

DANH SÁCH NHỮNG THÀNH VIÊN THAM GIA NGHIÊN CỨU ĐỀ TÀI VÀ ĐƠN VỊ PHỐI HỢP CHÍNH

TT.	Họ và tên	Đơn vị công tác	Lĩnh vực chuyên môn	Ghi chú
1	Dương Thị Hiền Thanh (Chủ nhiệm đề tài)	Trường DH Mỏ - Địa chất	Công nghệ thông tin	
2	Nguyễn Thu Hằng (Thành viên chính)	Trường DH Mỏ - Địa chất	Tin học kinh tế	
3	Vũ Thị Kim Liên (Thành viên)	Trường DH Mỏ - Địa chất	Công nghệ thông tin	
4	Võ Thị Thu Trang (Thành viên)	Trường DH Mỏ - Địa chất	Công nghệ thông tin	
5	Phạm Thị Nguyệt (Kỹ thuật viên)	Trường DH Mỏ - Địa chất	Tin học kinh tế	

MỤC LỤC

DANH SÁCH NHỮNG THÀNH VIÊN THAM GIA NGHIÊN CỨU ĐỀ TÀI VÀ ĐƠN VỊ PHỐI HỢP CHÍNH	i
MỤC LỤC	ii
DANH SÁCH BẢNG	iv
DANH SÁCH HÌNH VẼ	v
DANH MỤC CÁC TỪ VIẾT TẮT	vii
THÔNG TIN KẾT QUẢ NGHIÊN CỨU	viii
MỞ ĐẦU	1
 Chapter 1. TỔNG QUAN VỀ TÁCH NGUỒN ÂM THANH	 9
1.1 Mô hình tổng quát của bài toán tách nguồn âm thanh	9
1.2 Khái quát về các kỹ thuật tách nguồn âm thanh	10
1.2.1 Một số khái niệm cơ bản trong tách nguồn âm thanh	11
1.2.2 Các kỹ thuật tách nguồn âm thanh	13
 Chapter 2. PHÂN TÁCH TIẾNG NÓI SỬ DỤNG MÔ HÌNH NONNEGATIVE MA- TRIX FACTORIZATION	 17
2.1 Đặt vấn đề	17
2.2 Tổng quan về NMF	18
2.2.1 NMF là gì?	18
2.2.2 Hàm giá	20
2.3 Phân tách tiếng nói dựa trên mô hình NMF	20
2.3.1 Phân tích đặc trưng phổ của âm thanh	20
2.3.2 NMF-based audio source separation	21
2.4 Kết luận	25
 Chapter 3. THUẬT TOÁN PHÂN TÁCH TIẾNG NÓI KẾT HỢP MÔ HÌNH THỪA SỐ HÓA MA TRẬN KHÔNG ÂM VỚI RÀNG BUỘC THỪA	 27
3.1 Đặt vấn đề	27
3.2 Xây dựng mô hình phổ tổng quát GSSM	28
3.3 Các ràng buộc thừa Block sparsity và Component sparsity	29

3.3.1	Block sparsity	30
3.3.2	Component sparsity	31
3.4	Đề xuất công thức ràng buộc thừa kết hợp	31
3.5	Xây dựng mô hình thuật toán	32
3.6	Thuật toán đề xuất cho trường hợp Weakly-informed	32
3.7	Thuật toán đề xuất cho trường hợp Semi-supervised	33
3.8	Kết luận	35
Chapter 4. THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ		37
4.1	Phương pháp đánh giá kết quả phân tách âm thanh	37
4.2	Dữ liệu thử nghiệm	39
4.2.1	Bộ dữ liệu Synthetic	39
4.2.1.1	Bộ dữ liệu SiSEC-MUS	40
4.2.2	Bộ dữ liệu SiSEC-BNG	40
4.3	Kết quả thực nghiệm và đánh giá	41
4.3.1	Kết quả thử nghiệm thuật toán Weakly-informed	41
4.3.1.1	So sánh với các thuật toán cơ sở	41
4.3.1.2	Mở rộng so sánh với các thuật tách nguồn âm thanh khác	43
4.3.1.3	Khảo sát sự hội tụ và tính ổn định của thuật toán	45
4.3.2	Kết quả thử nghiệm thuật toán Semi-Supervised	46
4.4	Kết luận	49
KẾT LUẬN VÀ KIẾN NGHỊ		50
TÀI LIỆU THAM KHẢO		52
DANH SÁCH BÀI BÁO		62

DANH SÁCH BẢNG

4.1	Dữ liệu thử nghiệm SiSEC-MUS.	40
4.2	Kết quả tách nguồn trên bộ dữ liệu Synthetic và SiSEC-MUS.	42
4.3	Kết quả tách tiếng nói trên bộ dữ liệu SiSEC-BGN.	44
4.4	Kết quả phân tách tiếng nói trên bộ dữ liệu Synthetic với trường hợp Semisupervised.	47

DANH SÁCH HÌNH VẼ

1	Thu âm trong môi trường nhiễu	1
2	Kỹ thuật Beamforming [27]	3
3	Kỹ thuật triệt tiêu tiếng vọng.	4
1.1	Hệ thống tách nguồn âm thanh.	10
1.2	Các loại môi trường thu âm thực.	11
1.3	Hình ảnh minh họa môi trường thu âm thực.	11
1.4	Tín hiệu thu âm trong môi trường có và không có phản xạ âm.	12
1.5	Mô hình tổng quan về lịch sử phát triển các kỹ thuật tách nguồn âm thanh.	13
1.6	Chênh lệch cường độ của 2 tín hiệu tại 2 microphone trong môi trường không phản xạ.	14
1.7	Từ tách nguồn mù đến tách nguồn có hướng dẫn [76].	15
1.8	Từ tách nguồn mù đến tách nguồn có hướng dẫn [5].	15
2.1	Mô hình phân tách NMF [13].	19
2.2	Mô hình NMF trích chọn đặc trưng phổ của âm thanh ($K = 2$).	21
2.3	Mô hình tách tiếng nói và nhiễu theo thuật toán Supervised-NMF.	22
2.4	Mô hình NMF phân tách tín hiệu âm thanh.	25
3.1	Cách xây dựng mô hình phổ tổng quát GSSM.	30
3.2	Hình ảnh của ma trận activation $\mathbf{H}$: (a) NMF không dùng sparsity constraint, (b) NMF với block sparsity (3.6), (c) NMF với component sparsity (3.7), và (d) Kết hợp block sparsity và component sparsity (3.8).	31
3.3	Mô hình thuật toán UnSupervised-NMF phân tách tiếng nói và nhiễu sử dụng kết hợp mô hình NMF với nhóm ràng buộc thừa đề xuất.	33
3.4	Mô hình tách tiếng nói và nhiễu theo thuật toán SemiSupervised-NMF kết hợp với mixed sparsity constraints.	35
4.1	Kết quả phân tách tiếng nói trên bộ dữ liệu SiSEC-BGN-devset.	45
4.2	Kết quả phân tách tiếng nói trên bộ dữ liệu SiSEC-BGN-testset.	46
4.3	Khảo sát sự hội tụ của thuật toán theo số vòng lặp MU.	47

4.4	Khảo sát sự ảnh hưởng của hai tham số λ và γ với kết quả phân tách. .	48
4.5	Hình ảnh phổ tín hiệu trước và sau khi tách nguồn đơn kênh.	48
4.6	Hình ảnh phổ tín hiệu trước và sau khi tách nguồn đa kênh.	49

DANH MỤC CÁC TỪ VIẾT TẮT

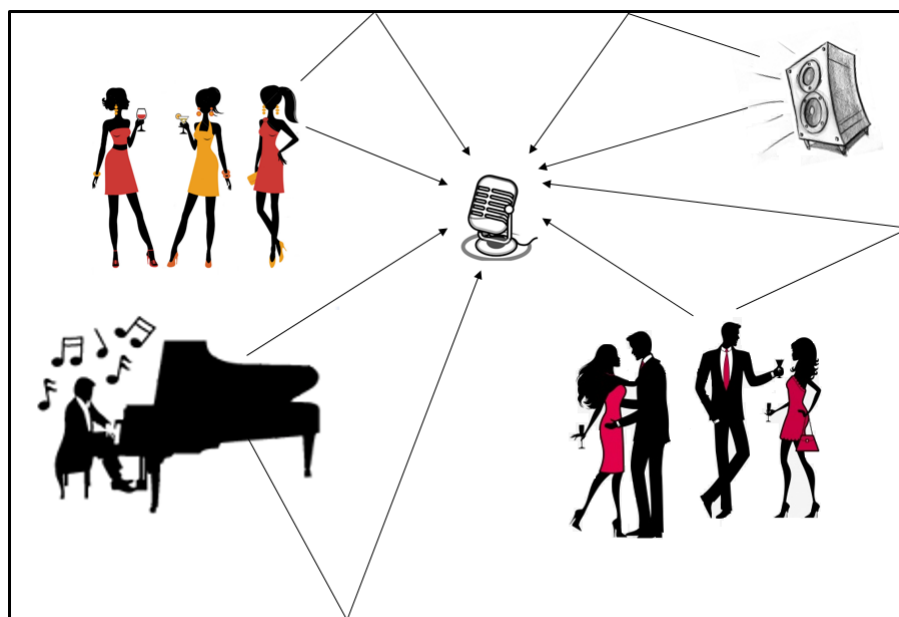
TT	Từ viết tắt	Từ nguyên mẫu	Nghĩa tiếng Việt
1	ASR	Automatic Speech Recognition	Nhận dạng tiếng nói tự động
2	DNN	Deep Neural Network	Mạng thần kinh sâu
3	EUC	Euclidean	Độ đo Euclidean
4	ICA	Independent Component Analysis	Phân tích thành phần độc lập
5	IS	Itakura-Saito	Độ đo Itakura-Saito
6	ISR	source Image to Spatial distortion Ratio	Tỷ lệ giữa tín hiệu và sai số do mô hình không gian
7	ISTFT	Inverse Short-Time Fourier Transform	Phép biến đổi Fourier ngược
8	IID (i.i.d)	Interchannel Intensity Difference	Chênh lệch cường độ tín hiệu giữa các kênh
9	ITD (i.t.d)	Interchannel Time Difference	Chênh lệch thời gian giữa các kênh
10	GSSM	Generic Source Spectral Model	Mô hình phổ nguồn tổng quát
11	KL	Kullback-Leibler	Độ đo Kullback-Leibler
12	MU	Multiplicative Update	Quy tắc cập nhật theo cấp số nhân
13	NMF	Non-negative Matrix Factorization	Thừa số hóa ma trận không âm
14	SAR	Signal to Artifacts Ratio	Tỷ lệ giữa tín hiệu và sai số do quá trình tính toán
15	SDR	Signal to Distortion Ratio	Tỷ lệ giữa tín hiệu và sai số tổng thể
16	SIR	Signal to Interference Ratio	Tỷ lệ giữa tín hiệu và nhiễu giao thoa từ các tín hiệu khác
17	SNR	Signal to Noise Ratio	Tỷ lệ giữa tín hiệu và nhiễu
18	STFT	Short-Time Fourier Transform	Phép biến đổi Fourier

THÔNG TIN KẾT QUẢ NGHIÊN CỨU

MỞ ĐẦU

1. Tổng quan về nâng cao chất lượng tiếng nói

Trong thực tế có rất nhiều tình huống mà tín hiệu tiếng nói thu âm được bị trộn lẫn với nhiều nhiễu từ môi trường xung quanh và tiếng vọng do phản xạ âm thanh mang lại (ví dụ: thu âm ở những nơi công cộng, nói chuyện điện thoại ở ngoài đường, hay giao tiếp trong buổi tiệc). Hình 1 mô tả một tình huống rất điển hình trong buổi tiệc, nơi có nhiều người tham dự, nhiều cuộc trò chuyện diễn ra đồng thời cùng với tiếng nhạc, tiếng chạm ly và nhiều âm thanh ồn ào khác. Trong điều kiện đó, con người với khả năng thính giác bình thường qua hai tai có thể dễ dàng định vị và phân tách tiếng nói mong muốn để nghe hiểu và xử lý thông tin. Tuy nhiên với học máy (machine learning) thì công việc này lại trở nên vô cùng khó khăn, đặc biệt là khi tín hiệu nghi âm có tiếng vọng và bị trộn bởi nhiều nguồn nhiễu khác nhau (nhiều nền, các giọng nói khác,...) [1].



