế nổi bật so với các công cụ họ Fourier. Ngoài ra, trong các giá trị của hệ số chirp, các hệ số âm có vai trò nổi trội hơn so với các hệ số dương trong quá trình biểu diễn thông tin. Chương tiếp theo của luận án sẽ khảo sát khả năng chống nhiễu của LCT khi tiến hành trích xuất đặc trưng.

Chương 4

Phép biến đổi LCT và ứng dụng vào nhận dạng dữ liệu nhiễu

Thế mạnh của phép biến đổi Chirplet nằm ở khả năng mô hình hoá sự biến đổi của tín hiệu tiếng nói, hay đơn giản hơn là sự lên giọng hay xuống giọng của tiếng nói. Vấn đề đặt ra trong chương 4 này là thế mạnh đó của phép biến đổi Chirplet có còn được đảm bảo với dữ liệu đầu vào bị nhiễu hay không. Phần lớn nội dung của chương này đã được công bố ở các công trình CT03, CT04, và CT05.

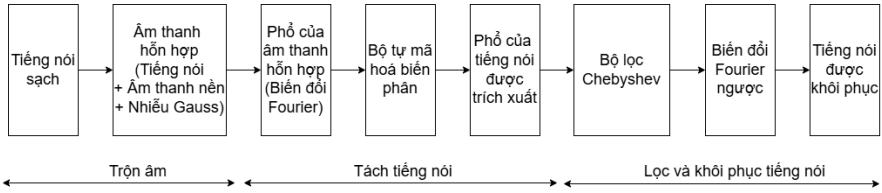
4.1 Khả năng chống nhiễu của đặc trưng LCT

Các kết quả thực nghiệm cho thấy rằng so với quá trình nhận dạng trên dữ liệu sạch, các mô hình nhận dạng trên dữ liệu nhiễu với đặc trưng LCT cho ra kết quả không tốt bằng.

4.2 Mô hình khử nhiễu và tăng cường tiếng nói

4.2.1 Xây dựng mô hình

Luận án đề xuất sử dụng bộ tự mã hoá biến phân (Variational AutoEncoder - VAE) kết hợp với bộ lọc thông dải Chebyshev để khử nhiễu và tăng cường tiếng nói. Cụ thể, VAE sẽ đảm nhận vai trò chính



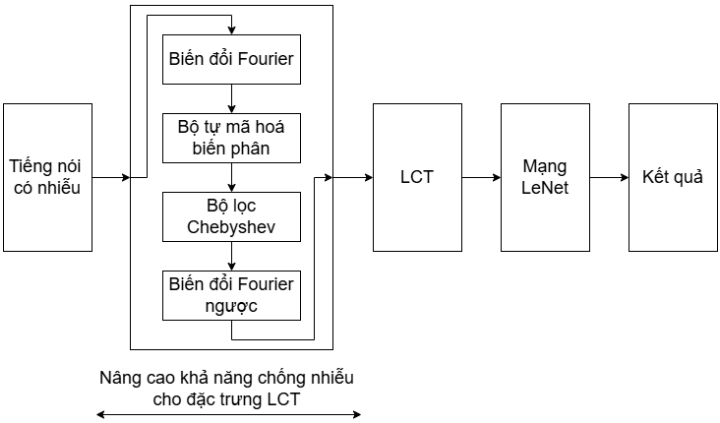
Hình 4.1: Lược đồ tổng quát mô tả các giai đoạn chính của quá trình khử nhiễu và tăng cường tiếng nói

trong quá trình xử lý, tách thành phần tiếng nói ra khỏi âm thanh hỗn hợp. Tiếng nói sau khi được VAE tách ra được đưa qua bộ lọc thông dải Chebyshev để loại bỏ các hài nằm ngoài dải tần số của tiếng nói, sau đó khôi phục lại dưới dạng waveform. Toàn bộ quá trình xử lý này được minh hoạ trong Hình 4.1.

