BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

DƯƠNG THỊ HIỀN THANH

TÁCH NGUỒN ÂM THANH SỬ DỤNG MÔ HÌNH PHỔ NGUỒN TỔNG QUÁT TRÊN CƠ SỞ THỪA SỐ HÓA MA TRẬN KHÔNG ÂM

Ngành: Khoa học máy tính
Mã số: 9480101

TÓM TẮT LUẬN ÁN TIẾN SĨ
KHOA HỌC MÁY TÍNH

Hà Nội - 2019

Công trình được hoàn thành tại:
Trường Đại học Bách khoa Hà Nội

Người hướng dẫn khoa học:
1. PGS. TS. Nguyễn Quốc Cường
2. TS. Nguyễn Công Phương

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Trường họp tại Trường đại học Bách khoa Hà Nội

Vào hồi....., ngày.....tháng.....năm.....

Có thể tìm hiểu luận án tại thư viện:
1. Thư viện Tạ Quang Bửu - Trường Đại học Bách khoa Hà Nội
2. Thư viện Quốc gia Việt Nam

MỞ ĐẦU

1. Đặt vấn đề

Trong thực tế cuộc sống có rất nhiều tình huống thu âm mà âm thanh mong muốn bị trộn lẫn với nhiều âm thanh khác, tiếng ồn từ môi trường xung quanh và tiếng vọng của hiện tượng phản xạ âm thanh mang lại. Con người với khả năng thính giác bình thường qua hai tai có thể dễ dàng định vị và phân tách âm thanh mong muốn để nghe, hiểu. Tuy nhiên đối với học máy thì việc đó lại trở nên vô cùng khó khăn. Vì lý do đó, nhiều ứng dụng thực tế (như hệ thống nhận dạng tiếng nói tự động, robotics, hội nghị truyền thanh/truyền hình, hệ thống hỗ trợ người khiếm thính, xử lý âm thanh hậu kỳ trong sản xuất phim ảnh,...) sử dụng kỹ thuật tách nguồn âm thanh [5] để phân tách, nâng cao chất lượng âm thanh mong muốn như một bước tiền xử lý quan trọng.

Những công bố gần đây về tách nguồn âm cho thấy trong điều kiện tỷ lệ nhiễu thấp và không có hiện tượng phản xạ âm thanh, một số thuật toán tách nguồn âm cho kết quả tương đối tốt. Nhưng với môi trường thu âm thực có mức nhiễu và tiếng vọng cao thì kết quả tách âm vẫn còn khá thấp. Các công bố cũng cho thấy thuật toán tách nguồn mù đạt kết quả phân tách chưa đủ tốt để đưa vào ứng dụng thực tế. Một số nghiên cứu sử dụng dữ liệu huấn luyện, hoặc những thông tin phụ trợ tương đối cụ thể (như tách âm nhạc khi biết trước bản nhạc, tách tiếng nói khi biết bản transcript,...) để hướng dẫn quá trình phân tách đã đạt được kết quả tốt hơn [4, 7, 8]. Tuy nhiên, dữ liệu huấn luyện hoặc những thông tin hướng dẫn cụ thể như thế thường không dễ dàng có được trong nhiều tình huống ứng dụng.

Từ những phân tích đó, chúng tôi tập trung phát triển thuật toán tách nguồn âm thanh trong trường hợp còn nhiều khó khăn thách thức: tín hiệu thu âm trong môi trường có phản xạ, chứa nhiễu ở mức cao, số lượng nguồn âm lớn hơn hoặc bằng số microphone (*determined/ underdetermined*) và không có dữ liệu huấn luyện cho các âm thanh cần phân tách. Tiếp cận theo hướng *weakly-informed*, chúng tôi sử dụng thông tin phụ trợ rất chung chung để hướng dẫn quá trình phân tách, đó là cần biết âm thanh có trong hỗn hợp là những loại nào (ví dụ như tiếng nói, âm thanh môi trường hay âm nhạc,...).

2. Mục tiêu và phạm vi nghiên cứu của luận án

- **Mục tiêu nghiên cứu**

  Mục tiêu của luận án là nghiên cứu phát triển thuật toán tách nguồn âm thanh có thể thực hiện phân tách nguồn hiệu quả trong điều kiện thu âm trong môi trường thực có phản xạ âm (*high reverberation*) và số nguồn âm nhiều hơn hoặc bằng số microphone (*determined/ underdetermined*).

Chúng tôi tìm hiểu các kỹ thuật phân tách âm thanh khác nhau, từ đó lựa chọn kỹ thuật phù hợp nhất với mục tiêu đã đặt ra để nghiên cứu phát triển. Chúng tôi đề xuất thuật toán mới cho cả hai trường hợp tách nguồn đơn kênh và đa kênh. Dựa vào thông tin về loại âm thanh xuất hiện trong tín hiệu trộn, chúng tôi tìm kiếm một số mẫu huấn luyện cho thuật toán đề xuất. Ví dụ, với tình huống nâng cao chất lượng tiếng nói trong môi trường thực, có thể xác định âm thanh cần tách là tiếng nói, thành phần còn lại là âm thanh môi trường. Từ đó có thể tìm kiếm vài tệp ngắn (khoảng 5 giây), chứa âm thanh môi trường (cafeteria, subway, square,...) và tiếng nói làm dữ liệu huấn luyện.

Thuật toán được đánh giá bằng các thí nghiệm với hai trường hợp: phân tách tiếng nói và nhiễu môi trường, và phân tách giọng hát và âm nhạc từ một bài hát. Để dễ dàng so sánh với những nghiên cứu khác trên thế giới, ngoài bộ dữ liệu tự xây dựng, chúng tôi sử dụng bộ dữ liệu chuẩn được công bố bởi SiSEC (Signal Separation Evaluation Campaign [1]).

- **Phạm vi nghiên cứu**

  Mục tiêu của nghiên cứu là khôi phục tín hiệu gốc của các nguồn thành phần (*original sources*) đối với trường hợp tách nguồn đơn kênh, và khôi phục tín hiệu thu được tại microphone (*spatial images*) của các nguồn thành phần trong trường hợp đa kênh.

  Hơn nữa, nghiên cứu của chúng tôi dựa trên giả định biết trước số nguồn thành phần và biết các nguồn đó thuộc loại âm thanh gì.

3. Những đóng góp của luận án

Chúng tôi đề xuất các thuật toán tách nguồn âm cho cả hai trường hợp đơn kênh và đa kênh. Kết quả nghiên cứu đã được công bố trong 7 bài báo. Kết quả của thuật toán đề xuất đã được gửi tới chiến dịch đánh giá tách nguồn âm quốc tế SiSEC 2016[2] và đạt kết quả tốt nhất với bộ tiêu chí đánh giá dựa trên năng lượng. Những đóng góp cụ thể của luận án như sau:

- Đề xuất thuật toán tách nguồn âm đơn kênh sử dụng tập mẫu huấn luyện là vài file âm thanh ngắn (khoảng 4 giây) cùng loại với các nguồn cần tách. Trong thuật toán đề xuất, mô hình phổ tổng quát GSSM của âm thanh được xây dựng bằng cách học các đặc trưng phổ từ tập mẫu huấn luyện, sau đó được sử dụng để hướng dẫn bước phân tách dùng mô hình thừa số hóa ma trận không âm (Nonnegative Matrix Factorization - NMF). Chúng tôi cũng đề xuất công thức ràng buộc thưa mới cho hàm giá trong quá trình ước lượng các nguồn thành phần ở bước phân

[1] http://sisec.inria.fr/

[2] http://sisec.inria.fr/sisec-2016/

tách. Thuật toán được xác thực về hiệu quả phân tách, khả năng hội tụ và tính ổn định đối với sự thay đổi của các tham số thông qua các thí nghiệm trên 3 bộ dữ liệu với các thiết lập unsupervised và semi-supervised.

- Đề xuất thuật toán tách nguồn đa kênh kết hợp NMF trong mô hình Gaussian cục bộ (Local Gaussian Model - LGM). Chúng tôi đề xuất hai tiêu chí tối ưu mới cho bước ước lượng thông tin phổ của các nguồn thành phần: (1) ước lượng đặc trưng phổ của từng nguồn riêng biệt và (2) ước lượng đồng thời trên tất cả các nguồn. Từ đó, chúng tôi tính toán công thức cập nhật tham số tương ứng với từng tiêu chí ước lượng và xây dựng thuật toán. Hiệu quả phân tách cũng như khả năng hội tụ và tính ổn định của thuật toán được xác thực bằng thí nghiệm trên bộ dữ liệu SiSEC (Signal Separation Evaluation Campaign), là bộ dữ liệu được dùng phổ biến trong cộng đồng tách nguồn âm trên thế giới.
- Ngoài hai đóng góp chính nêu trên, trong quá trình nghiên cứu và ứng dụng mô hình NMF trong xử lý âm thanh, chúng tôi đề xuất ba phương pháp tự động trích xuất những đoạn âm thanh bất thường từ tín hiệu thu âm ngoài trời kích thước lớn. Thí nghiệm đã chứng minh khả năng mô hình hóa tốt các đặc trưng phổ âm thanh của NMF. Thuật toán đề xuất đã được chuyển giao cho công ty RION (tại Tokyo-Nhật Bản) để phát triển và sử dụng hỗ trợ việc phát hiện, gán nhãn các sự kiện âm thanh.