Hình 1: Thu âm trong môi trường nhiễu

Chính vì thế phân tách và nâng cao chất lượng tiếng nói (speech separation and enhancement) là vấn đề quan trọng trong lĩnh vực xử lý tín hiệu và được tập trung

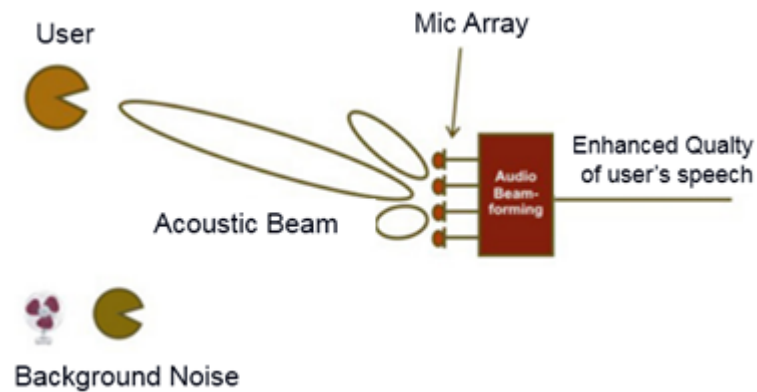
ngiên cứu nhiều trên thế giới. Chúng nói có vai trò quan trọng trong nhiều ứng dụng thực tế như:

- Nâng cao chất lượng giao tiếp qua mạng viễn thông (telephone, mobile phone) và hội nghị truyền hình (video/audio conferencing),
- Ứng dụng trong các hệ thống nhận dạng giọng nói tự động (Automatic Speech Recognition - ASR) như điều khiển robot, điều khiển các thiết bị bằng giọng nói,...
- Ứng dụng trong các thiết bị hỗ trợ người khiếm thính bị giảm khả năng nghe hiểu (hearing aids).
- Ứng dụng trong bước phân tách và xử lý âm thanh (information retrieval) của các hệ thống an ninh quốc phòng, chăm sóc y tế,...

1.1. Tình hình nghiên cứu trên thế giới

Có nhiều phương pháp nâng cao chất lượng tiếng nói đã và đang được nghiên cứu trên thế giới. Hầu hết các phương pháp sẽ biến đổi tín hiệu âm thanh từ miền thời gian (time domain) sang miền tần số (time-frequency domain), sau đó sử dụng các đặc trưng về phổ của tín hiệu (spectral cues) cho việc xử lý. Ngoài đặc trưng về phổ thì đặc trưng về hướng (spatial cues) của nguồn âm cũng đóng vai trò rất quan trọng trong các thuật toán sử dụng mảng microphone. Có thể tóm tắt về một số kỹ thuật nâng cao chất lượng tiếng nói được công bố gần đây như sau:

- **Kỹ thuật loại nhiễu (Noise Cancellation):** khi sử dụng 1 microphone để thu tín hiệu. Phương pháp này ước lượng mức tín hiệu nhiễu (noise spectral power) trong các khung thời gian (time frame) được xác định là không có tín hiệu tiếng nói, sau đó ở các khung thời gian có tín hiệu tiếng nói, phổ của âm thanh được các định khi trừ phổ của tín hiệu quan sát được cho phổ của tín hiệu nhiễu. Một số thuật toán cổ điển có thể kể đến là Spectral Subtraction, Wiener filtering,...Thuật toán tiêu biểu đang được sử dụng rộng rãi và được công nhận cho kết quả tốt trong nhiều điều kiện nhiễu khác nhau có thể kể đến là thuật toán OMLSP do I. Cohen đề xuất [9], thuật toán Wiener post-filtering loại bỏ nhiễu môi trường đồng đều (diffuse noise) do nhóm nghiên cứu của trường đại học Tokyo đề xuất năm 2010 [27].
- **Kỹ thuật tạo chùm tia Beamforming:** Khi sử dụng từ 2 microphone trở lên (microphone array) thì ngoài thông tin về đặc trưng phổ (spectral cues), thông tin về vị trí hay hướng của nguồn âm trong không gian (Direction of Arrival - DoA) có thể giúp ích rất lớn cho việc xử lý loại nhiễu và tăng cường tín hiệu



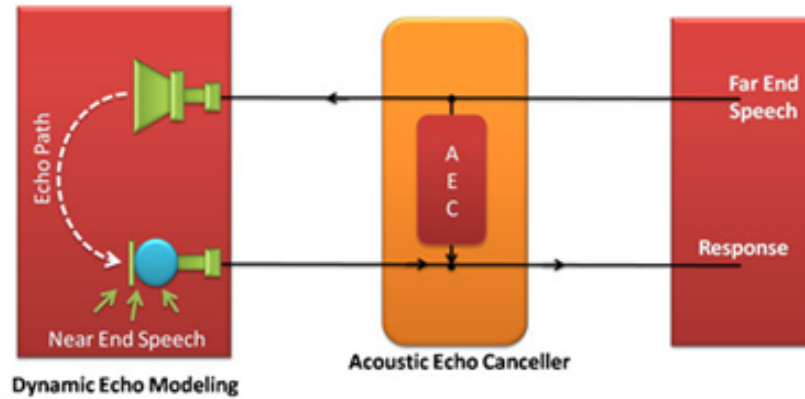
Hình 2: Kỹ thuật Beamforming [27]

mong muốn. Kỹ thuật tạo chùm tia Beamforming dựa vào hướng đến của tín hiệu nguồn âm (DoA) và khoảng cách của các microphone để thiết kế véc tơ định hướng (steering vector) giúp tăng cường hay triệt giảm tín hiệu theo hướng mong muốn. Các thuật toán Beamforming tiêu biểu có thể kể đến là MVDR hay McCowan [46]. Tuy nhiên trong môi trường thực tế với nhiều nguồn âm thì việc ước lượng chính xác hướng vẫn còn nhiều hạn chế, đặc biệt khi các nguồn âm di chuyển so với vị trí của microphone.

- Kỹ thuật tách nguồn âm thanh (Audio Source Separation):** Các thuật toán tách nguồn âm thanh thường kết hợp sử dụng cả thông tin về hướng của nguồn âm (spatial cues) và thông tin về đặc trưng phổ của nguồn (spectral cues) trong quá trình phân tách. Phương pháp được biết đến nhiều là ICA (Independent component analysis) áp dụng được khi số microphone lớn hơn số nguồn âm. Các phương pháp gần đây sử dụng các mô hình thống kê (statistical models) cho phép áp dụng hiệu quả và linh động hơn. Ví dụ, nghiên cứu của nhóm Telecoms ParisTech, Illinois, và nhóm của Adobe sử dụng mô hình Nonnegative Matrix Factorization (NMF) hay PLCA để xử lý đặc trưng phổ của tín hiệu âm thanh. Nhóm nghiên cứu của viện INRIA Pháp phát triển mô hình full-rank Gaussian [59] để xử lý đặc trưng về vị trí trong không gian của nguồn các âm. Nhóm nghiên cứu của NTT Nhật Bản sử dụng mặt nạ thời gian-tần số kết hợp phương pháp Wiener để tách nguồn [26],... Tách nguồn âm thanh đang thu hút rất nhiều nghiên cứu trong các năm gần đây, tuy nhiên các phương pháp công bố vẫn chưa đạt được hiệu quả tốt trong các tình huống thu âm thực tế như khi nguồn âm thanh không cố định vị trí, môi trường thu âm có nhiều tiếng vọng (reverberation), khi

số lượng nguồn cần tách nhiều hơn số lượng microphone, hay khi không có dữ liệu huấn luyện cho bài toán.

- **Kỹ thuật triệt tiêu tiếng vọng (Acoustic Echo Cancellation - AEC):**



Hình 3: Kỹ thuật triệt tiêu tiếng vọng.

AEC thường được sử dụng trong hệ thống giao nghị truyền thanh/truyền hình khi microphone thu lẫn cả tiếng người nói trực tiếp (gọi là near-end signal) và tiếng của người hội thoại phát ra từ loa (gọi là far-end signal). Trong trường hợp này ngoài việc loại bỏ nhiễu từ môi trường thông thường thì cần phải triệt tiêu tín hiệu far-end thu được ở microphone. Thuật toán cơ bản trong AEC là thiết kế bộ lọc tuyến tính mô phỏng đường truyền âm từ loa tới microphone, qua đó cho phép tính toán ước lượng tín hiệu far-end tại microphone. Tín hiệu mong muốn được tính bằng tín hiệu thực tế thu được trừ đi tín hiệu far-end ước lượng tại microphone, và tiếp tục trừ đi tín hiệu nhiễu ở môi trường xung quanh. AEC có thể sử dụng kết hợp với các kỹ thuật tăng cường chất lượng tiếng nói khác để tạo hiệu quả tổng thể cao. Ví dụ nhóm nghiên cứu của F. Tristan đề xuất kết hợp AEC với beamforming [34], hay nhóm nghiên cứu ở Israel sử dụng AEC kết hợp với kỹ thuật loại nhiễu (Noise Cancellation) [47].

- **Kỹ thuật giảm độ vọng (Speech Dereverberation):** thường được sử dụng với tín hiệu thu âm trong điều kiện có nhiều phản xạ từ tường, sàn, trần nhà. Kỹ thuật này có thể được dùng như bước tiền xử lý để giúp các thuật toán loại nhiễu, ước lượng hướng tín hiệu, tách âm hay AEC cho kết quả tốt hơn. Gần đây, nhóm nghiên cứu của tập đoàn Hitachi Nhật Bản đã đề xuất kết hợp đồng thời kỹ thuật Speech Dereverberation và triệt tiêu tiếng vọng (AEC) trên cùng một thuật toán để nâng cao hiệu quả tăng cường chất lượng tiếng nói [29].

Có thể thấy trong số các phương pháp nâng cao chất lượng tiếng nói nêu trên, tách nguồn âm thanh là kỹ thuật có thể áp dụng trong nhiều điều kiện khác nhau, do có thể kết hợp đồng thời hoặc sử dụng riêng rẽ các phương pháp xử lý thông tin phổ và thông tin không gian của tín hiệu [43, 58, 76]. Chính vì thế, đây là phương pháp được nhóm nghiên cứu lựa chọn giải quyết vấn đề đặt ra của đề tài.

1.2. Tình hình nghiên cứu trong nước

Cùng với sự phát triển của công nghệ thông tin và truyền thông, nhận thức được tầm quan trọng của việc nghiên cứu và ứng dụng công nghệ xử lý tiếng nói trong đời sống, từ khoảng hơn chục năm về trước, nhiều tổ chức, viện nghiên cứu, trường đại học trong nước đã tiến hành đầu tư nghiên cứu cho lĩnh vực này. Tuy nhiên các đơn vị này chủ yếu tập trung vào các bài toán: nhận dạng tiếng nói, tổng hợp tiếng nói, phân tích ngữ âm học [70, 71]. Bài toán về phân tách âm thanh cũng đã được nghiên cứu trong nước, chủ yếu thông qua các nhóm nghiên cứu tại các trường đại học. Đây là bài toán quan trọng thường được thực hiện nhằm nâng cao chất lượng tín hiệu âm thanh/tiếng nói thu được bởi máy tính, trước khi tiếp tục sử dụng các kỹ thuật xử lý tiếng nói khác như nhận dạng tiếng nói, nhận dạng âm thanh, v.v. Có thể kể tới các nhóm nghiên cứu sau:

Nhóm nghiên cứu của tác giả Trương Tấn Quang, Trường ĐH Khoa học tự nhiên, ĐHQG-HCM thực hiện các nghiên cứu tách nguồn âm áp dụng cho âm thanh, sử dụng phương pháp thống kê dựa trên phân tích thành phần độc lập ICA (Independent Component Analysis) gồm hai bước: ước lượng ma trận lai trộn và phân tách nguồn [62]. Nhóm đã có thử nghiệm đánh giá việc tách âm thanh trong trường hợp số tín hiệu nguồn cần ước lượng lớn hơn số cảm biến microphone (còn gọi là trường hợp under-determined). Thử nghiệm được thực hiện gồm phân tách 3 nguồn âm là 3 giọng nói độc lập hoặc 2 giọng nói và 1 tín hiệu âm nhạc. Kết quả phân tách là tương đối tốt tuy nhiên phương pháp ICA vẫn còn có tồn tại đó là yêu cầu nhiều điều kiện ràng buộc như các thành phần độc lập được xem là độc lập thống kê và có phân bố phi Gauss, ma trận lai trộn là ma trận vuông v.v.