4.3 Nâng cao khả năng trích xuất đặc trưng với dữ liệu có nhiễu cho Chirplet

4.3.1 Kết hợp LCT với mô hình khử nhiễu và tăng cường tiếng nói

Quá trình xử lý, mà cụ thể ở đây là nhận dạng một số thông tin của người nói trong môi trường có nhiễu được thực hiện qua ba giai đoạn chính với dữ liệu đầu vào là một đoạn âm thanh hỗn hợp. Đầu tiên, luận án nâng cao khả năng chống nhiễu cho hệ thống bằng mô hình đã xây dựng (VAE kết hợp với bộ lọc Chebyshev). Tiếp theo, tiếng nói đã được làm sạch sẽ được đưa qua khối phân tích đặc trưng sử dụng phép biến đổi LCT. Sau khi đặc trưng LCT được trích xuất, chúng sẽ được chuyển thành một vector đặc trưng đầu vào cho mạng nơ-ron dạng LeNet để xử lý và trả về kết quả cuối cùng.



Hình 4.2: Quá trình kết hợp phép biến đổi LCT với bộ khử nhiễu để nâng cao chất lượng cho vector đặc trưng

4.3.2 Một số kết quả thực nghiệm

Kết quả thực nghiệm cho thấy hiệu năng nhận dạng trên dữ liệu nhiễu sau khi đi qua quá trình lọc nhiễu và trích xuất đặc trưng LCT đã được cải thiện đáng kể, mặc dù vẫn thấp hơn một chút so với trường hợp nhận dạng trên dữ liệu sạch (không có nhiễu) ở chương trước.

Điều này chứng tỏ phép biến đổi LCT sau khi kết hợp với mô hình khử nhiễu như đề xuất đã giúp hệ thống trở nên bền vững hơn trước tác động của nhiễu, cải thiện hiệu năng so với trường hợp không xử lý.

Chương này đã nâng cao khả năng nhận dạng khi dùng đặc trưng LCT với các ứng dụng đơn giản. Tuy nhiên, với những ứng dụng lớn hơn và phức tạp hơn thì đặc trưng LCT là chưa đủ. Chương tiếp theo của luận án sẽ mở rộng LCT thành phép biến đổi tuyến tính tổng quát (General Linear Chirplet Transform - GLCT) nâng cao khả năng làm giàu thông tin cho vector đặc trưng.

Chương 5

Phép biến đổi Chirplet tuyến tính tổng quát và ứng dụng

Như đã trình bày trong Chương 3 và Chương 4, các đặc trưng thuộc họ Chirplet mặc dù rất giàu thông tin và có khả năng mô hình hóa sự biến đổi của tiếng nói, nhưng cũng tồn tại một số hạn chế trong quá trình trích chọn và biểu diễn đặc trưng, đặc biệt là chiến lược để lựa chọn ra các hệ số chirp tốt nhất, phù hợp với từng ứng dụng. Trong Chương 5 này, luận án đề xuất việc mở rộng phép biến đổi LCT thành phép biến đổi Chirplet tuyến tính tổng quát và Chirplet tuyến tính tổng quát rút gọn để nâng cao chất lượng đặc trưng. Một phần nội dung của chương này đã được công bố trong công trình CT06.

5.1 Phép biến Chirplet tuyến tính tổng quát

Tại mỗi điểm trên miền TFD, nếu chia góc định hướng tại đó thành N góc bằng nhau, sau đó dùng các hệ số góc này như là các hệ số chirp khác nhau, ta thu được các phép biến đổi LCT khác nhau. Khi đó:

$$\alpha_i = -\frac{\pi}{2} + i\frac{\pi}{N}, 0 < i < N. \tag{5.1}$$

Bằng cách kết hợp nhiều phép biến đổi Chirplet tuyến tính với các hệ số chirp khác nhau, GLCT cho phép mô hình hóa sự biến đổi tần số

Algorithm 2 General Linear Chirplet Transform (GLCT)

Require: Tín hiệu rời rạc $x[k]$ với $k = 0, 1, \ldots, N-1$, a (hệ số co dãn), T (Tập điểm thời gian), Ω (Tập điểm tần số), $\mathcal{A}$ (Tập hệ số chirp)