4. Cấu trúc của luận án

- ***Chương 1:*** Giới thiệu tổng quan về kỹ thuật tách nguồn âm thanh và những kết quả nghiên cứu liên quan đã được công bố, đồng thời mô hình hóa bài toán tách nguồn âm thanh mà luận án sẽ nghiên cứu giải quyết.
- ***Chương 2:*** Chương này giới thiệu mô hình NMF, được sử dụng rộng rãi trong xử lý âm thanh. Chúng tôi cũng trình bầy thuật toán tách nguồn âm thanh dựa trên NMF, là thuật toán cơ sở cho đề xuất của chúng tôi. Bên cạnh đó, chúng tôi đề xuất phương pháp trích xuất các đoạn âm thanh bất thường xuất hiện trong file ghi âm dài. Đề xuất cho thấy một hướng ứng dụng khác của NMF, đồng thời xác thực khả năng mã hóa các đặc trưng phổ âm thanh của mô hình NMF.
- ***Chương 3:*** Chúng tôi đề xuất thuật toán tách nguồn đơn kênh. Trong đó, mô hình phổ tổng quát GSSM được huấn luyện từ một vài ví dụ mẫu cùng loại với âm thanh cần phân tách bởi mô hình NMF. Chúng tôi cũng đề xuất hàm ràng buộc thưa thớt (sparsity-inducing penalty function) mới cho bước ước lượng các tham số. Đồng thời tính toán công thức cập nhật tham số theo hàm ràng buộc thưa mới đề xuất và xây dựng thuật toán. Hiệu quả của thuật toán đề xuất được xác thực bằng thí nghiệm trên ba bộ dữ liệu với các cài đặt khác nhau.

- ***Chương 4:*** Chương này mô tả thuật toán tách nguồn đa kênh mới, kết hợp mô hình phổ tổng quát GSSM với mô hình hiệp phương sai không gian của các nguồn âm trong khuôn khổ mô hình LGM. Để hướng dẫn ước lượng phương sai nguồn trung gian trong mỗi vòng lặp EM, chúng tôi đề xuất hai tiêu chí tối ưu hóa: (1) ước lượng phương sai của từng nguồn riêng biệt bằng mô hình NMF kết hợp với ràng buộc thưa đề xuất, (2) ước lượng phương sai của tất cả các nguồn đồng thời. Cuối cùng là thí nghiệm nhằm đánh giá hiệu suất phân tách của thuật toán đề xuất cũng như khả năng hội tụ và tính ổn định của thuật toán.

Phần cuối của luận án, chúng tôi nêu những đánh giá, kết luận về kết quả nghiên cứu đã đạt được và đề xuất định hướng nghiên cứu trong tương lai.

CHƯƠNG 1: TỔNG QUAN VỀ TÁCH NGUỒN ÂM THANH VÀ NHỮNG NGHIÊN CỨU LIÊN QUAN

1.1 Tổng quan về tách nguồn âm thanh

1.1.1 Mô hình chung của hệ thống tách nguồn âm

Tách nguồn âm thanh là kỹ thuật khôi phục những âm thanh thành phần (gọi là *nguồn âm*) từ tín hiệu chứa các âm thanh bị trộn lẫn (gọi là tín hiệu trộn (*mixture*)) đơn kênh hoặc đa kênh. Các hệ thống tách nguồn âm thanh thường ước lượng các nguồn thành phần trong miền thời gian - tần số (T-F), có thể dùng một trong hai hoặc cả hai mô hình sau: (1) mô hình phổ *spectral model* mã hóa và khai thác thông tin về đặc trưng phổ của âm thanh, (2) mô hình không gian *spatial model* mã hóa và khai thác thông tin về không gian. Sau quá trình ước lượng, các âm thanh thành phần được biến đổi về miền thời gian qua phép biến đổi Fourier ngược (ISTFT).

1.1.2 Xây dựng bài toán

Giả sử tín hiệu trộn từ J nguồn âm được thu âm bởi I microphone, với $j \in \{1, 2, \ldots, J\}$ là chỉ số của nguồn âm và $i \in \{1, 2, \ldots, I\}$ là chỉ số của microphone. Tín hiệu trộn $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^T \in \mathbb{R}^{I\times 1}$ được biểu diễn theo công thức sau [5]:

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t), \tag{1.1}$$

với $\mathbf{c}_j(t) = [c_{1j}(t), \ldots, x_{Ij}(t)]^T \in \mathbb{R}^{I\times 1}$ là tín hiệu thu được tại các microphone của nguồn thứ j, được gọi là *spatial image* của nguồn j, $.^T$ là phép toán chuyển vị của véc tơ hoặc ma trận, $t \in \{0, 1, \ldots, T-1\}$ là chỉ số khung thời gian và T là độ dài thời gian của tín hiệu. Công thức (1.1) trong miền thời gian - tần số (sau phép biến đổi Fourier STFT) được viết như sau:

$$\mathbf{x}(n, f) = \sum_{j=1}^{J} \mathbf{c}_j(n, f) \tag{1.3}$$

với $\mathbf{c}_j(n, f) \in \mathbb{C}^{I\times 1}$ và $\mathbf{x}(n, f) \in \mathbb{C}^{I\times 1}$ là biểu diễn trong miền T-F tương ứng của $\mathbf{c}_j(t)$ và $\mathbf{x}(t)$. $n = 1, 2, .., N$ là chỉ số khung thời gian và $f = 1, 2, ..., F$ biểu diễn số bin tần số. Mục tiêu của hệ thống tách nguồn âm thanh là khôi phục J tín hiệu nguồn thành phần $s_j(t)$ *(original source)*,hoặc khôi phục tín hiệu nguồn không gian *(spatial images)* $\mathbf{c}_j(t)$ từ tín hiêu trộn I kênh $\mathbf{x}(t)$.

1.2 Những nghiên cứu liên quan

- **Các mô hình phổ**: Phần này giới thiệu ba mô hình phổ biến, được dùng để mã hóa và khai thác thông tin phổ của âm thanh. Đó là mô hình Gaussian (Spectral GMM), mô hình thừa số hóa ma trận không âm (NMF), và deep neural network (DNN).
- **Các mô hình không gian**: Trong phần này, chúng tôi giới thiệu ba kỹ thuật mô hình hóa và khai thác các đặc tính về không gian và môi trường truyền âm. Đó là interchannel intensity/time difference (IID/ITD), rank-1 mixing vector, và mô hình mô hình hiệp phương sai không gian full-rank (full-rank spatial covariance model).

1.3 Các tiêu chí đánh giá nguồn tách

- **Energy-based criteria**: Nhóm tiêu chí dựa trên năng lượng gồm có 4 độ đo, được đo bằng đơn vị dB với giá trị càng cao càng tốt. Bốn độ đo đó là *Signal to Distortion Ratio* (SDR), *Signal to Artifacts Ratio* (SAR), *Signal to Interference Ratio* (SIR), và *source Image to Spatial distortion Ratio* (ISR).
- **Perceptually-based criteria**: Nhóm tiêu chí đánh giá dựa trên sự cảm thụ của tai người gồm 4 độ đo: *Overall Perceptual Score* (OPS), *Artifacts-related Perceptual Score* (APS), *Interference-related Perceptual Score* (IPS), và *Target-related Perceptual Score* (TPS). Các độ đo có giá trị từ 0 đến 100, giá trị cao biểu diễn hiệu quả phân tách tốt.

Tổng kết

Trong chương này, chúng tôi giới thiệu tổng quan về kỹ thuật tách nguồn âm thanh và những kiến thức liên quan, đồng thời xây dựng bài toán được tập trung nghiên cứu trong luận án.

CHƯƠNG 2: PHƯƠNG PHÁP THỪA SỐ HÓA MA TRẬN KHÔNG ÂM

2.1 Tổng quan về thừa số hóa ma trận không âm (Nonnegative Matrix Factorization - NMF)

2.1.1 NMF là gì?

Thừa số hóa ma trận không âm (NMF) là kỹ thuật giảm số chiều của ma trận được sử dụng phổ biến trong phân tích dữ liệu không âm.

Cho ma trận không âm $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ kích thước $F \times N$, NMF thực hiện phân tách $\mathbf{V}$ thành hai ma trận không âm $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ và $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ sao cho $\mathbf{V} \approx \mathbf{WH}$. NMF được dùng phổ biến trong xử lý tín hiệu, trong đó có lĩnh vực xử lý âm thanh [1].

2.1.2 Hàm giá

Việc phân tách ma trận $\mathbf{V}$ thành hai ma trận $\mathbf{W}$ và $\mathbf{H}$ được thực hiện bởi quá trình tối ưu hóa hàm mục tiêu [1]:

$$\min_{\mathbf{H} \geq 0, \mathbf{W} \geq 0} D(\mathbf{V} \| \mathbf{WH}), \tag{2.2}$$

với $D(\mathbf{V} \| \mathbf{WH}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d_{IS}(\mathbf{V}_{fm} \| [\mathbf{WH}]_{fm})$, $d_{IS}(x \| y) = \frac{x}{y} - \log(\frac{x}{y}) - 1$ là Itakura Saito divergence được sử dụng phổ biến với tín hiệu âm thanh.