Một công trình nghiên cứu khác nữa có thể kể đến là của tác giả Vương Hoàng Nam, trường ĐH Bách Khoa Hà Nội [81]. Trong công trình này, mô hình tách nguồn âm được nghiên cứu với giả thiết số nguồn tín hiệu bằng số sensor thu nhận là microphones. Nhìn chung công trình vẫn chủ yếu dựa trên việc sử dụng phương pháp ICA và thuật toán Fast-ICA, Fast-ICA phức tạp và đề xuất thêm một thuật toán DOAP sử dụng kết hợp góc tới của nguồn âm và bài toán phân việc để hỗ trợ việc phân tách. Các thuật toán

được chỉ ra có thể áp dụng cho phân tách âm thanh và ứng dụng được trong nhiều lĩnh vực khác nữa. Công trình nghiên cứu đã được mở rộng phát triển thành đề tài NCKH cấp Bộ về ứng dụng giải thuật mù vào lĩnh vực ảnh viễn thám, hoặc xử lý phân tích các tín hiệu y sinh học.

Một nhóm nghiên cứu cũng theo đuổi lĩnh vực bài toán này là nhóm nghiên cứu của tác giả Nguyễn Linh Trung. Trong đó bài toán xử lý phân tách tín hiệu dạng sóng âm từ các nguồn không ổn định, và cũng áp dụng cho trường hợp có nhiều nguồn âm hơn nguồn thu nhận. Giải thuật được dùng dựa trên phân bố thời gian tần số trên miền không gian (spatial time-frequency distributions) kết hợp với các phép chiếu chia nhỏ không gian (subspace projection) [40]. Nghiên cứu này sau đó được mở rộng sang phân tách nguồn tín hiệu điện não đồ EEG trong việc phát hiện các cơn động kinh ở người bệnh [69].

Có thể thấy bài toán phân tách nguồn âm thanh đã được nghiên cứu trong nước, tuy nhiên các nhóm nghiên cứu vẫn dừng lại ở việc thử nghiệm nhỏ lẻ các phương pháp cổ điển như ICA, TFD, v.v. mà chưa có những nghiên cứu ứng dụng các kỹ thuật khai phá dữ liệu tiên tiến hiện nay như phân tách ma trận không âm (NMF) hay mạng rơron (DNN). Hơn nữa, các nghiên cứu trong nước mới chỉ dừng lại ở mức thử nghiệm trên dữ liệu mô phỏng trong điều kiện ít nhiễu và tiếng vọng mà chưa tính đến các thách thức trong môi trường thu âm thực tế.

1.3. Tính cấp thiết của đề tài

Những phân tích trên cho thấy phân tách và nâng cao chất lượng tiếng nói là vấn đề có vai trò hết sức quan trọng và được ứng dụng trực tiếp trong nhiều hệ thống phục vụ đời sống của con người. Các đánh giá khoa học uy tín những năm gần đây [4, 43, 55, 75] cũng cho thấy kết quả phân tách tiếng nói trong các trường hợp khó khăn như: tín hiệu có tỷ lệ nhiễu cao ($SNR \leq 0dB$) và không có dữ liệu huấn luyện vẫn còn rất thấp ($SDR < 6dB$). Bên cạnh đó, các nghiên cứu trong nước về vấn đề này mới chỉ tiến hành với những kỹ thuật cổ điển và thí nghiệm trên những bộ dữ liệu mô phỏng chứ chưa thử nghiệm với dữ liệu thu âm trong môi trường thực có tỷ lệ nhiễu cao.

Chính vì thế việc nghiên cứu, đề xuất một thuật toán phân tách tín hiệu tiếng nói đạt hiệu quả tốt hơn những thuật toán đã công bố trên thế giới trong điều kiện tín hiệu thu âm thực chứa nhiễu ở mức cao là rất cần thiết và có tính ứng dụng thực tiễn cao.

2. Mục tiêu và cách tiếp cận của đề tài

- **Mục tiêu nghiên cứu:**

Mục tiêu của đề tài là nghiên cứu phát triển thuật toán tách tiếng nói từ tín hiệu thu âm có nhiễu ở mức cao. Thuật toán phải có đóng góp mới về học thuật và đạt hiệu quả phân tách cao hơn những thuật toán đã công bố trên thế giới trong cùng điều kiện thí nghiệm.

- **Cách tiếp cận của đề tài:**

Đề tài tiếp cận theo góc độ: xuất phát từ nghiên cứu lý thuyết, từ đó lựa chọn kỹ thuật để cải tiến và phát triển thuật toán, tiến hành thực nghiệm để đưa ra những kết luận và định hướng ứng dụng.

Theo cách tiếp cận đó, nhóm nghiên cứu đã tìm hiểu các kỹ thuật phân tách âm thanh khác nhau và nhận thấy:

- Các kỹ thuật tách nguồn mù cho kết quả còn rất hạn chế.
- Các phương pháp tách nguồn có dữ liệu huấn luyện, hay có những thông tin hướng dẫn cụ thể về nguồn âm như tách tiếng nói khi đã biết nội văn bản (speech transcript),... cho kết quả khá tốt. Tuy nhiên trong rất nhiều trường hợp, dữ liệu huấn luyện hay những thông tin cụ thể đó là không có sẵn.

Chính vì thế, nhóm nghiên cứu đã tiếp cận theo hướng "*weakly-informed*" để xây dựng thuật toán tách nguồn âm thanh. Thuật toán sử dụng thông tin hướng dẫn rất chung chung về loại của nguồn âm (như tiếng nói, tiếng nhạc hay nhiễu môi trường,...) để hướng dẫn quá trình phân tách.

3. Phương pháp, đối tượng và phạm vi nghiên cứu

- **Phương pháp nghiên cứu:**

- Nghiên cứu lý thuyết: Nghiên cứu và phân tích các phương pháp, các công trình nghiên cứu liên quan để đề xuất thuật toán phù hợp với ứng dụng của đề tài.
- Thực nghiệm: Thuật toán đề xuất được cài đặt và chạy thử nghiệm trên bộ dữ liệu thực để đánh giá kết quả và tối ưu hóa các tham số.

- **Đối tượng nghiên cứu:**

- Nghiên cứu kỹ thuật tách nguồn âm thanh sử dụng mô hình Nonnegative Matrix Factorization (NMF).

- Nghiên cứu các đặc trưng phổ của tín hiệu tiếng nói và âm thanh nhiễu trong môi trường thực.

- **Phạm vi nghiên cứu:**

- Tập trung nghiên cứu phương pháp tách tín hiệu tiếng nói (speech) trong điều kiện không có đầy đủ dữ liệu huấn luyện cho cả hai trường hợp thu âm đơn kênh và đa kênh.

- Nghiên cứu các đặc tính của tiếng nói, các loại âm thanh nhiễu môi trường và sự phản xạ âm thanh.

4. Nội dung nghiên cứu

- **Nội dung 1:** Nghiên cứu tổng quan: Nghiên cứu các phương pháp phân tách, nâng cao chất lượng tiếng nói, phân tích ưu nhược điểm của các kết quả nghiên cứu đã công bố liên quan đến đề tài.

- **Nội dung 2:** Đề xuất thuật toán mới phân tách tiếng nói từ tín hiệu chứa nhiễu trong đó phân tích rõ những đóng góp về học thuật của thuật toán đề xuất. Biểu diễn thuật toán bằng sơ đồ khối, thể hiện rõ các điều kiện ràng buộc về dữ liệu và các bước xử lý phân tách.

- **Nội dung 3:** Xây dựng bộ dữ liệu để thử nghiệm và đánh giá kết quả của thuật toán đề xuất. Tìm hiểu bộ công cụ (tools) đánh giá chất lượng của tín hiệu tiếng nói sau khi phân tách (bộ công cụ hiện đang được cộng đồng nghiên cứu trên thế giới về lĩnh vực công nhận và sử dụng rộng rãi).

- **Nội dung 4:** Cài đặt thuật toán phân tách tiếng nói mong muốn từ tín hiệu thu âm gồm nhiều âm thanh khác nhau bị trộn lẫn. Chạy thử nghiệm, tối ưu hóa các tham số, so sánh kết quả của thuật toán đề xuất với các phương pháp khác được công bố gần đây trên thế giới, đánh giá và công bố kết quả.

CHƯƠNG 1

TỔNG QUAN VỀ TÁCH NGUỒN ÂM THANH

1.1 Mô hình tổng quát của bài toán tách nguồn âm thanh

Với mục đích nâng cao chất lượng nguồn tiếng nói mong muốn trong tín hiệu thu âm có chứa nhiều nền, tiếng vọng và các âm thanh khác, kỹ thuật tách nguồn âm thanh (audio source separation) xem xét tiếng nói mong muốn là một trong số những nguồn âm cần tách từ tín hiệu trộn (mixture). Tùy theo từng tình huống cụ thể mà nhiều và các tín hiệu âm thanh khác (gọi chung là nhiễu) được xem xét như là một hay nhiều nguồn âm cần tách.

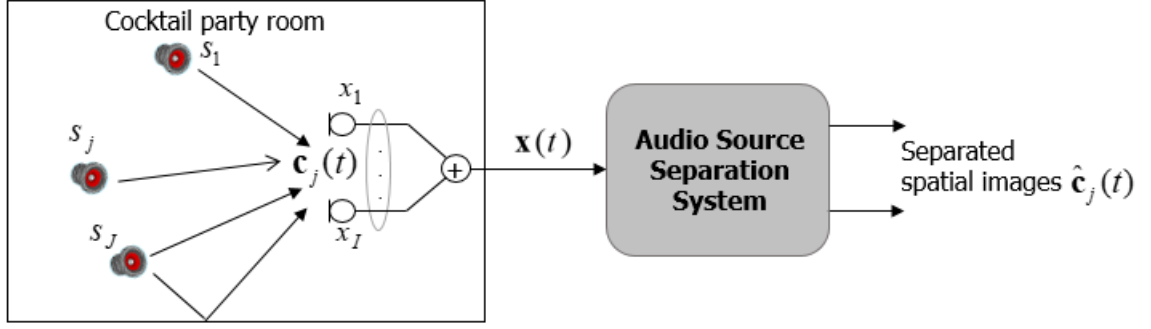
Tách nguồn âm thanh là kỹ thuật phân tách những âm thanh mong muốn từ tín hiệu thu âm bị trộn lẫn bởi nhiều âm thanh khác nhau [19, 42, 46, 49, 58]. Bài toán tách nguồn âm thanh tổng quát được mô tả như sau:

Mục đích của tách nguồn âm thanh là 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).