Ensure: Ma trận kết quả $GLCT[T, \Omega, \mathcal{A}]$

1: Khởi tạo ma trận:

$$GLCT[T, \Omega, \mathcal{A}] \leftarrow 0$$

2: **for** $\alpha_q \in \mathcal{A}$ **do** ▷ Với mỗi α_q, gọi lại thuật toán 1 một lần

$$GLCT[T, \Omega, q] \leftarrow LCT(T, \Omega, \alpha_q)$$

3: **return** Ma trận $GLCT[T, \Omega, \mathcal{A}]$

tức thời một cách linh hoạt và chi tiết. Quá trình thực hiện của GLCT được mô tả trong thuật toán 2. Với mỗi giá trị α_q, thuật toán 2 sẽ gọi lại thuật toán 1 tương ứng, trong đó thay giá trị α bằng giá trị α_q. Khi đó đầu ra của thuật toán sẽ là một ma trận 3 chiều, tương ứng với số lượng điểm thời gian, điểm tần số và số lượng giá trị của hệ số chirp.

5.2 Phép biến đổi Chirplet tuyến tính tổng quát rút gọn

Trong giai đoạn đầu tiên, luận án áp dụng GLCT với sự khác biệt nhỏ cho các góc để đảm bảo không mất mát thông tin quan trọng. Trong giai đoạn này, cần một không gian lớn với nhiều chiều để biểu diễn cả các đặc trưng quan sát được và ẩn của giọng nói con người. Tiếp theo, luận án sử dụng thuật toán giảm chiều tSVD để nén vector đặc trưng trong giai đoạn đầu tiên thành một không gian nhỏ hơn. Sau giai đoạn này, vector biểu diễn nằm trong không gian SVD với chiều thấp.

Algorithm 3 CGLCT - Sự kết hợp của GLCT và tSVD

Require: Tập dữ liệu $D = \{x_n\}$
Ensure: Tập dữ liệu đã được trích xuất đặc trưng D_{new}
1: $FtSet \leftarrow$ Một tập hợp rỗng
2: **for** $x_n \in D$ **do**
3: $\quad FtSet_n = GLCT(x_n)$
4: $D_{new} = tSVD(FtSet)$
5: **return** D_{new}

Với chiến lược này, thông tin có ý nghĩa nhất có thể được trích xuất và biểu diễn trong không gian chiều thấp, có thể sử dụng cho mô hình nhận dạng chính phía sau. Cả chất lượng, tương ứng với thông tin quan trọng nhất, và hình thức, tương ứng với chiều thấp của không gian, của vector đặc trưng đều được cải thiện đáng kể.

5.3 Kết quả thực nghiệm

Các kết quả thực nghiệm cho bài toán nhận dạng người nói trên tập dữ liệu TORGO cho thấy nhiều điểm quan trọng. Với CGLCT, việc kết hợp giữa GLCT và tSVD cũng tạo ra một sự cải thiện rõ nét hơn so với chỉ đơn thuần dùng GLCT. Với các dữ liệu càng bất thường, biến động, và thiếu tự nhiên, sự tham gia của tSVD vào quá trình biểu diễn đặc trưng quả thật rất quan trọng.

Xét về bản chất, những giải pháp đề xuất gồm GLCT và CGLCT cho ra các đặc trưng giàu thông tin, nâng cao hiệu quả nhận dạng, trong đó CGLCT có số chiều thấp, và dễ tinh chỉnh và thích nghi theo từng ứng dụng cụ thể. Từ đó khẳng định tính hiệu quả của việc kết hợp với các kỹ thuật giảm chiều, đồng thời mở ra hướng đi mới trong việc ứng dụng các phép biến đổi Chirplet trong xử lý tiếng nói.

Chương 6

Phép biến đổi Chirplet đa thức và ứng dụng

Chương này sẽ đi vào nghiên cứu một hướng mở rộng khác của LCT, đó là dùng một hoặc nhiều hàm đa thức để mô hình hoá sự biến đổi. Một phần nội dung của chương này đã được công bố trong công trình CT07, CT08.