2.1.3 Quy tắc cập nhật tham số MU rules

Để tối ưu hóa hàm mục tiêu (2.2), Lee và Seung đã đề xuất quy tắc cập nhật cho các thành phần NMF, được gọi là *multiplicative update (MU)* rules [2] và được viết như sau:

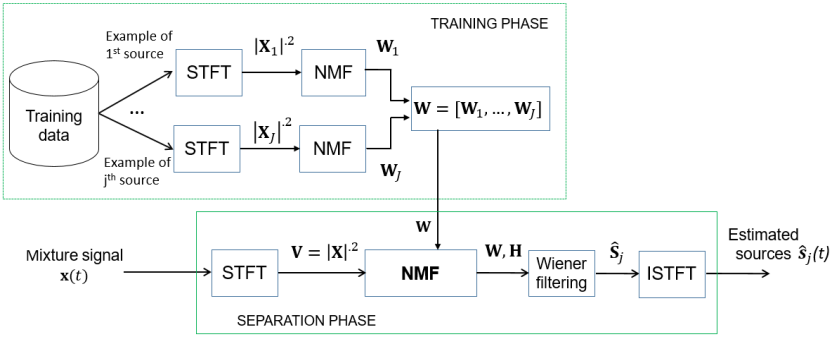
$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T \left((\mathbf{WH})^{\cdot(\beta-2)} \odot \mathbf{V} \right)}{\mathbf{W}^T (\mathbf{WH})^{\cdot(\beta-1)}}, \tag{2.13}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left((\mathbf{WH})^{\cdot(\beta-2)} \odot \mathbf{V} \right) \mathbf{H}^T}{(\mathbf{WH})^{\cdot(\beta-1)} \mathbf{H}^T}, \tag{2.14}$$

2.2 Áp dụng NMF trong bài toán tách nguồn âm

Mô hình chung của thuật toán tách nguồn âm thanh dựa trên NMF được mô tả trong hình 2.3 và gồm hai quá trình: (1) học các đặc tính phổ của các nguồn từ dữ liệu huấn

luyện bằng mô hình NMF, và (2) ước lượng tín hiệu các nguồn thành phần từ tín hiệu trộn dựa trên ma trận đặc trưng phổ đã được học trước đó.



Hình 2.3: Sơ đồ thuật toán tách nguồn âm thanh dựa trên NMF.

Ma trận đặc trưng phổ của từng nguồn thành phần, ký hiệu $\mathbf{W}_j, j = 1, \ldots, J$, được học từ dữ liệu huấn luyện qua quá trình tối ưu hóa hàm (2.2) của mô hình NMF. Từ đó, ma trận đặc trưng phổ của tất cả các nguồn thành phần $\mathbf{W}$ được xác định và là tham số đầu vào cho pha tách nguồn. Trong pha tách nguồn, thuật toán sẽ ước lượng ma trận kích hoạt $\mathbf{H}$ theo công thức cập nhật tham số MU. Sau khi ước lượng các ma trận tham số $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$, tín hiệu nguồn thành phần thứ j trong miền T-F được tính toán bằng công thức Wiener filtering: $\hat{\mathbf{S}}_j = \frac{\mathbf{W}_j \mathbf{H}_j}{\mathbf{WH}} \odot \mathbf{X}$, trong đó $\odot$ là ký hiệu phép nhân element-wise Hadamard. Cuối cùng, các tín hiệu nguồn thành phần được biến đổi về miền thời gian qua phép biến đổi ISTFT.

Lưu ý rằng thuật toán nêu trên ước lượng các thành phần theo quy tắc cập nhật tham số MU rules với sự hướng dẫn của ma trận đặc trưng phổ $\mathbf{W}$ đã được học trước từ dữ liệu huấn luyện. Do đó, thuật toán sẽ hoạt động tốt khi có dữ liệu huấn luyện và kết quả phân tách sẽ kém khi không có dữ liệu huấn luyện. Điều này sẽ được xác thực qua kết quả thí nghiệm trong chương 3.

2.3 Áp dụng NMF trong bài toán phát hiện những âm thanh bất thường

2.3.1 Mô tả bài toán

Trong phần này, chúng tôi trình bầy cách áp dụng NMF để phát hiện những đoạn âm thanh bất thường trong tín hiệu thu âm thực. Chúng tôi đề xuất thuật toán tự động

trích xuất những đoạn âm thanh bất thường từ tín hiệu thu âm dài (nhiều giờ) mà không dùng bất kỳ dữ liệu hay thông tin hướng dẫn nào.

Trong thực tế, âm thanh nhiễu môi trường (background sound) luôn tồn tại trong suốt thời gian thu âm và các sự kiện âm thanh thường xuất hiện với thời gian ngắn hơn. Ví dụ: với tín hiệu thu âm ở công viên vào mùa hè và ban ngày thì tiếng ve và tiếng gió sẽ xuất hiện thường xuyên và được coi là âm thanh nền; trong khi đó tiếng còi xe, tiếng bước chân, hay tiếng người nói,... là những sự kiện âm thanh có thể xuất hiện không thường xuyên.

NMF có khả năng mô hình hóa những đặc trưng phổ của âm thanh. Nếu số lượng đặc trưng phổ nhỏ (K nhỏ), NMF sẽ mô hình hóa những đặc trưng xuất hiện thường xuyên hơn trong tín hiệu đầu vào.

Từ nhận định đó, để kiểm chứng khả năng mô hình hóa đặc trưng âm thanh của mô hình NMF, chúng tôi đề xuất 3 thuật toán tự động trích xuất những sự kiện âm thanh, hay còn gọi là "âm thanh bất thường".

2.3.2 Thuật toán đề xuất

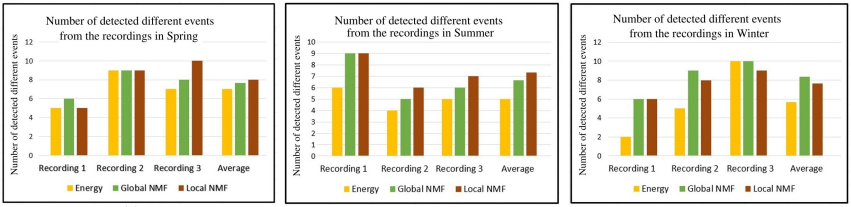
- **Signal energy-based method**: Nhận thấy âm thanh nền thường có năng lượng phổ nhỏ hơn các sự kiện âm thanh. Thuật toán sẽ tính toán năng lượng phổ của từng đoạn âm thanh ngắn từ ma trận phổ $\mathbf{V}$, sau đó trích xuất những đoạn âm thanh có năng lượng phổ cao với mong muốn đó sẽ là các sự kiện âm thanh.
- **Global NMF-based method**: Thuật toán sử dụng NMF với 1 thành phần phổ cơ sở duy nhất ($K = 1$) để mô mình hóa đặc trưng âm thanh xuất hiện thường xuyên nhất, với mong muốn đó chính là đặc trưng của âm thanh nền. Sau khi tính toán ma trận divergence, những phân đoạn âm thanh tại vị trí divergence cao sẽ được trích xuất với mong muốn đó sẽ là các sự kiện âm thanh.
- **Local NMF-based method**: Với những file ghi âm dài nhiều giờ, âm thanh nền có thể thay đổi. Khi đó áp dụng NMF trên từng phân đoạn ngắn hơn của file âm thanh có thể mang lại kết chính xác hơn. Chúng tôi đề xuất giải pháp áp dụng NMF trên từng phân đoạn ngắn (ví dụ 10 phút). Sau đó ma trận divergence được tính toán và các phân đoạn được trích xuất giống như phương pháp Global NMF-based.

2.3.3 Thí nghiệm

Chúng tôi sử dụng 9 file âm thanh đơn kênh được ghi âm ngoài trời vào 3 mùa khác nhau trong năm tại các địa điểm: công viên, bãi đỗ xe, góc đường. Mỗi file dài 1 giờ[1]. Kết quả thí nghiệm (hình 2.5) cho thấy: hai phương pháp sử dụng NMF cho kết quả

[1]Test data are provided by RION Co., Ltd., in Japan.

trích xuất tốt hơn phương pháp dựa trên năng lượng. Với file âm thanh mà âm thanh nền không thay đổi, kết quả của global NMF-based method là tốt nhất (ví dụ, vào mùa đông, âm thanh nền là tiếng gió). Với file có âm thanh nền thay đổi (như vào mùa hè, âm thanh nền thay đổi gồm tiếng chim, tiếng ve, tiếng gió xài xạc) thì kết quả của local NMF-based method là tốt hơn. Thí nghiệm cho thấy NMF với 1 thành phần phổ cơ sở có khả năng mô hình hóa tốt đặc trưng của âm thanh nền xuất hiện thường xuyên nhất trong tín hiệu. Điều này một lần nữa xác thực khả năng mô hình hóa tốt đặc trưng phổ âm thanh của mô hình NMF.



Hình 2.6: Số lượng sự kiện âm thanh được phát hiện của ba phương pháp.

2.4 Tổng kết

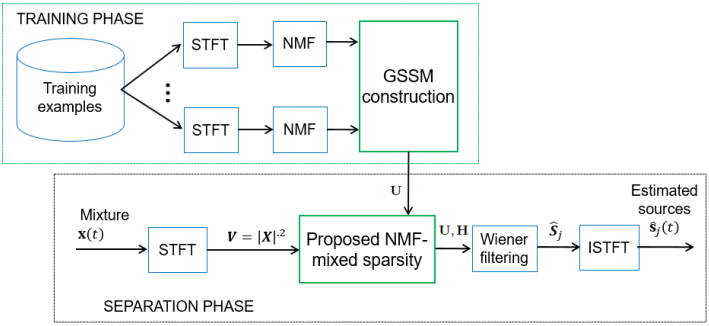
Chương này giới thiệu về NMF, kỹ thuật được sử dụng rộng rãi trong lĩnh vực xử lý âm thanh. Chúng tôi cũng trình bầy thuật toán tách nguồn âm thanh dựa trên NMF và coi đó là thuật toán cơ sở để phát triển nghiên cứu của mình. Bên cạnh đó, để kiểm chứng khả năng mô hình hóa đặc trưng phổ âm thanh của NMF, chúng tôi đề xuất phương pháp trích xuất các âm thanh bất thường xuất hiện trong file ghi âm dài. Đề xuất cho thấy một hướng ứng dụng khác của NMF, đồng thời xác thực khả năng mô hình hóa các đặc trưng phổ của tín hiệu âm thanh của NMF. Từ nhận định đó, chúng tôi sẽ đề xuất thuật toán tách nguồn đơn kênh sử dụng NMF theo hướng tiếp cận weakly-informed trong những chương sau.