Giả sử tín hiệu trộn từ J nguồn âm được thu âm bởi I microphone, với $j \in \{1, 2, \dots, J\}$ là chỉ số của nguồn âm và $i \in \{1, 2, \dots, I\}$ là chỉ số của microphone. Tín hiệu trộn $\mathbf{x}(t) = [x_1(t), \dots, x_I(t)]^T \in \mathbb{R}^{I \times 1}$ được biểu diễn theo công thức sau [46]:

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

với $\mathbf{c}_j(t) = [c_{1j}(t), \dots, c_{Ij}(t)]^T \in \mathbb{R}^{I \times 1}$ 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



Hình 1.1: Hệ thống tách nguồn âm thanh.

tơ hoặc ma trận, $t \in \{0, 1, \dots, T-1\}$ 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:

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

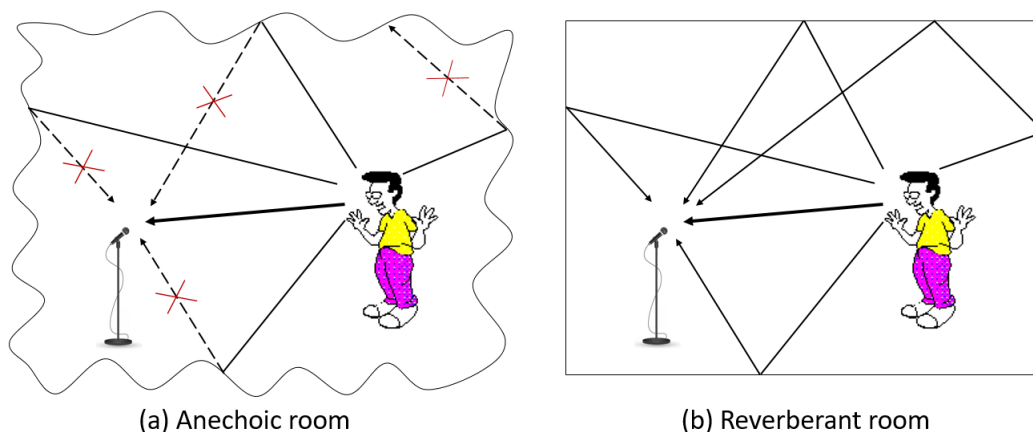
với $\mathbf{c}_j(n, f) \in \mathbb{C}^{I \times 1}$ và $\mathbf{x}(n, f) \in \mathbb{C}^{I \times 1}$ là biểu diễn trong miền T-F tương ứng của $\mathbf{c}_j(t)$ và $\mathbf{x}(t)$. $n = 1, 2, \dots, N$ là chỉ số khung thời gian và $f = 1, 2, \dots, F$ biểu diễn số bin tần số. Mục tiêu của hệ thống 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 hiệu trộn I kênh $\mathbf{x}(t)$.

1.2 Khái quát về các kỹ thuật tách nguồn âm thanh

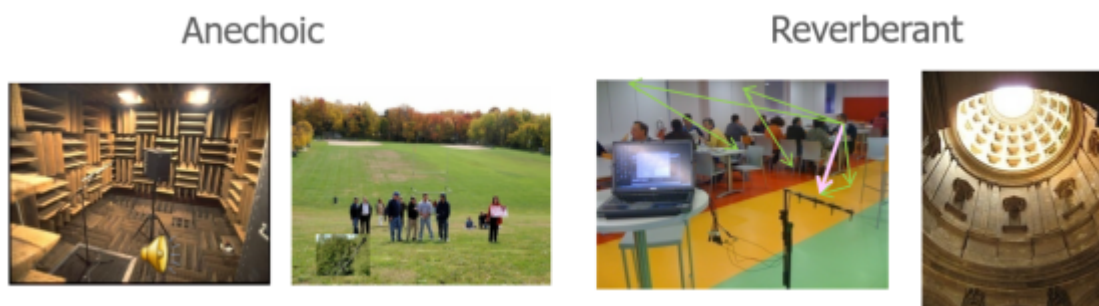
Bắt đầu bằng phương pháp tách nguồn mù (blind -source separation) vào giữa những năm 90 [7], các nghiên cứu về tách nguồn âm thanh đã liên tục phát triển cho đến nay sử dụng nhiều phương pháp khác nhau, với nhiều kịch bản ứng dụng thực tế khác nhau như: ứng dụng trong nhận dạng giọng nói tự động, chỉnh sửa âm nhạc, truy vấn và phát hiện thông tin âm thanh trong các hệ thống chăm sóc y tế, hệ thống an ninh, ... Các phương pháp tách nguồn âm thanh đã được nghiên cứu ở nhiều điều kiện ngữ cảnh và mức độ phức tạp khác nhau và được phân loại theo nhiều tiêu chí khác nhau dựa trên điều kiện thu âm và các thông tin có thể khai thác trong quá trình phân tách.

1.2.1 Một số khái niệm cơ bản trong tách nguồn âm thanh

* Môi trường thu âm hay cách tạo ra tín hiệu trộn là vấn đề ảnh hưởng trực tiếp đến việc lựa chọn phương pháp tách nguồn âm, môi trường thu âm được chia thành ba loại như sau:



Hình 1.2: Các loại môi trường thu âm thực.

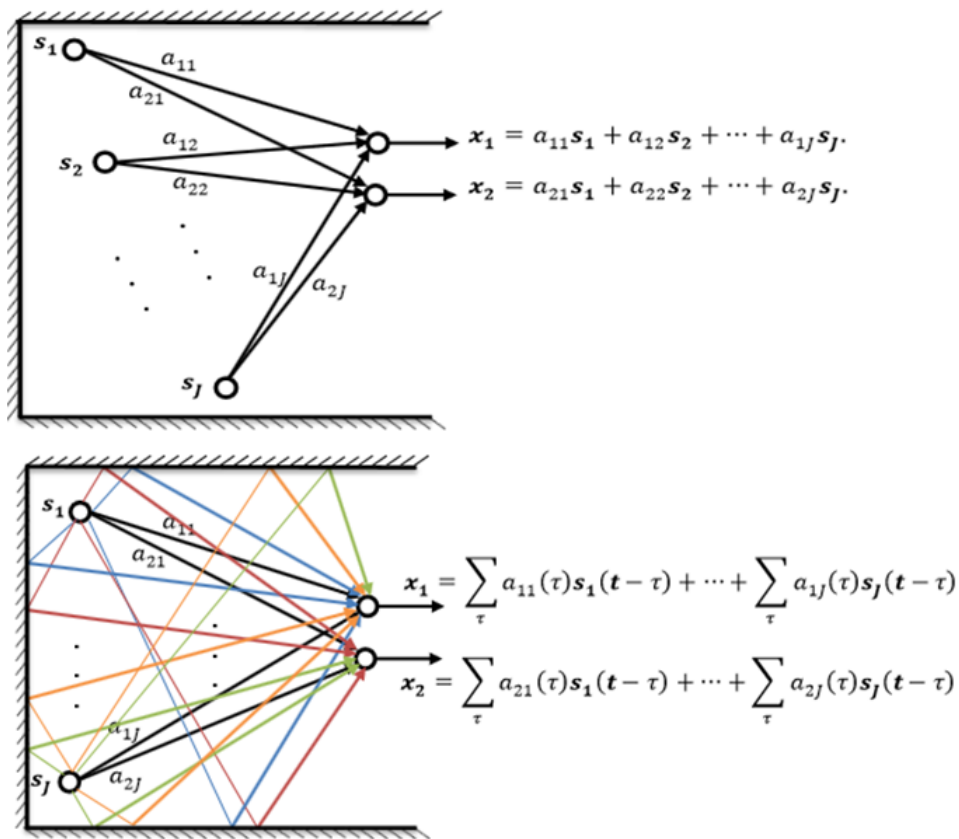


Hình 1.3: Hình ảnh minh họa môi trường thu âm thực.

- **Instantaneous:** Các tín hiệu thành phần được trộn trực tiếp với nhau bằng cách cộng biên độ tín hiệu theo bộ trọng số tỷ lệ nào đó. Tín hiệu tạo ra bằng cách này sẽ không có độ trễ về thời gian và đây là trường hợp đơn giản nhất cho các thuật toán phân tách tín hiệu.

- **Anechoic** là những môi trường không có phản xạ âm thanh, ví dụ như khi thu âm trong phòng thu hay ngoài môi trường không gian rộng không có vật chắn âm. Trong điều kiện đó, tín hiệu thu được có độ trễ về thời gian nhưng hầu như không có tiếng vọng và việc tách nguồn âm sẽ trở nên dễ dàng hơn.

- **Reverberant** là môi trường thu âm thực có phản xạ âm, như khi thu âm trong phòng, nhà thờ, hay các địa điểm công cộng. Khi đó âm thanh thu được sẽ vừa có độ trễ thời gian, vừa bao gồm phản xạ âm (tiếng vọng) và đòi hỏi các thuật toán tách nguồn phải xử lý loại bỏ âm thanh phản xạ đó.



Hình 1.4: Tín hiệu thu âm trong môi trường có và không có phản xạ âm.

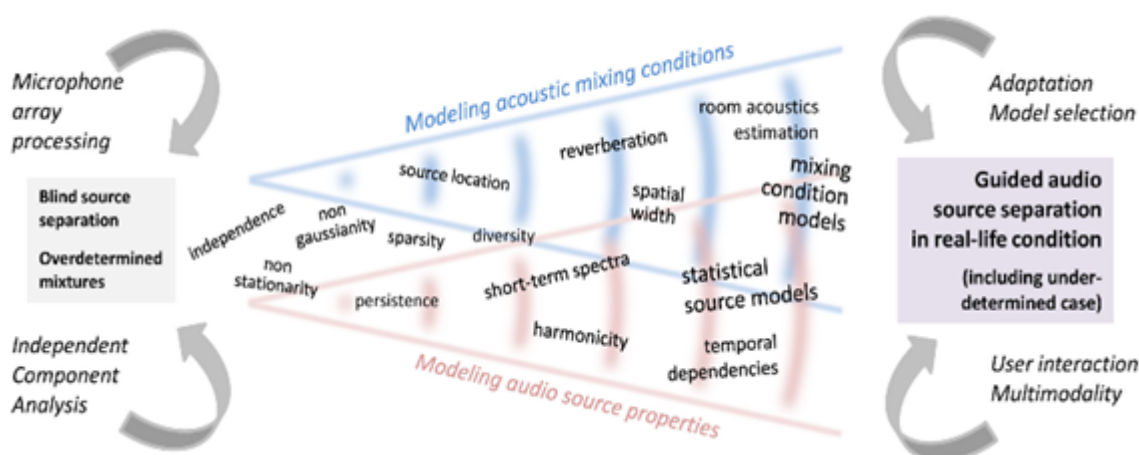
* Phân loại theo tương quan giữa số nguồn âm và số microphone sẽ có ba trường hợp tương ứng sau:

- **Overdetermined** là trường hợp số lượng microphone lớn hơn số nguồn âm. Các thuật toán hiện nay đã có thể cho kết quả phân tách khá tốt trong trường hợp này.
- **Determined** là trường hợp số lượng microphone bằng số nguồn âm.
- **Underdetermined** là trường hợp số lượng microphone nhỏ hơn số nguồn âm. Đây là trường hợp còn rất nhiều khó khăn thách thức cho cộng đồng tách nguồn âm thanh.

Ngoài ra, khi chỉ dùng một microphone để thu âm ta sẽ có tín hiệu đơn kênh (single-channel). Tín hiệu đa kênh (multichannel) có được khi dùng từ hai microphone trở lên để thu âm.