6.1 Phép biến đổi Chirplet đa thức

Phép biến đổi Chirplet đa thức sử dụng một hàm đa thức bậc D cho quy luật IF thay vì một hàm tuyến tính:

$$\phi'(t) = \sum_{i=1}^{D} \alpha_i t^i. \tag{6.1}$$

Khi đó, từ phương trình (6.1), pha tức thời được xác định bởi phương trình (6.2) như sau:

$$\phi(t) = \sum_{i=1}^{D} \alpha_i \frac{t^{i+1}}{i+1}. \tag{6.2}$$

Do đó, hàm Chirplet mẹ $\psi(a, t_0, \omega_0, \mathcal{A})$ sẽ có dạng hàm mũ của e với số mũ là một đa thức bậc $D+1$ với các hệ số xác định bởi $\mathcal{A}$. Khi tín hiệu phân tích $z[n]$ được biểu diễn dưới dạng rời rạc, phép biến đổi

PCT rời rạc được thể hiện bởi phương trình (6.3) sau đây:

$$PCT_z[\alpha, n, p, \mathcal{A}] = \sum_{k=0}^{N-1} x[k]\psi^*_{\alpha,t_n,\omega_p,\mathcal{A}}[k], \tag{6.3}$$

trong đó, hàm Chirplet đa thức cơ bản được xác định bởi việc thay các giá trị t_0, ω_0 thành các giá trị rời rạc.

Quá trình trích xuất đặc trưng cho tiếng nói bằng phép biến đổi PCT được trình bày trong thuật toán 4.

Algorithm 4 Polynomial Chirplet Transform (PCT)

Require: Tín hiệu rời rạc $x[k]$ với $k = 0, 1, \ldots, N-1$, a (hệ số co dãn), T (Tập điểm thời gian), Ω (Tập điểm tần số), $\mathcal{A} = (\alpha_1, \alpha_2, \ldots, \alpha_D)$ (Tập hệ số của hàm đa thức chirp)

Ensure: Ma trận kết quả $PCT[T, \Omega]$

Khởi tạo ma trận: $PCT[T, \Omega, \mathcal{A}] \leftarrow 0$
for $t_n \in T$ **do**
  for $\omega_p \in \Omega$ **do**
    Khởi tạo giá trị biến đổi: $C_{\text{value}} \leftarrow 0$
    for $k \in [0, 1, ...N-1]$ **do**
      $\Delta t \leftarrow k - t_n$
      $P(\Delta t) \leftarrow \alpha_1 \cdot \frac{(\Delta t)^2}{2} + \alpha_2 \cdot \frac{(\Delta t)^3}{3} + \ldots + \alpha_D \cdot \frac{(\Delta t)^{D+1}}{D+1}$
      $\phi_{\text{real}} \leftarrow -\frac{\pi}{a^2} \cdot (\Delta t)^2$
      $\phi_{\text{imag}} \leftarrow \omega_0 \cdot \Delta t + P(\Delta t)$
      $\text{exponent} \leftarrow \phi_{\text{real}} + j \cdot \phi_{\text{imag}}$
      $\psi[k] \leftarrow \frac{1}{\sqrt{a}} \cdot e^{\text{exponent}}$
      $\psi^*[k] \leftarrow \overline{\psi[k]}$
      $C_{\text{value}} \leftarrow C_{\text{value}} + x[k] \cdot \psi^*[k]$
    Lưu kết quả vào ma trận: $PCT[n, p] \leftarrow C_{\text{value}}$
return Ma trận $PCT[T, \Omega]$

6.2 Phép biến đổi Chirplet nhiều hàm đa thức

Trong việc xử lý và phân tích tín hiệu tiếng nói bằng phép biến đổi PCT, việc chọn ra các hàm đa thức đặc trưng cho tập dữ liệu đóng vai trò quan trọng. Luận án đề xuất sử dụng DBSCAN để lựa chọn ra các hàm đa thức đặc trưng này. Lý do cho việc này là các hành vi trong tiếng nói là rất phức tạp và nhiều ngoại lệ, rất phù hợp với các thế mạnh của DBSCAN. Với một tập dữ liệu cho trước, luận án thiết kế thuật toán 5 để tìm ra các hàm đa thức tối ưu, sau đó sẽ dùng các hàm này như là phương trình của đường IF để thực thi các phép biến đổi PCT tương ứng.