Những kết quả của chương 2 được công bố trong bài báo [3] trong "**Danh mục các công trình đã công bố**" của luận án. Thuật toán trích xuất các âm thanh bất thường đề xuất đã được chuyển giao cho RION Co., Ltd., tiếp tục phát triển và sử dụng cho bài toán phát hiện và gán nhãn các sự kiện âm thanh.

CHƯƠNG 3: TÁCH NGUỒN ÂM THANH ĐƠN KÊNH SỬ DỤNG NMF VÀ RÀNG BUỘC THƯA ĐỂ KHAI THÁC MA TRẬN PHỔ TỔNG QUÁT GSSM

3.1 Sơ đồ thuật toán đề xuất

Những công bố gần đây về tách nguồn âm cho thấy thuật toán tách nguồn mù cho kết quả phân tách chưa đủ tốt để đưa vào ứng dụng thực tế. Một số thuật toán sử dụng thông tin hướng dẫn tương đối cụ thể (như tách âm nhạc khi biết trước bản nhạc, tách tiếng nói khi biết bản transcript,...) cho kết quả phân tách tốt hơn [4, 7, 8]. Tuy nhiên những thông tin chính xác đó thường không có sẵn trong nhiều tình huống. Hướng tiếp cận sử dụng thông tin hướng dẫn yếu (*weakly-informed*) là một giải pháp hiệu quả nhằm nâng cao hiệu quả tách nguồn âm trong tình huống thiếu dữ liệu huấn luyện. Trong nghiên cứu của mình, chúng tôi chỉ cần biết các tín hiệu cần tách thuộc loại âm thanh gì (như tiếng nói, âm nhạc, nhiễu môi trường,...) để tìm kiếm những mẫu âm thanh cùng loại làm dữ liệu huấn luyện. Tập mẫu huấn luyện đó được dùng để xây dựng ma trận phổ tổng quát GSSM (*general source spectral model*) của các nguồn thành phần, sau đó GSSM được dùng để hướng dẫn quá trình phân tách.



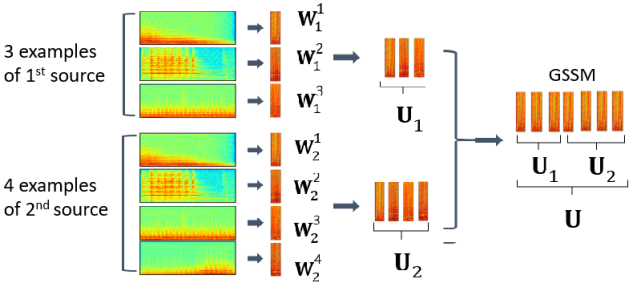
Hình 3.1: Sơ đồ thuật toán tách nguồn đơn kênh đề xuất.

Giả sử cần phân tách tín hiệu trộn bởi J nguồn, ký hiệu $\mathbf{X} \in \mathbb{C}^{F \times N}$ và $\mathbf{S}_j \in \mathbb{C}^{F \times N}$ là các ma trận phức biểu diễn tín hiệu trộn $\mathbf{x}(t)$ và tín hiệu nguồn thứ j $\mathbf{c}_j(t)$ trong miền thời gian - tần số, mục tiêu của thuật toán là ước lượng tín hiệu nguồn $\mathbf{c}_j(t)$ từ tín hiệu trộn đơn kênh $\mathbf{x}(t)$ khi không có dữ liệu huấn luyện.

Từ thông tin đã biết về loại nguồn cần phân tách, chúng tôi thu thập các mẫu huấn

luyện cùng loại. Ví dụ, tách tiếng nói bị trộn lẫn với âm thanh nhiễu môi trường, chúng tôi thu thập 3 file tiếng nói, 4 file âm thanh nhiễu khác nhau, mỗi file dài khoảng từ 5 đến 10 giây làm dữ liệu huấn luyện. Các bước của thuật toán đề xuất được mô tả trong hình 3.1: (1) học ma trận phổ tổng quát GSSM từ các mẫu huấn luyện bởi NMF, (2) phân tách các nguồn thành phần từ tín hiệu trộn qua quá trình ước lượng $\mathbf{H}$ bằng mô hình NMF hết hợp với hàm ràng buộc thưa.

3.2 Học mô hình phổ tổng quát GSSM



Hình 3.2: Ma trận phổ tổng quát GSSM.

Gọi $s_j^l(t)$ là mẫu huấn luyện thứ l của nguồn cần tách $s_j(t)$. Ở bước huấn luyện, NMF mã hóa đặc trưng phổ của từng mẫu $s_j^l(t)$ bởi ma trận $\mathbf{W}_j^l$. Sau đó, ma trận phổ tổng quát $\mathbf{U}$ được xây dựng từ các thành phần $\mathbf{W}_j^l$ như mô tả trong hình 3.2.

3.3 Ước lượng H với công thức ràng buộc thưa đề xuất

Ma trận phổ tổng quát $\mathbf{U}$ sẽ có kích thước lớn khi số mẫu huấn luyện tăng. Hơn nữa, do các mẫu huấn luyện chỉ là âm thanh cùng loại với nguồn cần tách, nên $\mathbf{U}$ có thể có nhiều đặc trưng không phù hợp với bất kỳ nguồn cần tách nào. Vì vậy, ở bước phân tách tín hiệu nguồn thành phần, ràng buộc thưa được sử dụng nhằm hướng dẫn quá trình ước lượng $\mathbf{H}$ chỉ kích hoạt những phần nhỏ từ ma trận lớn $\mathbf{U}$ chứa đặc tính phổ phù hợp với nguồn cần tách. Hàm mục tiêu khi có ràng buộc thưa được viết như sau [3]:

$$\min_{\mathbf{H}\geq 0} D(\mathbf{V}\|\mathbf{U}\mathbf{H}) + \lambda\Omega(\mathbf{H}), \tag{3.4}$$

với $\Omega(\mathbf{H})$ là hàm ràng buộc thưa tác động lên ma trận $\mathbf{H}$, λ là hằng số không âm thể hiện mức độ ảnh hưởng của ràng buộc thưa. Có hai nhóm ràng buộc thưa đã được công

Algorithm 4 Unsupervised NMF with mixed sparsity-inducing penalty

Require: $\mathbf{V}$, $\mathbf{W}$, λ, γ

Ensure: $\mathbf{H}$

Initialize $\mathbf{H}$ randomly

$\hat{\mathbf{V}} = \mathbf{UH}$

repeat

// Taking into account block sparsity-inducing penalty

for $g = 1, ..., G$ **do**

$\mathbf{Y}_{(g)} \leftarrow \frac{1}{\epsilon + \|\mathbf{H}_{(g)}\|_1}$

end for

$\mathbf{Y} = [\mathbf{Y}_{(1)}^T, \ldots, \mathbf{Y}_{(G)}^T]^T$

// Taking into account component sparsity-inducing penalty

for $k = 1, ..., K$ **do**

$\mathbf{z}_k \leftarrow \frac{1}{\epsilon + \|\mathbf{h}_k\|}$

end for

$\mathbf{Z} = [\mathbf{z}_1^T, \ldots, \mathbf{z}_K^T]^T$

// Updating activation matrix

$\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{U}^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2})}{\mathbf{U}^T(\hat{\mathbf{V}}^{\cdot -1}) + \lambda(\gamma \mathbf{Y} + (1-\gamma)\mathbf{Z})} \right)^{\cdot \frac{1}{2}}$

$\hat{\mathbf{V}} \leftarrow \mathbf{UH}$

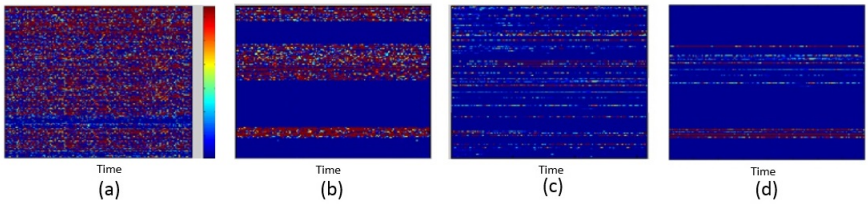
until convergence

bố là *block sparsity* và *component sparsity* như sau:

- Ràng buộc thưa Block: $\Omega_1(\mathbf{H}) = \sum_{g=1}^{G} \log(\epsilon + \|\mathbf{H}_{(g)}\|_1)$
- Ràng buộc thưa Component: $\Omega_2(\mathbf{H}) = \sum_{k=1}^{K} \log(\epsilon + \|\mathbf{h}_k\|_1)$

Chúng tôi đề xuất kết hợp hai nhóm ràng buộc thưa nêu trên bằng công thức khái quát hóa (3.7), với γ là tham số thể hiện sự đóng góp của mỗi thành phần ràng buộc thưa trong công thức kết hợp. Hình ảnh ma trận kích hoạt $mathbf H$ khi sử dụng các ràng buộc thưa khác nhau được thể hiện trong hình 3.3.

$$\Omega_{\text{new}}(\mathbf{H}) = \gamma \sum_{g=1}^{G} \log(\epsilon + \|\mathbf{H}_{(g)}\|_1) + (1-\gamma) \sum_{k=1}^{K} \log(\epsilon + \|\mathbf{h}_k\|_1), \quad (3.7)$$



Hình 3.3: Hình ảnh ma trận $\mathbf{H}$: (a) không sử dụng ràng buộc thưa, (b) với ràng buộc thưa Block, (c) với ràng buộc thưa Component, and (d) với ràng buộc thưa đề xuất.