1.2.2 Các kỹ thuật tách nguồn âm thanh

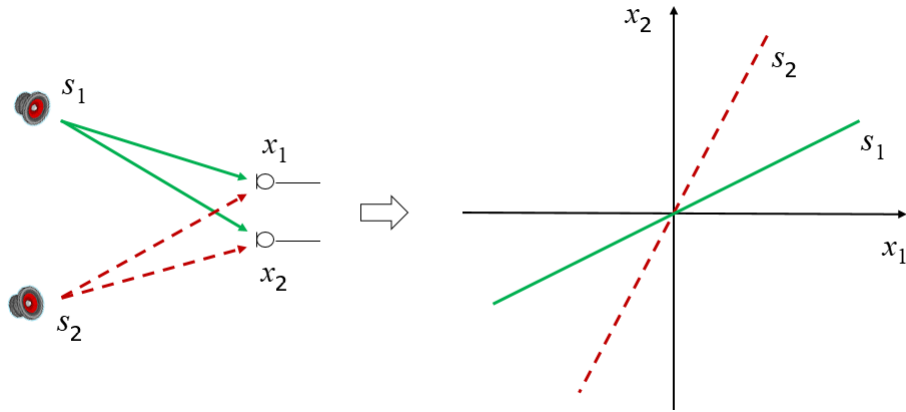
Hình 1.2 là bức tranh minh họa sự phát triển của các phương pháp tách nguồn âm thanh trên thế giới trong hơn hai thập kỷ qua [76]. Các phương pháp phân tách được thực hiện dựa trên việc mô hình hóa và khai thác các đặc tính không gian của nguồn âm và môi trường thu âm (spatial cues) (là mối quan hệ giữa các kênh theo thời gian) và/hoặc các đặc tính phổ của nguồn âm (spectral cues), tức là sự phân phối năng lượng phổ của các nguồn theo thời gian và tần số.



Hình 1.5: Mô hình tổng quan về lịch sử phát triển các kỹ thuật tách nguồn âm thanh.

Các phương pháp tách nguồn khai thác đặc tính không gian sử dụng mảng microphone với giả định các nguồn có hướng đến khác nhau so với các microphone. Các nhóm nghiên cứu theo hướng khai thác đặc tính không gian của nguồn âm và môi trường thu âm này dựa trên các kỹ thuật như: định vị nguồn âm (source localization) [22]; phân tích sự chênh lệch về thời gian đến (Interchannel Time Difference - ITD) và chênh lệch về cường độ tín hiệu (Interchannel Intensity Difference IID) đến các microphone [6, 57] (ví dụ trong hình 1.6) Một số nghiên cứu khác mã hóa và ước lượng mô hình không gian của môi trường thu âm bằng một ma trận bậc 1 (rank-1) trong miền tần số với giả định băng hẹp [65, 84]. Nghiên cứu của Nikunen và các cộng sự công bố năm 2014 [51] sử dụng ma trận hiệp phương sai bậc đầy đủ full-rank Gaussian để biểu diễn môi trường thu âm có phản xạ âm.

Các nghiên cứu cho thấy trong số các kỹ thuật nêu trên, sử dụng ITD/IID hay mô hình rank-1 là giải pháp phù hợp khi môi trường thu âm không có phản xạ âm (trong phòng thu), nhưng với môi trường thu âm thực có độ phản xạ âm cao (trong phòng họp,



Hình 1.6: Chênh lệch cường độ của 2 tín hiệu tại 2 microphone trong môi trường không phản xạ.

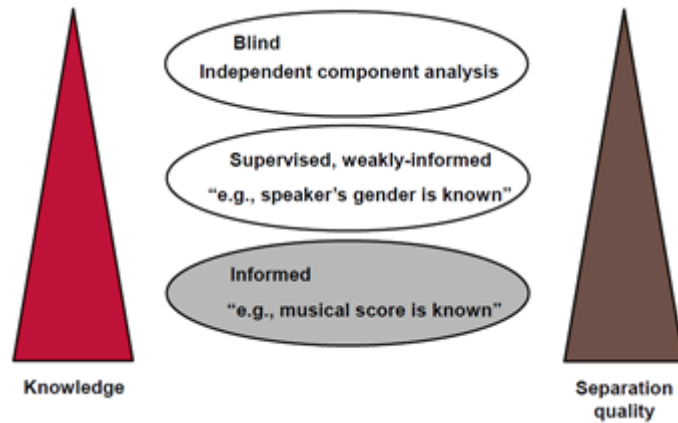
nhà thờ, nơi công cộng,...) thì các phương pháp này cho kết quả rất hạn chế. Mô hình full-rank Gaussian có độ phức tạp tính toán cao hơn nhưng là mô hình khái quát có thể biểu diễn được môi trường thu âm thực có phản xạ âm.

Nhóm phương pháp phân tách dựa trên thông tin phổ khai thác tính chất đa dạng của phổ tín hiệu với những giả định ràng buộc khác nhau. Phương pháp phân tích các thành phần độc lập (independent component analysis - ICA) [50, 82], hay phân cụm trong miền tần số (T-F components clustering) [41] giả định các tín hiệu nguồn là các thành phần độc lập với nhau. Với giả định năng lượng phổ của các nguồn âm có sự phân bố xác suất đặc biệt, một số nghiên cứu sử dụng mô hình Gaussian để mã hóa và ước lượng phân bố năng lượng phổ của các nguồn âm [7].

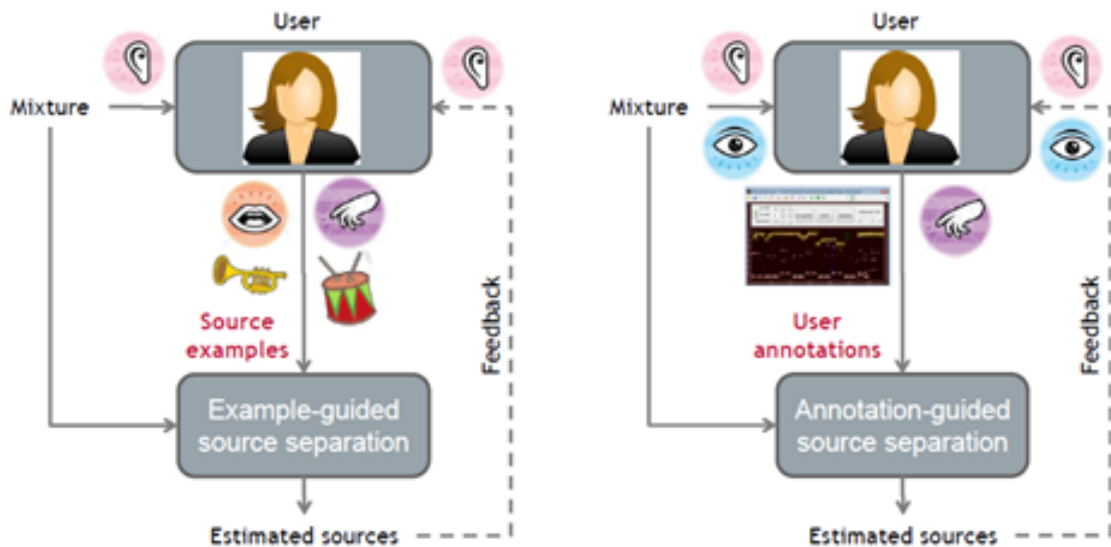
Các phương pháp dựa trên đặc tính không gian cho kết quả tương đối tốt trong trường hợp số nguồn âm ít hơn hoặc bằng số microphone (over-determined/determined) nhưng cho kết quả rất kém trong trường hợp số nguồn âm nhiều hơn số microphone (under-determined). Chúng cũng rất khó phân biệt các nguồn âm khi chúng có vị trí gần nhau và có cùng hướng đến microphones. Phương pháp tách nguồn dựa trên thông tin về phổ cho phép phân tách trong trường hợp under-determined và khi các nguồn có cùng hướng đến. Tuy nhiên đối với những nguồn âm có đặc trưng phổ tương đối giống nhau thì kết quả đạt được sẽ không tốt.

Trong những năm gần đây, nhiều mô hình và thuật toán phức tạp hơn đã được nghiên cứu phát triển theo hướng sử dụng tích hợp mô hình không gian và mô hình phổ [23, 52, 86], sử dụng dữ liệu huấn luyện lớn [52], hay sử dụng các thông tin hướng dẫn cụ thể về nguồn âm (tách âm nhạc khi đã biết bản nhạc, tách tiếng nói khi

biết bản transcript,...) và điều kiện của môi trường thu âm để nâng cao hiệu quả tách nguồn [35, 57]. Các nghiên cứu sử dụng kỹ thuật học sâu DNNs [8, 52, 83], hay sử dụng phương pháp thừa số hóa ma trận không âm (Nonnegative Matrix Factorization - NMF) với các ràng buộc cải tiến [32, 35] đã cho kết quả tốt hơn các phương pháp cổ điển.



Hình 1.7: Từ tách nguồn mù đến tách nguồn có hướng dẫn [76].



Hình 1.8: Từ tách nguồn mù đến tách nguồn có hướng dẫn [5].

Tuy nhiên việc sử dụng DNNs đòi hỏi một lượng dữ liệu huấn luyện lớn cũng như chi phí tính toán lớn. Trong khi đó, sử dụng NMF có thể giải quyết được bài toán phân tách trong trường hợp thiếu dữ liệu huấn luyện. NMF cũng thể hiện được ưu thế so với phương pháp học sâu về tốc độ tính toán và sự tường minh trong việc mô tả quá trình

phân tách tín hiệu. Ngoài ra, các đánh giá uy tín công bố gần đây [19, 58] cũng cho thấy, kết quả phân tách nguồn âm của những phương pháp hiện nay vẫn còn rất thấp trong những trường hợp khó khăn thách thức như: không có dữ liệu huấn luyện hoặc dữ liệu huấn luyện rất hạn chế, khi các nguồn âm di chuyển, khi thu âm đơn kênh (không khai thác được thông tin về đặc tính không gian của môi trường thu âm).

Chính vì thế, nhóm nghiên cứu lựa chọn mô hình NMF để xây dựng thuật toán phân tách tiếng nói từ tín hiệu thu âm trong môi trường thực có mức nhiễu cao. Thuật toán xây dựng có thể thực hiện tách nguồn âm ngay cả khi không có sẵn dữ liệu huấn luyện cho tín hiệu cần phân tách.

CHƯƠNG 2

PHÂN TÁCH TIẾNG NÓI SỬ DỤNG MÔ HÌNH NONNEGATIVE MATRIX FACTORIZATION

2.1 Đặt vấn đề

Trong lĩnh vực nghiên cứu về xử lý âm thanh hiện nay, vấn đề phân tách nhằm nâng cao chất lượng tiếng nói mong muốn (speech enhancement) trong điều kiện thu âm trong môi trường thực tế có nhiều tiếng ồn và nhiễu vẫn còn rất nhiều khó khăn thách thức. Đặc biệt là khi tín hiệu trộn chứa nhiễu (noise) ở mức cao, ví dụ khi tỷ lệ nhiễu cao hơn hoặc bằng tín hiệu tiếng nói mong muốn ($SNR \leq 0dB$). Bên cạnh đó kết quả phân tách cũng còn rất thấp trong trường hợp thu đơn kênh (single-channel) vì khi đó không khai thác được thông tin không gian trong quá trình phân tách [30, 74].

Đối với vấn đề nâng cao chất lượng tiếng nói nói riêng, trong môi trường thu âm thực tế, tín hiệu tiếng nói (speech) thường bị trộn lẫn với nhiễu từ môi trường xung quanh và nhiều âm thanh không mong muốn khác như tiếng nhạc, tiếng xe cộ trên đường,... Nhiệm vụ của speech enhancement là cần loại bỏ các tín hiệu nhiễu và những âm thanh không mong muốn, làm cho tín hiệu tiếng nói trở nên tốt hơn [4, 74]. Speech enhancement có vai trò rất quan trọng và được ứng dụng trực tiếp trong các ứng dụng truyền thông (communication), xử lý âm thanh hậu kỳ trong truyền hình, phim ảnh (sound post-production), và trong điều khiển rô bốt (robotics). Để nâng cao hiệu quả của hệ thống nhận dạng tiếng nói tự động (automatic speech recognition -ASR) trong môi trường có nhiễu, thuật toán speech enhancement thường được kết hợp sử dụng để loại bỏ nhiễu trước khi nhận dạng [3, 30].