6.3 Kết quả thực nghiệm

Các kết quả thực nghiệm nhận dạng tiếng nói trên tập LibriSpeech (tiếng Anh) và trên tập VIVOS (tiếng Việt) Các đặc trưng mạnh hơn mà luận án đề xuất như PCT và MPCT cho hiệu quả nhỉnh hơn về tỉ lệ lỗi so với MFCC hay phổ và phổ mel.

Ngoài ra, kết quả nhận dạng cảm xúc trên IEMOCAP với nhiều đặc trưng khác nhau thể hiện rằng MPCT là đặc trưng tốt nhất. Bên cạnh đó, sự vượt trội của MPCT là rất rõ ràng so với các trong các bài toán khác. Điều này có thể được giải thích là cảm xúc được nhận biết chủ yếu bởi quá trình biến đổi của cao độ trên miền TFD, chứ không phải do bản thân giá trị của tín hiệu.

Nhìn chung, MPCT nổi bật với ưu điểm là khả năng duy trì độ chính xác cao của phân tích tín hiệu khi áp dụng vào các loại dữ liệu phức tạp. Đây là một cải tiến đáng kể so với phương pháp PCT truyền thống, đặc biệt là trong ứng dụng nhận dạng cảm xúc người nói.

Algorithm 5 Phương pháp chọn hàm đa thức tối ưu bằng DBSCAN

Require: Một tập hợp các tín hiệu cần phân tích D
Ensure: Tập hợp các bộ trọng số đa thức $\mathcal{A}_D = \{\mathcal{A}_1, \mathcal{A}_2, ...\}$

Bước 1: Xây dựng ma trận Dữ Liệu
for all $x_i \in D$ **do**
    $\mathcal{A}_i \leftarrow$ tìm hàm đa thức cho x_i bằng thuật toán **??**
    $[\alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,N}] \leftarrow \mathcal{A}_i$
Xây dựng ma trận hệ số H chứa các hệ số đa thức:

$$H \leftarrow \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,N} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{M,1} & \alpha_{M,2} & \cdots & \alpha_{M,N} \end{bmatrix} \tag{6.4}$$

Bước 2: Áp Dụng Thuật Toán DBSCAN
$C \leftarrow DBSCAN_{\epsilon, MinPts}(H)$. ▷ Thực hiện bằng thuật toán **??**
Bước 3: Xác Định Tâm Của Các Cụm
for $C_k \in C$ **do**
    Tính toán tâm của cụm C_k:

$$\mu_k \leftarrow \frac{1}{|C_k|} \sum_{i \in C_k} H_i \tag{6.5}$$

    Chuyển toạ độ của tâm cụm sang bộ trọng số của đa thức:

$$\mathcal{A}_k \leftarrow \mu_k \tag{6.6}$$

Bước 4: Xác Nhận Kết Quả
return $\mathcal{A}_D = \{\mathcal{A}_1, \mathcal{A}_2, ...\}$

Chương 7

Tổng kết

7.1 Những nội dung chính của luận án

Luận án này đã giới thiệu một framework tổng quát để giải quyết bài toán nhận dạng trong xử lý tiếng nói bằng các phương pháp học máy. Framework bao gồm các giai đoạn chính: tiền xử lý và khử nhiễu, trích xuất đặc trưng, biến đổi đặc trưng, nhận dạng, và hậu xử lý để đưa ra kết quả cuối cùng. Mỗi giai đoạn đóng vai trò quan trọng trong việc nâng cao hiệu suất và độ chính xác của hệ thống nhận dạng tiếng nói, trong đó luận án tập trung nghiên cứu chủ yếu ở các giai đoạn trích xuất và biến đổi đặc trưng theo tiếp cận đa phân giải với các phép biến đổi họ Chirplet.