3.4 Thuật toán tách nguồn âm thanh với hàm ràng buộc thưa mới

Sau quá trình biến đổi đạo hàm hàm giá (3.4) với hàm ràng buộc thưa đề xuất (3.7), công thức cập nhật ma trận $\mathbf{H}$ là: $\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{U}^\top (\widehat{\mathbf{V}} \odot \mathbf{V}^{\cdot -2})}{\mathbf{U}^\top (\widehat{\mathbf{V}}^{\cdot -1}) + \lambda(\gamma \mathbf{Y} + (1-\gamma)\mathbf{Z})} \right)^{\cdot \frac{1}{2}}$.

Thuật toán tách nguồn âm thanh đề xuất sử dụng mô hình phổ tổng quát và hàm ràng buộc thưa (2.7) được mô tả trong Algorithm 3. Trong đó, $\mathbf{Y}_{(g)}$ là ma trận có cùng kích thước với ma trận $\mathbf{H}_{(g)}$, $\mathbf{z}_k$ và véc tơ cùng kích thước với $\mathbf{h}_k$.

3.5 Thí nghiệm

3.5.1 Dữ liệu thí nghiệm

Bảng 3.2: Kết quả tách nguồn trên hai bộ dữ liệu Synthetic và SiSEC-MUS.

Methods		Speech/Vocals			Noise/Music		
		SDR (dB)	SIR (dB)	SAR (dB)	SDR (dB)	SIR (dB)	SAR (dB)
Synthetic	NMF non-sparsity	2.7	6.9	11.7	3.6	14.3	5.2
	NMF – Block sparsity ($\lambda = 25$)	7.4	10.2	16.4	6.9	19.8	8.5
	NMF – Component sparsity ($\lambda = 50$)	7.4	10.9	16.2	7.6	16.3	9.3
	Proposed ($\boldsymbol{\lambda = 50, \gamma = 0.2}$)	**7.7**	**10.8**	**17.8**	**7.8**	**18.7**	**9.4**
SiSEC 2016-MUS	NMF non-sparsity	1.3	3.7	7.1	3.8	9.5	11.2
	NMF – Block sparsity ($\lambda = 50$)	2.5	4.9	8.1	6.2	7.7	13.3
	NMF – Component sparsity ($\lambda = 25$)	2.7	5.6	7.3	6.2	7.7	13.5
	Proposed ($\boldsymbol{\lambda = 50, \gamma = 0.4}$)	**3.2**	**6.2**	**7.9**	**6.4**	**7.9**	**14.2**

Chúng tôi lựa chọn các file âm thanh từ 2 cơ sở dữ liệu được công bố và sử dụng rộng rãi trong cộng đồng xử lý âm thanh là DEMAND[1] và SISEC[2] cho bước học mô hình GSSM. Thuật toán được đánh giá với 3 tập dữ liệu thử nghiệm khác nhau. Trong đó tập *Synthetic* được tự tạo bằng cách trộn tín hiệu tiếng nói và âm thanh nhiễu môi trường theo tỷ lệ tín hiệu/nhiễu SNR=0. Hai tập còn lại, *SiSEC-MUS* và *SiSEC-BNG*, là dữ liệu thử nghiệm được công bố và sử dụng phổ biến trong cộng đồng tách nguồn âm.

3.5.2 Kết quả thử nghiệm

Bảng 3.3: Kết quả phân tách giọng nói thu được trên tập dữ liệu SiSEC-BGN.

Method		devset				testset						
		Ca1	Sq1	Su1	Average	Ca1	Ca2	Sq1	Sq2	Su1	Su2	Average
Martinez-Munoz (SiSEC 2013)	SDR	5.4	9.6	1.5	6.4	3.4	3.7	9.0	10.9	5.0	2.2	6.1
	SIR	15.4	17.3	5.8	14.1	14.6	17.1	18.6	20.5	23.2	5.9	17.1
	SAR	6.1	10.7	5.8	7.9	4.2	4.0	9.9	11.5	5.2	6.0	7.0
Bryan [17] (SiSEC 2013)	SDR	5.6	10.2	4.2	**7.3**	3.7	3.8	13.1	12.9	5.6	5.6	**7.8**
	SIR	18.4	15.6	13.6	**16.1**	13.9	16.5	21.8	18.2	21.4	23.0	**18.5**
	SAR	5.9	12.1	4.9	**8.4**	4.5	4.2	13.7	14.6	5.7	5.7	**8.5**
López (SiSEC 2015)	SDR	-	-	-	-	4.0	4.5	5.1	11.0	-3.8	3.9	4.9
	SIR	-	-	-	-	14.9	16.1	9.6	16.3	-1.6	8.8	12.1
	SAR	-	-	-	-	4.7	5.0	8.6	13.0	4.3	6.3	7.3
Liu (SiSEC 2016)	SDR	1.9	-3	-10.6	-3.1	1.6	2.7	-4.4	1.9	-12.6	-1.2	-1.0
	SIR	4	-2.9	-9.7	-2.1	4.5	7.7	-4.3	2.4	-12.2	0.1	0.9
	SAR	7.5	16.4	6.9	11.3	6.5	5.5	18.8	16.9	10.3	8	11.4
Proposed (SiSEC 2016)	SDR	**5.6**	**9.3**	**4.1**	**6.9**	**3.7**	**4.3**	**10.1**	**11.6**	**5.3**	**4.2**	**6.9**
	SIR	**14.9**	**15.4**	**12.1**	**14.5**	**13.2**	**15**	**17.9**	**18.2**	**19.3**	**9.3**	**15.7**
	SAR	**6.3**	**10.7**	**5.3**	**8.0**	**4.8**	**4.9**	**11.1**	**12.7**	**5.5**	**6.6**	**7.9**

Kết quả thí nghiệm trên hai tập dữ liệu Synthetic và SiSEC-MUS trong bảng 3.2 cho thấy: Kết quả của thuật toán "NMF -without training" là thấp nhất, chứng tỏ thuật toán tách nguồn âm thanh dựa trên NMF cơ bản được mô tả trong chương 2 không phân tách tốt khi thiếu dữ liệu huấn luyện. Kết quả của 3 thuật toán sử dụng nhóm ràng buộc thưa tốt hơn nhiều so với thuật toán "NMF non-sparsity". Điều đó cho thấy vai trò quan trọng của nhóm ràng buộc thưa trong quá trình ước lượng nguồn thành phần. Cuối cùng, thuật toán đề xuất cho kết quả tốt nhất và tốt hơn 2 thuật toán sử dụng hai hàm ràng buộc thưa trước đó. Kết quả này khẳng định đề xuất kết hợp hai thành phần ràng buộc thưa đã nâng cao đáng kể hiệu quả tách nguồn âm.

[1]http://parole.loria.fr/DEMAND/
[2]http://sisec.wiki.irisa.fr.

Kết quả của thuật toán đề xuất đã được gửi tham gia SiSEC năm 2016. So sánh với thuật toán của Liu cùng tham gia năm đó, thuật toán đề xuất cho kết quả tốt hơn ở hai độ đo SDR và SIR, đặc biệt là cho kết quả vượt trội trên độ đo tổng thể quan trọng nhất SDR. Thuật toán đề xuất được đánh giá tốt hơn thuật toán của Liu bởi ban tổ chức SiSEC 2016 [4].

Mở rộng so sánh với các thuật toán tách nguồn đơn kênh khác đã tham gia SiSEC từ năm 2013 cho đến nay, bảng 3.3 cho thấy kết quả của thuật toán đề xuất kém hơn so với thuật toán của López nhưng tốt hơn tất cả các thuật toán còn lại. Tuy nhiên thuật toán của López sử dụng chú thích của người dùng trên phổ của tín hiệu trộn để hướng dẫn tách nguồn. Thuật toán này sẽ không thể thực hiện được nếu không có sự tham gia của một chuyên gia âm thanh.

3.6 Tổng kết

Trong chương 3, chúng tôi đã đề xuất một thuật toán tách nguồn âm thanh đơn kênh khi không có dữ liệu huấn luyện chính xác cho các nguồn cần tách. Những đóng góp cụ thể hơn gồm:

- Đề xuất thuật toán mới phân tách các âm thanh thành phần từ tín hiệu trộn đơn kênh.
- Đề xuất công thức kết hợp hai nhóm ràng buộc thưa thành dạng tổng quát, có sự đóng góp của cả hai thành phần ràng buộc thưa trước đó.
- Chúng tôi đã xem xét khả năng hội tụ của thuật toán đề xuất theo số vòng lặp MU, tính ổn đinh cũng như hiệu quả phân tách của thuật toán thông qua 3 bộ dữ liệu thí nghiệm. Kết quả của thuật toán đề xuất đã được gửi tham gia chiến dịch SiSEC năm 2016.