Tiếp cận speech enhancement theo hướng source separation, tiếng nói (speech) và nhiễu (noise) được coi là hai nguồn khác nhau cần được tách từ tín hiệu thu âm [46]. Trong phần này, nhóm nghiên cứu tập trung giải quyết bài toán phân tách tiếng nói trong trường hợp thu âm đơn kênh và không có dữ liệu huấn luyện chính xác cho các nguồn cần tách. Liên quan đến bài toán đặt ra, một số nghiên cứu mới đây đã đề xuất huấn luyện và đưa ra mô hình phổ tổng quát cho tín hiệu tiếng nói (universal speech

model) đối với trường hợp không chỉ có dữ liệu huấn luyện cho nhiều mà không có dữ liệu huấn luyện chính xác cho tiếng nói [68], hoặc xây dựng mô hình phổ tổng quát cho cả tín hiệu tiếng nói và nhiễu trong trường hợp không có dữ liệu huấn luyện chính xác cho cả hai từ các mẫu tín hiệu tiếng nói và nhiễu khác nhau thu thập được (tìm kiếm dữ liệu huấn luyện tương tự từ Internet cho các loại tín hiệu khác nhau) [?]. Tuy nhiên những công bố kể trên chỉ áp dụng mô hình Nonnegative Matrix Factorization (NMF) cơ bản, hoặc sử dụng NMF với một điều kiện ràng buộc thưa đơn lẻ. Trong nghiên cứu của mình, chúng tôi giới thiệu về mô hình áp dụng NMF để nâng cao chất lượng tiếng nói, đề xuất phương pháp kết hợp mô hình NMF với hai nhóm ràng buộc thưa (sparsity constraint) để nâng cao hiệu quả tách tín hiệu tiếng nói mong muốn trong trường hợp không có dữ liệu huấn luyện chính xác cho các tín hiệu cần tách (unsupervised).

2.2 Tổng quan về NMF

NMF là kỹ thuật giảm số chiều của ma trận thường được dùng trong phân tích dữ liệu không âm có kích thước lớn. NMF được sử dụng phổ biến từ sau công bố cả tiến công thức cập nhật tham số của Lee và Seung năm 1999 [36, 37]. Nhưng thuật toán gốc với tên gọi *nonnegative rank factorization* [28] và *positive matrix factorization* [61] đã xuất hiện từ trước đó 20 năm. Cho đến nay, NMF được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau như tin sinh học (bioinformatics) [39], xử lý ảnh [64], nhận dạng ảnh [24], xử lý âm thanh và tiếng nói [14, 48],...

Có nhiều tài liệu mô tả rất chi tiết về mô hình NMF và các vấn đề liên quan, có thể kể đến là [85, 87]. Trong phần này, nhóm nghiên cứu sẽ trình bày một số nội dung cơ bản về NMF: kỹ thuật đó là gì và được áp dụng như thế nào trong xử lý và phân tách tín hiệu âm thanh.

2.2.1 NMF là gì?

Giả sử cho ma trận đầu vào $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ có kích thước $F \times N$ chứa các phần tử không âm, NMF sẽ phân tách ma trận $\mathbf{V}$ thành hai ma trận thành phần $\mathbf{W}$ và $\mathbf{H}$ sao cho $\mathbf{WH}$ xấp xỉ bằng $\mathbf{V}$ theo công thức ước lượng sau [37]:

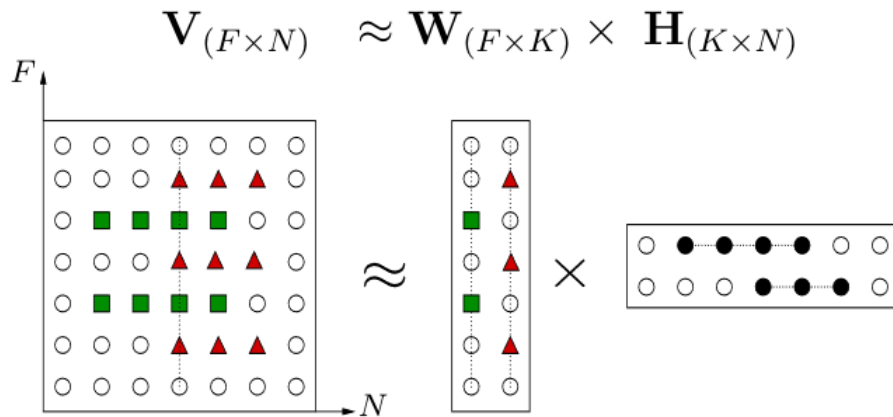
$$\mathbf{V} \approx \mathbf{WH}, \quad (2.1)$$

trong đó $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ và $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ cũng là các ma trận không âm có kích thước tương ứng là $F \times K$ và $K \times N$.

Các nghiên cứu cho thấy một lượng dữ liệu dù rất lớn nhưng nếu có cùng nguồn gốc thì dữ liệu đó sẽ chỉ gồm một số lượng nhỏ các đặc tính cơ sở, còn gọi là thành phần gốc hay véc tơ cơ sở. Ví dụ giọng nói của mỗi người, âm thanh phát ra của mỗi nhạc cụ sẽ có các đặc trưng riêng, một bức ảnh có thể được tái hiện bằng một tập véc tơ điểm xác định, ... Từ nhận định đó, có thể dùng NMF trong các bài toán phân tích dữ liệu đa biến theo cách sau: Từ một tập các véc tơ dữ liệu đa biến n chiều, các véc tơ là các cột của ma trận $\mathbf{V}$ kích thước $F \times N$, với F là các thuộc tính đặc trưng của dữ liệu, N là kích thước dữ liệu đầu vào. NMF sẽ phân tách $\mathbf{V}$ thành hai ma trận $\mathbf{W}$ kích thước $F \times K$ và ma trận $\mathbf{H}$ kích thước $K \times N$ như mô tả trong hình 2.1, trong đó K là số véc tơ cơ sở. Thông thường K được chọn sao cho nhỏ hơn F và N [18, 37], do đó $\mathbf{W}$ và $\mathbf{H}$ có kích thước nhỏ hơn nhiều ma trận ban đầu $\mathbf{V}$.

Công thức (2.1) được viết dưới dạng véc tơ thành $\mathbf{v} \approx \mathbf{W}\mathbf{h}$, với $\mathbf{v}$ và $\mathbf{h}$ là số cột của $\mathbf{V}$ và $\mathbf{H}$. Nói cách khác, mỗi véc tơ $\mathbf{v}$ được xác định bởi tổ hợp tuyến tính của các véc tơ cột của $\mathbf{W}$ với trọng số là véc tơ $\mathbf{h}$. Như vậy, $\mathbf{W}$ được coi là ma trận từ điển chứa các véc tơ đặc trưng của dữ liệu đầu vào cần phân tích, $\mathbf{H}$ là ma trận trọng số, cho biết xác suất xuất hiện các thành phần đặc trưng trong tín hiệu đầu vào tại mỗi thời điểm. Dữ liệu đầu vào kích thước lớn có thể được biểu diễn bởi số lượng ít các vectơ đặc trưng, do đó phép phân tách NMF sẽ đạt kết quả tốt khi véc tơ đặc trưng được chọn đủ để mô tả đầy đủ các đặc trưng của dữ liệu.

Tóm lại, NMF nhằm mục đích tìm ra các yếu tố đại diện cơ bản không âm có thể được sử dụng để trích xuất các đặc trưng của dữ liệu, nhằm giảm kích thước và loại bỏ thông tin dư thừa trong dữ liệu.



Hình 2.1: Mô hình phân tách NMF [13].

2.2.2 Hàm giá

Quá trình phân tách ma trận $\mathbf{V}$ thành $\mathbf{W}$ và $\mathbf{H}$ mong muốn ước lượng được các ma trận thành phần sao cho công thức (2.1) có độ xấp xỉ gần đúng tốt nhất có thể. Quá trình ước lượng được thực hiện bằng thuật toán lặp và cập nhật liên tục các tham số nhằm cực tiểu hàm giá sau [16]:

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

trong đó $D(\mathbf{V} \parallel \mathbf{WH})$ là hàm giá xác định độ sai khác giữa $\mathbf{V}$ và $\mathbf{WH}$. Ký hiệu $\hat{\mathbf{V}} = \mathbf{WH}$, hàm giá được viết thành

$$D(\mathbf{V} \parallel \hat{\mathbf{V}}) = \sum_{f=1}^F \sum_{n=1}^N d(\mathbf{V}_{fm} \parallel \hat{\mathbf{V}}_{fm}). \quad (2.3)$$

Trong công thức (2.3), $d(x \parallel y)$ là hàm xác định độ lệch hay độ sai khác giữa x và y . Có nhiều độ đo khác nhau được dùng tùy theo điều kiện áp dụng và đặc tính của dữ liệu, có thể dùng độ đo Euclidean (EUC) [60], độ đo KullBack-Leibler (KL divergence) [37], hay độ đo Itakura-Saito (IS divergence) [16]. Các độ đo đó có công thức tính như sau:

- EUC distance:

$$d_{EUC}(x \parallel y) = \frac{1}{2}(x - y)^2. \quad (2.4)$$

- KL divergence:

$$d_{KL}(x \parallel y) = x \log\left(\frac{x}{y}\right) - x - y. \quad (2.5)$$

- IS divergence:

$$d_{IS}(x \parallel y) = \frac{x}{y} - \log\left(\frac{x}{y}\right) - 1. \quad (2.6)$$

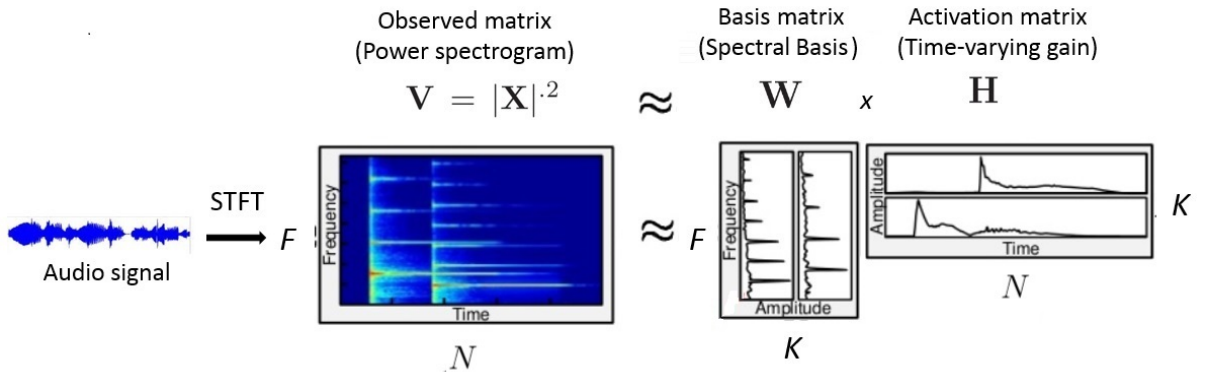
Tùy theo đặc điểm của dữ liệu mà các thuật toán lựa chọn độ đo khác nhau, dẫn đến các hàm giá và công thức cập nhật tham số trong mỗi bước lặp sẽ khác nhau. Các nghiên cứu [16, 19, 67] đã chỉ ra rằng độ đo IS divergence là lựa chọn phù hợp nhất đối với dữ liệu âm thanh.

2.3 Phân tách tiếng nói dựa trên mô hình NMF

2.3.1 Phân tích đặc trưng phổ của âm thanh

NMF đã được sử dụng rộng rãi để phân tách nguồn âm có giám sát dựa trên thông tin phổ của tín hiệu. Theo đó, tín hiệu âm thanh từ miền thời gian $\mathbf{x}(t)$ được biến đổi

sang miền tần số nhờ phép biến đổi Fourier (STFT). Năng lượng phổ của tín hiệu trong miền tần số được tính toán và biểu diễn bằng ma trận không âm V . NMF coi ma trận V là sự kết hợp của ma trận phổ cơ sở W và ma trận kích hoạt H dưới dạng $V = WH$. Các cột của W là các véc tơ phổ cơ sở, biểu diễn sự phân bố năng lượng phổ theo tần số, các âm thanh khác nhau sẽ có đặc điểm phân bố khác nhau do đó ma trận phổ cơ sở của chúng sẽ khác nhau. Các hàng trong H cho biết tỷ lệ kết hợp của các phổ cơ sở theo thời gian. Gọi K , F và N số lượng phổ cơ sở, số bin tần số và kích thước dữ liệu (số lượng mẫu theo thời gian), mô hình NMF đơn giản với hai véc tơ phổ cơ sở được biểu thị trong Hình. 2.2 [33]. Trong ví dụ này, hai véc tơ phổ cơ sở đại diện cho hai đặc trưng phân bố năng lượng phổ khác nhau tạo thành ma trận từ điển W .