Đầu tiên, luận án đề xuất xây dựng bộ trích xuất đặc trưng cho tiếng nói bằng cách sử dụng Phép biến đổi Chirplet tuyến tính (LCT) và áp dụng vào một số bài toán nhận dạng cơ bản như nhận dạng giới tính và vùng miền của người nói. Sau đó, nhằm nâng cao khả năng nhận dạng với dữ liệu có nhiễu, luận án đã đề xuất giải pháp kết hợp LCT với Bộ tự mã hóa biến phân (VAE) và bộ lọc Chebyshev.

Phần tiếp theo của luận án cũng đã đề xuất hai phương pháp trích xuất và biểu diễn đặc trưng cho tín hiệu tiếng nói dựa trên các phép biến đổi họ Chirplet, bao gồm: Phép biến đổi Chirplet tuyến tính tổng quát (GLCT) và Phép biến đổi Chirplet tuyến tính tổng quát rút gọn

(CGLCT)

Cuối cùng, Để nâng cao chất lượng của các đặc trưng đa phân giải họ Chirplet trong trường hợp dữ liệu biến đổi phức tạp, luận án đã đề xuất hai phương pháp gồm phép biến đổi tuyến tính đa thức (PCT) và phép biến đổi Chirplet với nhiều hàm đa thức (MPCT) bằng cách kết hợp PCT với DBSCAN.

7.2 Kết luận của luận án

Luận án đã nghiên cứu và đề xuất các giải pháp nhằm nâng cao hiệu suất của hệ thống nhận dạng tiếng nói thông qua việc khử nhiễu và trích xuất đặc trưng hiệu quả.

Thứ nhất, luận án đã đề xuất và triển khai các đặc trưng đa phân giải họ Chirplet, bao gồm LCT, GLCT và PCT. Các đặc trưng này có chất lượng tốt, giàu thông tin, và đã nâng cao độ chính xác của mô hình nhận dạng tiếng nói. Việc sử dụng các phép biến đổi Chirplet cho phép mô hình hóa hiệu quả các biến đổi phức tạp của tín hiệu tiếng nói trong miền Thời gian - Tần số, giúp hệ thống nhận dạng tiếp cận thông tin một cách toàn diện hơn.

Thứ hai, luận án đã phát triển giải pháp nâng cao khả năng chống nhiễu cho phép biến đổi Chirplet bằng cách kết hợp Chirplet với Bộ tự mã hóa biến phân (VAE) và bộ lọc Chebyshev. Giải pháp này đã chứng minh khả năng nhận dạng đối với dữ liệu có nhiễu trong các bài toán khác nhau.

Cuối cùng, luận án đã phát triển các đặc trưng nâng cao như CGLCT và MPCT, kết hợp các kỹ thuật giảm chiều dữ liệu và học máy. Các đặc trưng này có chất lượng cao, biểu diễn một lượng lớn thông tin trong không gian có số chiều thấp, qua đó giảm độ phức tạp

không gian và thời gian của giai đoạn trích xuất và biến đổi đặc trưng. Điều này không chỉ cải thiện hiệu suất tính toán mà còn nâng cao khả năng ứng dụng thực tế của các đặc trưng đề xuất, khi chúng có thể được tích hợp dễ dàng vào các hệ thống nhận dạng tiếng nói với yêu cầu tài nguyên hạn chế. Bên cạnh đó, với sự đóng góp của các mô hình giảm chiều, các đặc trưng này có khả năng biến đổi tuỳ theo nhu cầu của từng bài toán cụ thể, trên từng tập dữ liệu cụ thể.

Những kết quả đạt được trong luận án không chỉ khẳng định tính hiệu quả của các giải pháp đề xuất mà còn mở ra hướng nghiên cứu mới trong việc kết hợp các kỹ thuật học máy và xử lý tín hiệu để giải quyết các vấn đề phức tạp trong lĩnh vực xử lý tiếng nói.