Trong chương tiếp theo, chúng tôi sẽ đề xuất mở rộng thuật toán cho trường hợp đa kênh bằng cách kết hợp mô hình NMF với mô hình Gaussian cục bộ.

Những kết quả của chương 3 được công bố trong 4 bài báo [1], [2], [4] và [5] trong "**Danh mục các công trình đã công bố**" của luận án.

CHƯƠNG 3: TÁCH NGUỒN ÂM THANH ĐA KÊNH SỬ DỤNG KẾT HỢP NMF TRONG MÔ HÌNH GAUSSIAN CỤC BỘ

4.1 Mô hình hóa bài toán tách nguồn đa kênh

4.1.1 Mô hình Gaussian cục bộ

Gọi $\mathbf{x}(t)$ là tín hiệu trộn của J nguồn âm được thu âm bởi mảng I microphones được biểu diễn trong công thức (1.1), tách nguồn âm thanh đa kênh là vấn đề ước lượng các tín hiệu nguồn thành phần $\mathbf{c}_j(t)$ từ tín hiệu đầu vào $\mathbf{x}(t)$.

Trong mô hình Gaussian cục bộ (LGM), tín hiệu nguồn thành phần trong miền T-F, ký hiệu là $\mathbf{c}_j(n,f)$, được biểu diễn theo chuẩn phân bố Gaussian với trung bình bằng 0 và ma trận hiệp phương sai $\mathbf{\Sigma}_j(n,f) = \mathbb{E}(\mathbf{c}_j(n,f)\mathbf{c}_j^H(n,f))$ như sau:

$$\mathbf{c}_j(n,f) \sim \mathcal{N}_c(\mathbf{0}, \mathbf{\Sigma}_j(n,f)), \tag{4.1}$$

với $\mathbf{0}$ là véc tơ 0 kích thước $I \times 1$, $(.)^H$ biểu diễn phép chuyển vị liên hợp (conjugate transposition). Ma trận hiệp phương sai được xác định gồm hai thành phần:

$$\mathbf{\Sigma}_j(n,f) = v_j(n,f)\,\mathbf{R}_j(f), \tag{4.2}$$

trong đó $v_j(n,f)$ là phương sai nguồn (*source variance*) mã hóa sự thay đổi về năng lượng phổ của nguồn âm và là tham số phụ thuộc thời gian t. $\mathbf{R}_j(f)$ là ma trận hiệp phương sai không gian (*spatial covariance*) kích thước $I \times I$ mã hóa các đặc tính không gian giữa nguồn và microphone, tham số này không phụ thuộc t khi các nguồn và microphone không di chuyển. Việc ước lượng nguồn thành phần $\mathbf{c}_j(t)$ được thực hiện bằng cách ước lương hai thành phần $v_j(n,f)$ và $\mathbf{R}_j(f)$.

4.1.2 Mô hình phương sai nguồn dựa trên NMF

Khi kết hợp NMF trong mô hình LGM, phương sai nguồn $v_j(n,f)$ được phân tách theo NMF bởi công thức $v_j(n,f) = \sum_{k=1}^{K_j} w_{jfk} h_{jkn}$. Trong đó w_{jfk} là phần tử của ma trận đặc trưng phổ $\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$, h_{jkn} là phần tử của ma trận kích hoạt $\mathbf{H}_j \in \mathbb{R}_+^{K_j \times N}$, K_j là số lượng đặc trưng phổ được mã hóa.

4.1.3 Ước lượng các tham số

Các thành phần $v_j(n,f)$ và $\mathbf{R}_j(f)$ được ước lượng qua các vòng lặp EM, mỗi vòng lặp gồm hai bước xử lý: bước E và bước M. Trong bước E, thuật toán cập nhật các tham

số $\theta = \{v_j(n,f), \mathbf{R}_j(f)\}_{j,n,f}$ theo công thức:

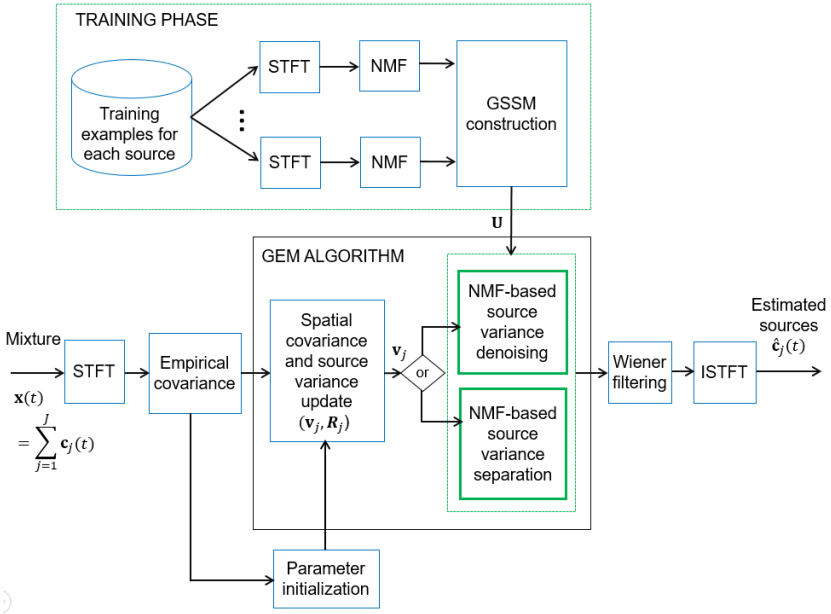
$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_j(n,f)}\widehat{\mathbf{\Sigma}}_j(n,f), \tag{4.11}$$

$$v_j(n,f) = \frac{1}{I}\text{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f)). \tag{4.12}$$

Khi kết hợp NMF trong mô hình LGM, tại bước M của mỗi vòng lặp EM, vòng lặp MU của mô hình NMF sẽ cập nhật $v_j(n,f)$ theo công thức $v_j(n,f) = \sum_{k=1}^{K_j} w_{jfk}h_{jkn}$.

4.2 Thuật toán tách nguồn đa kênh đề xuất

Mô hình thuật toán đề xuất được thể hiện trong hình. 4.1. Trong pha huấn luyện, ma trận phổ tổng quát GSSM được học từ các mẫu huấn luyện như mô tả trong phần 3.2. Ở pha phân tách, hai thành phần $v_j(n,f)$ và $\mathbf{R}_j(f)$ được ước lượng bằng thuật toán tối ưu hóa kỳ vọng tổng quát (generalized expectation minimization - GEM), trong đó có sự kết hợp của mô hình NMF khai thác ma trận GSSM trong bước M.



Hình 4.1: Sơ đồ thuật toán tách nguồn đa kênh đề xuất.

Trong chương 3, chúng tôi đã đề xuất công thức kết hợp hai nhóm ràng buộc thưa trong bước ước lượng ma trận $\mathbf{H}$ bởi NMF. Kết hợp với mô hình LGM, chúng tôi đề

xuất hai tiêu chí tối ưu hóa mới để hướng dẫn ước lượng phương sai nguồn trung gian trong mỗi vòng lặp EM như sau:

- **Source variance denoising**: ước lượng phương sai của từng nguồn riêng biệt bằng NMF kết hợp với ràng buộc thưa đề xuất, công thức tối ưu hóa ma trận phương sai của từng nguồn được viết như sau:

$$\min_{\widetilde{\mathbf{H}}_j \geq 0} D(\mathbf{V}_j \| \mathbf{U}_j \widetilde{\mathbf{H}}_j) + \lambda \Omega(\widetilde{\mathbf{H}}_j). \tag{4.19}$$

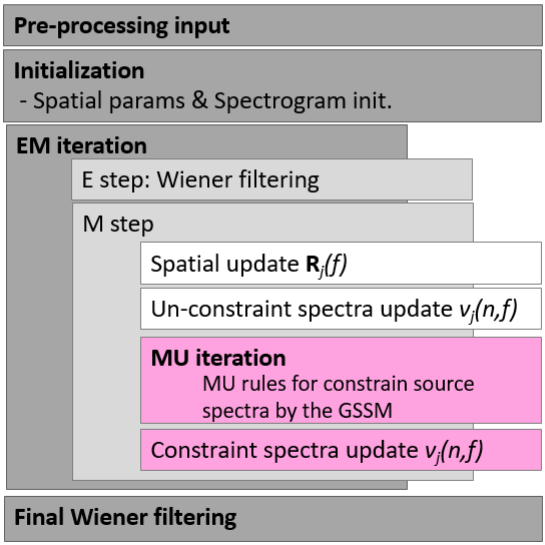
- **Source variance separation**: Gọi $\widetilde{\mathbf{V}} = \sum_{j=1}^{J} \mathbf{V}_j$ là ma trận phương sai của tổng các nguồn thành phần, tiêu chí thứ hai tối ưu hóa ma trận phương sai tổng thể của tất cả các nguồn thành phần như sau:

$$\min_{\widetilde{\mathbf{H}} \geq 0} D(\widetilde{\mathbf{V}} \| \mathbf{U}\widetilde{\mathbf{H}}) + \lambda \Omega(\widetilde{\mathbf{H}}). \tag{4.20}$$

Algorithm 6 Proposed GSSM + SV separation algorithm

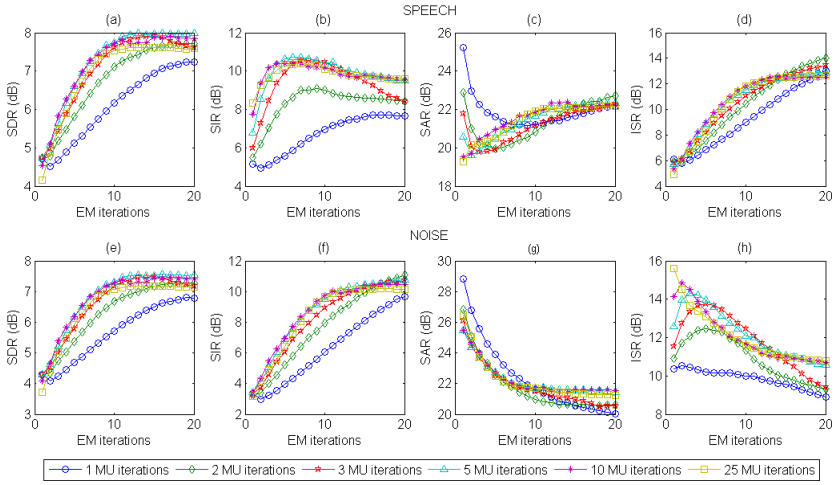
Require:
Mixture signal $\mathbf{x}(t)$
List of examples of each source in the mixture $\{s_j^l(t)\}_{j=1:J,l=1:L_j}$
Hyper-parameters λ, γ, MU-iteration

Ensure: Source images $\hat{\mathbf{c}}_j(t)$ separated from $\mathbf{x}(t)$



Công thức cập nhật $\widetilde{\mathbf{H}}$ cho công thức tối ưu hóa thứ 2 là $\widetilde{\mathbf{H}} \leftarrow \widetilde{\mathbf{H}} \odot \Big(\frac{\mathbf{U}^\top (\widehat{\mathbf{V}} \odot \widehat{\mathbf{V}}^{\cdot -2})}{\mathbf{U}^\top (\widehat{\mathbf{V}}^{\cdot -1}) + \lambda(\gamma \mathbf{Y} + (1-$
Công thức này dùng để cập nhật $v_j(n,f)$ trong vòng lặp MU tại bước M. Các bước chi tiết của thuật toán đề xuất được thể hiện trong Algorithm 6.

4.3 Thí nghiệm



Hình 4.2: Sơ đồ tương quan của hiệu suất tách nguồn theo số vòng lặp EM và MU.

4.3.1 Dữ liệu thí nghiệm

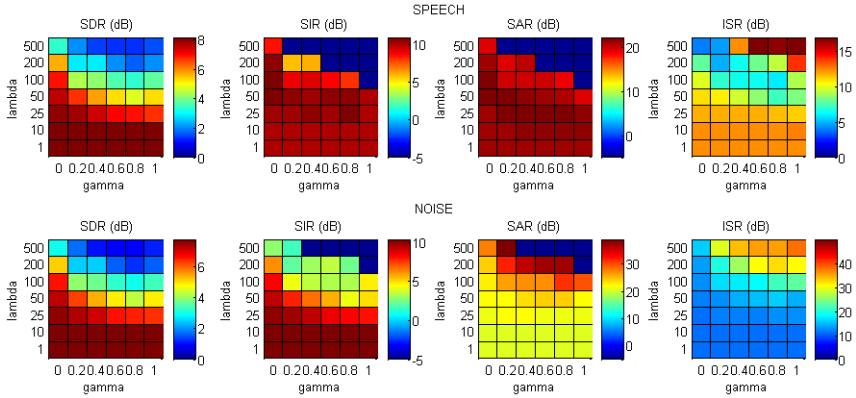
Thuật toán đề xuất được đánh giá bằng thí nghiệm trên tập dữ liệu devset của SiSEC2016-BGN[1]. Tập devset gồm 9 file tín hiệu trộn của tiếng nói và âm thanh nhiễu môi trường, mỗi file dài 10 giây.

4.3.2 Sự hội tụ và ổn định của thuật toán

Sự hội tụ của thuật toán: Hình 4.2 cho thấy thuật toán hội tụ khá tốt với 10 hoặc 25 vòng lặp MU, và đạt giá trị bão hòa sau khoảng 10 vòng lặp EM. Điều này thể hiện sự ảnh hưởng tốt của mô hình NMF trong mô hình LGM.

Kết quả phân tách với các giá trị khác nhau của λ và γ: Hình 4.3 cho thấy kết quả tách giảm nhanh chóng với $\lambda > 25$. Kết quả tốt nhất trên SDR được quan sát thấy

[1] https://sisec.inria.fr/sisec-2016/bgn-2016/



Hình 4.3: Sơ đồ tương quan của hiệu suất tách nguồn theo các tham số λ và γ.

ứng với $\lambda = 10$ và $\gamma = 0.2$. Với nhưng giá trị λ nhỏ, sự thay đổi của γ ít ảnh hưởng đối với kết quả tách và thuật toán cho kết quả khá ổn định. Những phân tích trên thể hiện công thức kết hợp hai nhóm ràng buộc thưa đề xuất làm việc hiệu quả trong mô hình LGM.

4.3.3 Kết quả thí nghiệm

Kết quả thu được của thuật toán đề xuất được so sánh với kết quả của thuật toán Arberet's (là thuật toán cơ sở của thuật toán đề xuất) và những thuật toán từng tham gia SiSEC từ năm 2013 cho đến nay.

Điều thú vị là thuật toán đề xuất khi không có điều kiện ràng buộc thưa cho kết quả thấp hơn thuật toán của Arberet. Điều này một lần nữa khẳng định sự dư thừa của GSSM. Thuật toán "GSSM + SV denoising" cho kết quả tốt hơn Arberet (ngoại trừ ISR và TPS) cho thất việc khai thác GSSM trong bước phân tách giúp tăng đáng kể hiệu quả tách nguồn. Thuật toán "GSSM + SV separation" cho kết quả tốt nhất với SDR, SIR, OPS, IPS, khi so sánh với "GSSM + SV denoising" và "GSSM' + component sparsity", khẳng định hiệu quả của tiêu chí tối ưu hóa trên tổng thể các nguồn (4.20).

Khi so sánh với các thuật toán khác tham gia SiSEC trong nhiều năm, kết qủa cho thấy thuật toán đề xuất tốt với nhóm tiêu chí dựa trên năng lượng, nhưng kém hơn với nhóm tiêu chí dựa trên sự cảm thụ của tai người. Xem xét độ đo quan trọng nhất SDR, thuật toán "GSSM + SV separation" cho kết quả kém hơn thuật toán của Wang nhưng tốt hơn các thuật toán còn lại. Điều này khẳng định thuật toán đề xuất đã nâng cao hiệu xuất tách nguồn âm như mục tiêu đặt ra và khẳng định sự kết hợp thành công của NMF và LGM. Lưu ý rằng sau khi dùng thuật toán phân tách, Wang đã sử dụng kỹ thuật xử lý lọc nhiễu để nâng cao chất lượng tín hiệu tiếng nói tách được. Hơn nữa, thuật toán

Bảng 4.1. Kết quả phân tách giọng nói trên tập dữ liệu SiSEC-BGN.

Methods		BSS Eval				PEASS			
		SDR	SIR	SAR	ISR	OPS	IPS	APS	TPS
Wang*	SiSEC 2013	**9.8**	20.0	12.0	13.5	**37.9**	54.8	50.7	54.1
Le Magoarou*		3.7	5.6	8.8	17.8	32.8	35.3	43.1	**65.9**
Rafii*		5.1	8.0	9.0	11.6	30.4	31.1	55.5	56.9
Ito*	SiSEC 2015	7.4	**22.6**	7.7	-	-	-	-	-
Liu*	SiSEC 2016	-7.0	-1.4	15.0	3.1	14.2	17.5	70.3	42.3
Wood*		1.9	3.6	3.7	5.1	34.1	**57.7**	39.3	44.5
Arberet (2012)		4.4	4.6	12.1	15.9	10.4	9.1	72.6	43.2
GSSM + SV separation (No sparsity constraint)		1.8	1.5	19.1	8.9	18.8	22.0	58.2	53.3
GSSM + SV separation (GSSM + component sparsity)		4.9	6.5	17.7	8.3	21.3	22.3	60.3	49.1
GSSM + SV denoising ($\lambda = 10, \gamma = 0.2$)		**7.7**	**9.0**	**23.6**	**11.6**	**18.1**	**12.5**	**75.9**	**40.1**
GSSM + SV separation ($\lambda = 10, \gamma = 0.2$)		**8.1**	**11.0**	**21.3**	**14.1**	**23.1**	**24.5**	**65.2**	**54.2**

của Wang sử dụng kỹ thuật phân tích ICA, do đó không áp dụng được cho trường hợp số nguồn âm nhiều hơn số microphone. Trong khi thuật toán đề xuất vẫn có thể áp dụng được trong trường hợp này.

4.4 Tổng kết

Chương 4 mô tả thuật toán tách nguồn âm đa kênh mới theo hướng tiếp cận "weakly-informed". Thuật toán đề xuất sử dụng mô hình phổ tổng quát được học bởi NMF kết hợp trong mô hình LGM. Kết quả cụ thể như sau:

- Chúng tôi đã đề xuất hai tiêu chí tối ưu hóa mới cho quá trình ước lượng của vòng lặp EM, tính toán công thức cập nhật tham số tương ứng với từng tiêu chí và xây dựng thuật toán tách nguồn đa kênh.