7.3 Một số hướng phát triển trong tương lai

Trong tương lai, việc phát triển phương pháp phân tích đa phân giải với phép biến đổi Chirplet cần tập trung giải quyết một số hạn chế hiện tại. Thứ nhất là thiết kế thuật toán tính toán nhanh, nhằm nâng cao tốc độ thực thi của phép biến đổi Chirplet, giúp ứng dụng trong thời gian thực và các hệ thống yêu cầu hiệu năng cao. Thứ hai là nâng cao khả năng giải thích của đặc trưng, có thể thông qua việc kết hợp với các kỹ thuật học máy giải thích được hoặc tối ưu hóa quá trình giảm chiều dữ liệu mà không làm mất mát thông tin quan trọng. Thứ ba là xử lý các vấn đề về thời gian thực thi và thu nhỏ vector đặc trưng. Những hướng phát triển này sẽ giúp nâng cao hiệu quả và khả năng ứng dụng thực tế của các phương pháp phân tích đa phân giải với phép biến đổi Chirplet.

Danh mục công trình của NCS

Các công trình mà NCS là tác giả chính, đã công bố, và liên quan quan trực tiếp đến nội dung chính của luận án

CT01 Hao D. Do, Duc T. Chau, and Son T. Tran. "Speech Representation Using Linear Chirplet Transform and Its Application in Speaker-Related Recognition". In: Computational Collective Intelligence. Ed. by Ngoc Thanh Nguyen et al. Cham: Springer International Publishing, 2022, pp. 719–729. isbn: 978-3-031-16014-1. doi: 10.1007/978-3-031-16014-1_56. (Rank B)

CT02 Hao Duc Do, Duc Thanh Chau and Son Thai Tran. "Speech feature extraction using linear Chirplet transform and its applications". In: Journal of Information and Telecommunication 7.3 (2023), pp. 376– 391. doi: 10.1080/24751839.2023.2207267. (ESCI, Scopus Q2, IF 2.7)

CT03 Hao D. Do, Son T. Tran, and Duc T. Chau. "A Variational Autoen-coder Approach for Speech Signal Separation". In: Computational Collective Intelligence. Ed. by Ngoc Thanh Nguyen et al. Cham: Springer International Publishing, 2020, pp. 558–567. isbn: 978-3-030-63007-2. doi: 10.1007/978-3-030-63007-2_43. (Rank B)

CT04 Hao Duc Do, Son Thai Tran, and Duc Thanh Chau. "Speech Separation in the Frequency Domain with Autoencoder". In:

Journal of Communications 15.11 (2020), pp. 841–848. doi: 10.12720/jcm.15. 11.841-848. (Scopus Q3)

CT05 Hao Duc Do, Son Thai Tran, and Duc Thanh Chau. "Speech Source Separation Using Variational Autoencoder and Bandpass Filter". In: IEEE Access 8 (2020), pp. 156219–156231. doi: 10.1109/ACCESS.2020.3019495. (SCIE, Scopus Q1, IF 3.336)

CT06 Hao Duc Do, Duc Thanh Chau and Son Thai Tran. "Speech Feature Enhancement based on Time-frequency Analysis". In: ACM Transactions on Asian and Low-Resource Language Information Processing 22.8 (2023), pp. 1–14. doi: 10.1145/3605549. (SCIE, Scopus Q2, IF 1.8)

CT07 Hao Duc Do, Duc Thanh Chau and Son Thai Tran. "A New Algorithm for Speech Feature Extraction Using Polynomial Chirplet Transform". Circuits Syst Signal Process 43 (2024), pp. 2320–2340, doi: 10.1007/s00034-023-02561-6. (SCIE, Scopus Q2, IF 1.8)

CT08 Hao D. Do, Son T. Tran, and Duc T. Chau. "Enriching information for speech emotion feature using polynomial chirplet transform and density-based clustering algorithm". In: 17th Asian Conference on Intelligent Information and Database (ACIIDS 2025 - Accepted). (Rank B)