- Thí nghiệm được thực hiện trên tập dữ liệu từ website uy tín SiSEC đã xác thực tính ổn định, sự hội tụ và hiệu quả tách nguồn của thuật toán đề xuất. Chúng tôi cũng gửi kết quả thuật toán tham gia chiến dịch SiSEC 2016, đánh giá từ ban tổ chức cho thấy thuật toán đề xuất cho kết quả tốt nhất với bộ tiêu chí dự trên năng lượng, so với các thuật toán cùng tham gia năm đó.

Những kết quả của chương 4 được công bố trong 2 bài báo [6] và [7] trong "**Danh mục các công trình đã công bố**" của luận án.

KẾT LUẬN

Có rất nhiều tình huống trong thực tế mà âm thanh thu được là hỗn hợp trộn của nhiều nguồn âm thanh khác nhau. Con người với khả năng thính giác bình thường có thể dễ dàng xác định được âm thanh mục tiêu để nghe, hiểu. Nhưng đối với học máy thì nhiệm vụ này lại vô cùng khó khăn.

Chúng tôi nghiên cứu hướng tiếp cận sử dụng thông tin hướng dẫn ít (weakly-informed approach) để phân tách các âm thanh bị trộn lẫn trong hỗn hợp. Trong đó, mô hình phổ tổng quát GSSM được huấn luyện từ một vài ví dụ mẫu cùng loại với âm thanh cần phân tách bởi quá trình ước lượng của thuật toán NMF. Chúng tôi đề xuất một công thức ràng buộc thưa mới cho bước ước lượng các tham số. Đồng thời chúng tôi tính toán công thức cập nhật tham số theo hàm ràng buộc thưa mới đề xuất và xây dựng thuật toán tách các âm thanh thành phần từ tín hiệu trộn đơn kênh. Thí nghiệm được thực hiện với các cài đặt khác nhau trên ba bộ dữ liệu đã cho thấy hiệu quả của thuật toán đơn kênh đề xuất.

Từ thuật toán đơn kênh, chúng tôi phát triển cho trường hợp đa kênh, kết hợp mô hình phổ tổng quát GSSM với mô hình hiệp phương sai không gian của các nguồn âm trong khuôn khổ mô hình Gaussian (LGM). Trong mô hình LGM, các tham số được ước lượng bằng thuật toán tối ưu hóa kỳ vọng EM. Để hướng dẫn ước lượng phương sai nguồn trung gian trong mỗi vòng lặp EM, chúng tôi đề xuất hai tiêu chí tối ưu hóa: (1) ước lượng phương sai của từng nguồn riêng biệt bằng mô hình NMF kết hợp với ràng buộc thưa đề xuất, (2) ước lượng phương sai của tất cả các nguồn đồng thời bằng mô hình NMF kết hợp với ràng buộc thưa đề xuất. Tiêu chí thứ hai được xem như một bước tách được thực hiện bổ sung cho phương sai nguồn. Hiệu suất phân tách của thuật toán đề xuất cũng như khả năng hội tụ và tính ổn định của thuật toán được kiểm chứng qua thí nghiệm được thực hiện trên bộ dữ liệu SiSEC, được công bố và sử dụng rộng rãi trong cộng đồng xử lý âm thanh.

Bên cạnh hai đóng góp chính nêu trên, chúng tôi đề xuất thuật toán sử dụng NMF tự động trích xuất những đoạn âm thanh bất thường từ tín hiệu thu âm đơn kênh kích thước lớn. Đóng góp này nhằm mục đích hỗ trợ quá trình phát hiện và gán nhãn các sự kiện âm thanh. Sau khi trích xuất được những sự kiện âm từ dữ liệu, người gán nhãn sẽ chỉ cần nghe và gán nhãn tại vị trí xuất hiện đoạn âm thanh bất thường đã được thuật toán phát hiện, thay vì nghe toàn bộ file âm thanh dài. Thí nghiệm thực hiện trên bộ dữ liệu thu âm trong môi trường ngoài trời, được cũng cấp bởi công ty RION, Nhật Bản. Kết quả thí nghiệm đã kiểm chứng khả năng mô hình hóa tốt các đặc tính phổ của NMF.

Hướng phát triển trong tương lai:

- Kiểm chứng hiệu quả của các thuật toán đề xuất trên hệ thống nhận dạng tiếng nói tự động ASR.

- Phát triển ý tưởng phân tách phương sai nguồn thành phần sử dụng mô hình DNN, dựa trên việc tìm hiểu kết quả nghiên cứu của nhóm Nugraha [6].
- Phát triển từ ý tưởng sử dụng mô hình phổ tổng quát GSSM, nghiên cứu xây dựng mô hình hiệp phương sai không gian tổng quát cho các nguồn trong hỗn hợp.
- Kết hợp thuật toán đề xuất với các kỹ thuật khác như: loại nhiễu (dereverberation), source localization, post-filtering,...nhằm xây dựng hệ thống phân tách âm thanh đạt hiệu quả phân tách cao hơn nữa.

TÀI LIỆU THAM KHẢO

[1] Févotte, C., Bertin, N., and Durrieu, J. (2009). *Non-negative matrix factorization with the itakura-saito divergence. With application to music analysis*. Neural Computation, 21(3):793–830.

[2] Lee, D. D. and Seung, H. S. (2001). *Algorithms for non-negative matrix factorization*. In Advances in Neural and Information Processing Systems 13, pages 556–562.

[3] Lefèvre, A., Bach, F., and Févotte, C. (2011). *Itakura-Saito non-negative matrix factorization with group sparsity*. In IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pages 21–24.

[4] Liutkus, A., St¨oter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., and Fontecave, J. (2017). *The 2016 Signal Separation Evaluation Campaign*. In Latent Variable Analysis and Signal Separation, volume 10169, pages 323–332. Springer International Publishing, Cham.

[5] Makino, S., Lee, T.-W., and Sawada, H. (2007). *Blind Speech Separation*. Springer.

[6] Nugraha, A., Liutkus, A., and Vincent, E. (2016). *Multichannel audio source separation with deep neural networks*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 14(9):1652–1664.

[7] Ono, N., Koldovský, Z., Miyabe, S., and Ito, N. (2013). *The 2013 Signal Separation Evaluation Campaign*. In 2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pages 1–6.

[8] Ono, N., Rafii, Z., Kitamura, D., Ito, N., and Liutkus, A. (2015). *The 2015 Signal Separation Evaluation Campaign*. In Latent Variable Analysis and Signal Separation, volume 9237, pages 387–395. Springer International Publishing, Cham.

DANH MỤC
CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ CỦA LUẬN ÁN

1. **Hien-Thanh Thi Duong**, Quoc-Cuong Nguyen, Cong-Phuong Nguyen, Thanh Huan Tran, and Ngoc Q. K. Duong (2015). *Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint.* Proc. ACM International Symposium on Information and Communication Technology (SoICT 2015), pp. 247-251, Hue, Vietnam. ISBN: 978-1-4503-3843-1, DOI:10.1145/2833258.2833276.
2. **Hien-Thanh Thi Duong**, Quoc-Cuong Nguyen, Cong-Phuong Nguyen, and Ngoc Q. K. Duong (2016). *Single-channel speaker-dependent speech enhancement exploiting generic noise model learned by non-negative matrix factorization.* Proc. IEEE International Conference on Electronics, Information and Communication (ICEIC 2016), pp. 268-271, Danang, Vietnam, ISBN 978-1-4673-8016-4, DOI 10.1109/ELINFOCOM.2016.7562952.
3. **Thanh Thi Hien Duong**, Nobutaka Ono, Yasutaka Nakajima and Toshiya Ohshima (2016). *Non-stationary Segment Detection Methods based on Single-basis Non-negative Matrix Factorization for Effective Annotation.* Proc. IEEE Asia-Pacific Signal and Information Processing Association Annual Summit Conference (APSIPA ASC 2016), pp. 1-6, Jeju, Korea, ISBN 978-9-8814-7682-1, DOI 10.1109/APSIPA.2016.7820760.
4. **Thanh Thi Hien Duong**, Phuong Cong Nguyen, and Cuong Quoc Nguyen (2018). *Exploiting Nonnegative Matrix Factorization with Mixed Group Sparsity Constraint to Separate Speech Signal from Single-channel Mixture with Unknown Ambient Noise.* EAI Endorsed Transactions on Context-Aware Systems and Applications. Vol. 18(13), pp: 1-8. ISSN 2409-0026.
5. **Dương Thị Hiền Thanh**, Nguyễn Công Phương, Nguyễn Quốc Cường (2018). *Kết hợp mô hình thừa số hóa ma trận không âm với các ràng buộc thưa để khai thác mô hình phổ tổng quát trong bài toán tách nguồn âm thanh đơn kênh.* Tạp chí Nghiên cứu Khoa học và Công nghệ quân sự. Số 45, tháng 4 năm 2018, trang 83 - 94. ISSN 1859-1043.
6. **Thanh Thi Hien Duong**, Ngoc Q. K. Duong, Phuong Cong Nguyen, and Cuong Quoc Nguyen (2018). *Multichannel source separation exploiting NMF-based generic source spectral model in Gaussian mod-*

eling framework. In Latent Variable Analysis and Signal Separation, vol. 10891, pp. 547-557. Springer International Publishing. DOI 10.10-07/978-3-319-93764-9_50 (SCOPUS).

7. **Thanh Thi Hien Duong**, Ngoc Q. K. Duong, Phuong Cong Nguyen, and Cuong Quoc Nguyen (2019). *Gaussian modeling-based multi-channel audio source separation exploiting generic source spectral model.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27(1), pp. 32-43. ISSN 2329-9304, DOI 10.1109/TASLP.-2018.28 69692 (ISI - Q1).