Hình 2.2: Mô hình NMF trích chọn đặc trưng phổ của âm thanh ($K = 2$).

Số lượng phổ cơ sở K được chọn ảnh hưởng trực tiếp đến hiệu quả phân tích các đặc tính phổ của tín hiệu. Nếu K lớn, các đặc tính phổ có thể được mô hình hóa một cách chi tiết, nhưng việc ước tính tham số sẽ đòi hỏi chi phí thời gian cao hơn. Ngược lại, nếu K quá nhỏ, một số đặc tính quan trọng của âm thanh cần được mô hình hóa có thể sẽ bị bỏ sót. Do đó, việc lựa chọn K là rất quan trọng và nó phụ thuộc vào từng loại âm thanh khác nhau. Trong đề tài, chúng tôi thiết lập tham số này dựa theo các tài liệu uy tín đã được công bố [19, 67, 86], chúng tôi chọn $K = 32$ đối với tiếng nói, $K = 16$ đối với âm thanh nhiễu môi trường và $K = 25$ đối với âm nhạc (music).

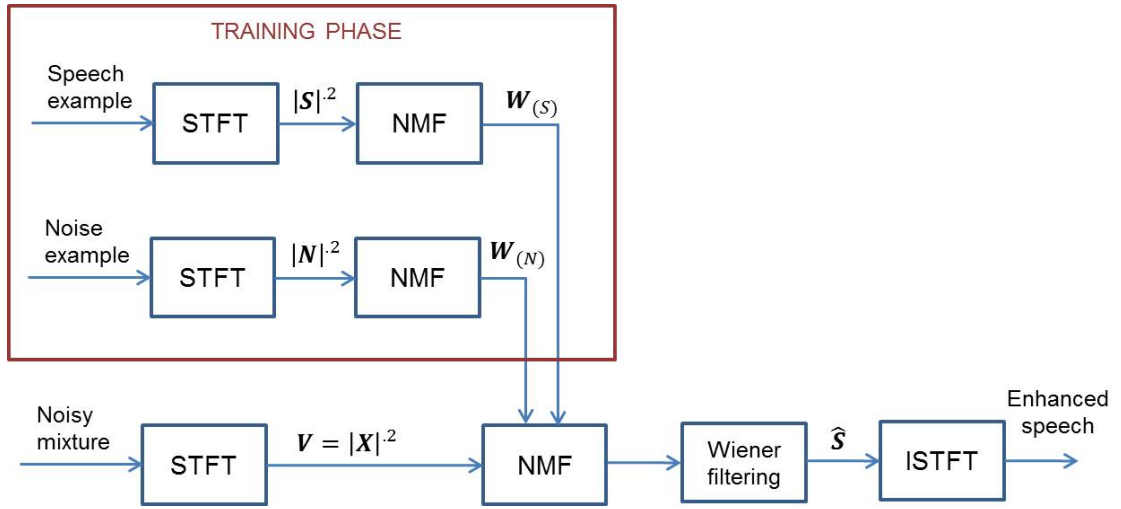
2.3.2 NMF-based audio source separation

Nhiều nghiên cứu trong lĩnh vực xử lý tín hiệu âm thanh đã sử dụng mô hình NMF để phân tích âm thanh trong các tình huống và điều kiện khác nhau [15, 66, 68]. Chúng tôi xem xét bài toán tách tín hiệu trộn thành hai thành phần: tiếng nói mong muốn

(speech) và nhiễu (noise). Lưu ý rằng âm thanh nhiễu có thể gồm nhiều âm thanh thành phần, nhiễu cũng có thể chứa các giọng nói khác.

Một cách khái quát, mô hình phân tách âm thanh làm việc với tín hiệu trong miền thời gian- tần số sau phép biến đổi Short-Time Fourier Transform (STFT) và gồm có hai pha như trong Hình. 2.3:

- Pha huấn luyện (1): ước lượng các đặc trưng phổ của tín hiệu từ các mẫu huấn luyện,
- Pha tách (2): Tách các tín hiệu tiếng nói và nhiễu từ tín hiệu mixture theo "hướng dẫn" từ các đặc trưng phổ đã được huấn luyện trong bước (1).



Hình 2.3: Mô hình tách tiếng nói và nhiễu theo thuật toán Supervised-NMF.

Gọi $\mathbf{X} \in \mathbb{C}^{F \times M}$, $\mathbf{S}_1 \in \mathbb{C}^{F \times M}$, và $\mathbf{S}_2 \in \mathbb{C}^{F \times M}$ lần lượt là các ma trận biểu diễn tín hiệu mixture, speech và noise sau phép biến đổi STFT trên miền số thực, trong đó F là số bin tần số (frequency bins) và M là số khung thời gian (time frames). Công thức quan hệ giữa chúng như sau:

$$\mathbf{X} = \mathbf{S}_1 + \mathbf{S}_2. \quad (2.7)$$

Gọi $\mathbf{V} = |\mathbf{X}|^2$ là ma trận năng lượng phổ của tín hiệu mixture (power spectrogram), với $|\mathbf{X}|^p$ là ma trận với các phần tử là $[\mathbf{X}]_{il}^p$. Khi đó NMF sẽ phân tách ma trận không âm $\mathbf{V}$ kích thước $F \times M$ thành hai ma trận không âm thành phần $\mathbf{W} \in \mathbb{R}^{F \times K}$ và $\mathbf{H} \in \mathbb{R}^{K \times M}$. K thường được chọn sao cho nhỏ hơn F và M . Để ước lượng các ma trận thành phần, độ sai khác giữa $\mathbf{V}$ và $\mathbf{WH}$ sẽ được cực tiểu hóa theo độ đo IS-divergence [15].

Việc phân tách các ma trận được thực hiện theo tiêu chí cực tiểu hàm giá 2.2 với điều

kiện ràng buộc $\mathbf{W}, \mathbf{H} \geq 0$ [15]. Trong đó $D(\mathbf{V} \|\hat{\mathbf{V}}) = \sum_{f=1}^F \sum_{m=1}^M d_{IS}(\mathbf{V}_{fm} \|\hat{\mathbf{V}}_{fm})$, $\hat{\mathbf{V}} = \mathbf{WH}$, f và m lần lượt là chỉ số bin tần số (frequency bin) và chỉ số khung thời gian lấy mẫu (time frame). Các tham số $\theta = \{\mathbf{W}, \mathbf{H}\}$ thường được khởi tạo với các giá trị không âm ngẫu nhiên và được cập nhật trong từng bước lặp bằng quy tắc cập nhật multiplicative update (MU rules) [15] theo công thức sau:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T ((\mathbf{WH})^{\cdot -2} \odot \mathbf{V})}{\mathbf{W}^T (\mathbf{WH})^{\cdot -1}} \quad (2.8)$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{((\mathbf{WH})^{\cdot -2} \odot \mathbf{V}) \mathbf{H}^T}{(\mathbf{WH})^{\cdot -1} \mathbf{H}^T} \quad (2.9)$$

với $\mathbf{A}^T$ là ma trận chuyển vị của ma trận $\mathbf{A}$, $\odot$ là phép element-wise Hadamard.

Thuật toán NMF ước lượng hai ma trận thành phần $\mathbf{W}$ and $\mathbf{H}$ từ ma trận đầu vào $\mathbf{V}$ được mô tả trong Algorithm 1. Đầu vào của thuật toán là ma trận $\mathbf{V}$ và số lượng phổ cơ sở K . Tham số β quy định độ đo dùng cho hàm giá: $\beta = 0$ tương ứng với độ đo IS-divergence, $\beta = 1$ là KL-divergence, *niter* là số vòng lặp của thuật toán.

Ký hiệu $\mathbf{W}_1$ và $\mathbf{W}_2$ lần lượt là ma trận đặc trưng phổ của tín hiệu tiếng nói và nhiễu. Trong trường hợp học có hướng dẫn (supervised), $\mathbf{W}_1$ và $\mathbf{W}_2$ được ước lượng qua bước huấn luyện từ các dữ liệu mẫu tương ứng. Sau đó ma trận phổ của hai nguồn $\mathbf{W}$ được xây dựng bằng công thức

$$\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2]. \quad (2.10)$$

Trong pha test (speech enhancement process), NMF sẽ cố định ma trận đặc trưng phổ của tín hiệu $\mathbf{W}$, và ước lượng ma trận time activation $\mathbf{H}$ bằng công thức cập nhật MU-rule theo công thức (2.8), và $\mathbf{H}$ cũng là ma trận gồm hai phần tương ứng với hai nguồn cần tách

$$\mathbf{H} = [\mathbf{H}_1^T, \mathbf{H}_2^T]^T, \quad (2.11)$$

trong đó $\mathbf{H}_1$ và $\mathbf{H}_2$ là các ma trận kích hoạt tương ứng của tín hiệu tiếng nói và nhiễu, $\mathbf{A}^T$ là ma trận chuyển vị của ma trận $\mathbf{A}$.

Mô hình hoạt động của NMF trong bước phân tách tín hiệu trộn thành hai âm thanh thành phần được mô tả trong Hình 2.4.

Algorithm 1 NMF algorithm

Require: $\mathbf{V}$, K , β , $niter$.

Ensure: $\mathbf{W}$, $\mathbf{H}$.

```
// Parameters initialization
Initialize  $\mathbf{H}^{(0)}$ ,  $\mathbf{W}^{(0)}$  randomly with nonnegative values
 $t = 0$ ;
 $\hat{\mathbf{V}} \leftarrow \mathbf{W}^{(0)}\mathbf{H}^{(0)}$ 
// Parameters estimation
repeat
   $t = t + 1$ 
  //Update  $\mathbf{W}^{(t)}$ 
   $\mathbf{W} \leftarrow \mathbf{W} \odot \left( \frac{(\hat{\mathbf{V}}^{(\beta-2)} \odot \mathbf{V})\mathbf{H}^T}{(\hat{\mathbf{V}}^{(\beta-1)})\mathbf{H}^T} \right)$ 
  //Update  $\mathbf{H}^{(t)}$ 
   $\mathbf{H} \leftarrow \mathbf{H} \odot \left( \frac{\mathbf{W}^T(\hat{\mathbf{V}}^{(\beta-2)} \odot \mathbf{V})}{\mathbf{W}^T(\hat{\mathbf{V}}^{(\beta-1)})} \right)$ 
   $\hat{\mathbf{V}} \leftarrow \mathbf{W}\mathbf{H}$ 
until convergence
return  $\mathbf{W}$ ,  $\mathbf{H}$ 
```

Algorithm 2 Baseline NMF-based audio source separation algorithm

Require:

Mixture signal $\mathbf{x}(t)$

Training data of all source $\{s_j(t)\}, j = 1 : J$

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

//TRAINING PHASE

for $j = 1, \dots, J$ **do**

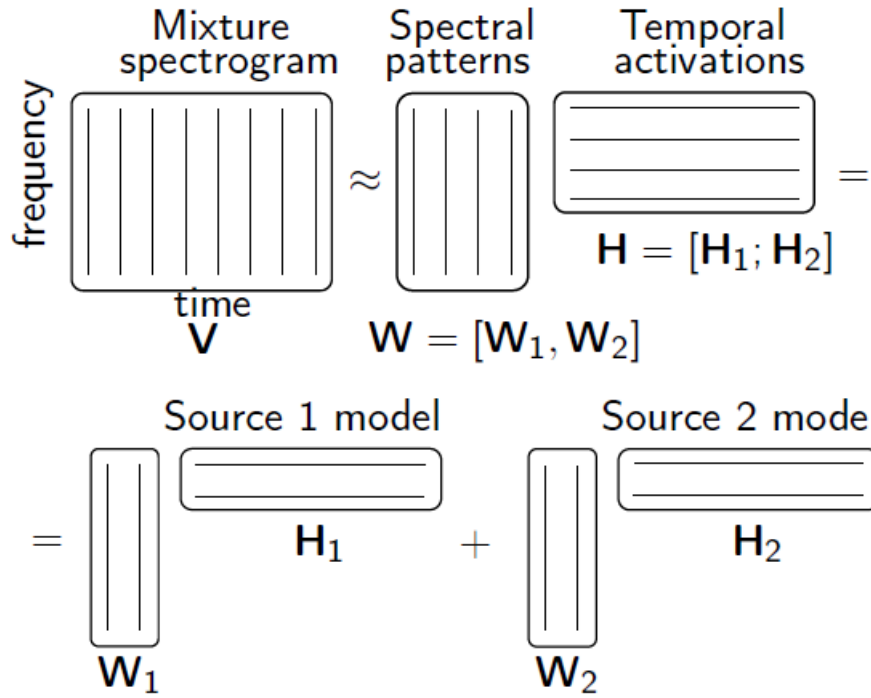
- Estimating the spectral basis matrix $\mathbf{W}_j$ for j -th source from training example $s_j(t)$ by Algorithm 1

end for

$\mathbf{W} = [\mathbf{W}_1, \dots, \mathbf{W}_J]$

//SEPARATION PHASE

- Initialize the nonnegative matrix $\mathbf{H}$
 - Estimating $\mathbf{H}$ from mixture signal $\mathbf{x}(t)$ by Algorithm 1 ($\mathbf{W}$ is fixed).
 - Estimating $\hat{\mathbf{S}}, \hat{\mathbf{N}}$ by Wiener filtering (2.13)
 - $\hat{\mathbf{s}}(t) = \text{ISTFT}(\hat{\mathbf{S}})$
-



Hình 2.4: Mô hình NMF phân tách tín hiệu âm thanh.

Sau khi ước lượng các thành phần $\theta = \{W, H\}$, tín hiệu speech và noise được tính bởi công thức Wiener filtering như sau:

$$\hat{S}_1 = \frac{W_{(1)} H_{(1)}}{WH} \odot X, \quad (2.12)$$

$$\hat{S}_2 = \frac{W_{(2)} H_{(2)}}{WH} \odot X. \quad (2.13)$$

Cuối cùng các tín hiệu speech và noise tách được sẽ được chuyển sang miền thời gian (time domain) bằng phép biến đổi STFT ngược (inverse STFT). Thuật toán được mô tả như trong Algorithm 2.

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 W đã được học trước từ dữ liệu huấn luyện. Thuật toán sẽ cho kết quả tốt khi có dữ liệu huấn luyện, tuy nhiên trong trường hợp không có dữ liệu huấn luyện thì kết quả phân tách nguồn đạt được là rất thấp. Điều này được thể hiện trong các kết quả thí nghiệm trong chương 4.

2.4 Kết luận

Chương này giới thiệu về NMF, kỹ thuật được sử dụng rộng rãi trong vấn đề xử lý dữ liệu đa chiều nói chung cũng như trong lĩnh vực xử lý âm thanh nói riêng. Chúng

tôi cũng trình bày mô hình phân tích đặc trưng phổ âm thanh bằng NMF, thuật toán tách nguồn âm thanh dựa trên NMF và coi đó là thuật toán cơ sở để phát triển nghiên cứu của mình. Trong chương sau, chúng tôi sẽ đề xuất thuật toán tách nguồn đơn kênh sử dụng NMF kết hợp với điều kiện ràng buộc thưa giải quyết bài toán tách tiếng nói trong điều kiện thông tin hướng dẫn ít (weakly-informed).

CHƯƠNG 3

THUẬT TOÁN PHÂN TÁCH TIẾNG NÓI KẾT HỢP MÔ HÌNH THỪA SỐ HÓA MA TRẬN KHÔNG ÂM VỚI RẰNG BUỘC THỪA

3.1 Đặt vấn đề

Những công bố gần đây về tách nguồn âm cho thấy thuật toán tách nguồn mù, hoàn toàn không dùng dữ liệu huấn luyện hay các thông tin phụ trợ hướng dẫn phân tách, cho kết quả chưa đủ tốt để đưa vào ứng dụng thực tế. Các thuật toán sử dụng dữ liệu huấn luyện chính xác cho các nguồn cần tách, ví dụ như thuật toán Supervised-NMF được trình bày trong phần 2.3, hoặc dùng các thông tin hướng dẫn 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 [44, 53, 55]. Tuy nhiên trong nhiều tình huống thực tế, dữ liệu huấn luyện hay những thông tin chính xác đó thường không có sẵn. Ví dụ, hệ thống nhận dạng giọng nói cần hoạt động trong nhiều môi trường khác nhau, do đó tín hiệu nhiễu và/hoặc giọng nói sẽ không xác định.

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 vắng 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, hay nhiễu môi trường,...) để tìm kiếm những mẫu âm thanh cùng loại làm dữ liệu huấn luyện. Số lượng mẫu huấn luyện cho mỗi loại tín hiệu chúng tôi thử nghiệm là từ 3 đến 5 mẫu, kích thước mỗi mẫu là rất nhỏ từ 5 đến 10 giây. 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.

Cụ thể trong chương 3, nhóm nghiên cứu trình bày cách xây dựng mô hình phổ tổng quát GSSM cho tín hiệu speech and noise trong phần 3.2. Sau đó giới thiệu về hai sparsity constraint cho tín hiệu âm thanh đã được công bố trong các nghiên cứu gần đây [11, 63, 68] trong phần 3.3. Cuối cùng nhóm nghiên cứu đề xuất phương pháp kết hợp hai sparsity constraint đó trong một mô hình duy nhất để kết hợp với NMF ước

lượng tham số $\mathbf{H}$ trong phần 3.4. Trong phần này, nhóm nghiên cứu cũng xây dựng hai mô hình thuật toán cho hai tình huống:

- Unsupervised: không có dữ liệu huấn luyện chính xác cho speech và noise, khi đó mô hình phổ tổng quát của cả speech và noise sẽ được xây dựng bằng cách học từ các mẫu speech và noise "tương tự" thu thập được.

- Semi-supervised: chỉ có dữ liệu huấn luyện chính xác cho speech và không có dữ liệu huấn luyện chính xác cho noise, khi đó mô hình phổ tổng quát của noise được xây dựng từ các tín hiệu noise tương tự thu thập được.

Chú ý rằng trong cả hai tình huống nêu trên, các mô hình phổ tổng quát được xây dựng từ một tập hợp nhiều tín hiệu cùng loại nhưng khác với tín hiệu cần tách. Ví dụ, mô hình phổ cho speech được xây dựng bằng cách dùng 3 file speech (2 giọng nữ và 1 giọng nam) trong pha huấn luyện nhưng các giọng nói này đều khác giọng nói cần tách, tương tự có thể dùng 5 file tín hiệu nhiễu môi trường khác nhau để huấn luyện mô hình phổ tổng quát của tín hiệu nhiễu và chúng cũng khác với tín hiệu nhiễu cần tách.

3.2 Xây dựng mô hình phổ tổng quát GSSM

Từ thông tin đã biết về loại nguồn cần phân tách, chúng tôi thu thập các mẫu huấn luyện cùng loại (có thể dễ dàng thu thập được từ các kho dữ liệu được cung cấp miễn phí). 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 3.3: (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 kết hợp với hàm ràng buộc thưa.

Ký hiệu $\mathbf{V}_S^p$ và $\mathbf{V}_N^q$ lần lượt là ma trận phổ của mẫu huấn luyện speech thứ p và noise thứ q . $\mathbf{V}_S^p$ và $\mathbf{V}_N^q$ được dùng để huấn luyện mô hình đặc trưng phổ $\mathbf{W}_{(S)}^p$ and $\mathbf{W}_{(N)}^q$ qua NMF theo công thức (2.1):

$$\min_{\mathbf{H}_{(S)}^p \geq 0, \mathbf{W}_{(S)}^p \geq 0} D(\mathbf{V}_S^p \| \mathbf{W}_{(S)}^p \mathbf{H}_{(S)}^p), \quad (3.1)$$

$$\min_{\mathbf{H}_{(N)}^q \geq 0, \mathbf{W}_{(N)}^q \geq 0} D(\mathbf{V}_N^q \| \mathbf{W}_{(N)}^q \mathbf{H}_{(N)}^q), \quad (3.2)$$

ở đó $\mathbf{H}_{(S)}^p$ và $\mathbf{H}_{(N)}^q$ là các ma trận time activation.

Sau khi ước lượng các thành phần $\mathbf{W}_{(S)}^p$ và $\mathbf{W}_{(N)}^q$ theo tập dữ liệu mẫu với số lượng mẫu speech là $p = 1, \dots, P$, số lượng mẫu noise là $q = 1, \dots, Q$, mô hình phổ tổng quát của speech và noise được xây dựng như sau:

$$\mathbf{W}_S = [\mathbf{W}_{(S)}^1, \dots, \mathbf{W}_{(S)}^P] \quad (3.3)$$

$$\mathbf{W}_N = [\mathbf{W}_{(N)}^1, \dots, \mathbf{W}_{(N)}^Q]. \quad (3.4)$$

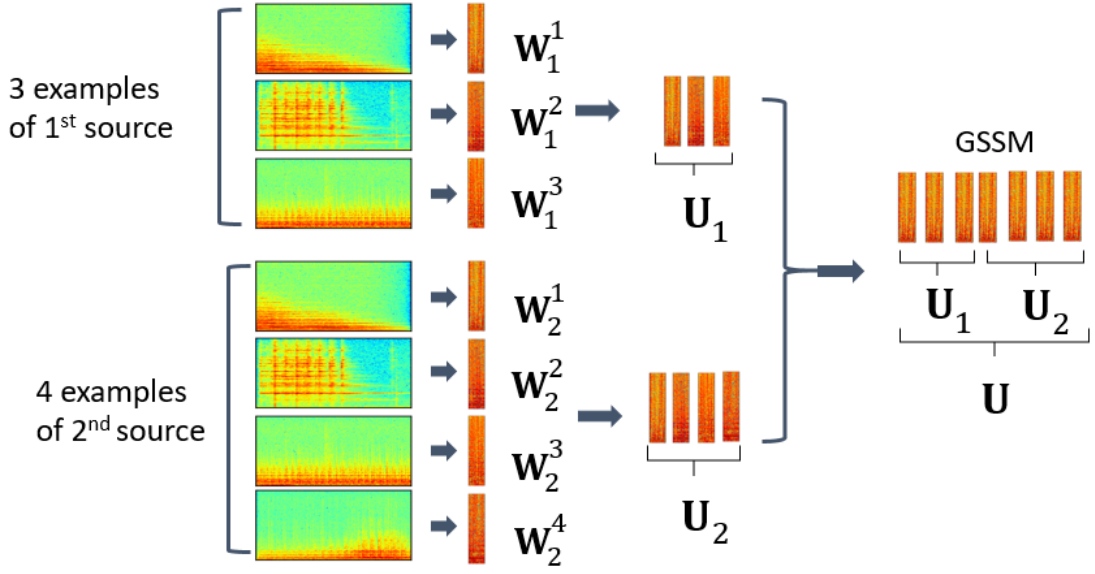
Trong thí nghiệm, nhóm nghiên cứu đã thực hiện với dữ liệu huấn luyện gồm một số giọng nam và nữ khác nhau (ví dụ $P = 3$), và một số loại nhiễu môi trường như tiếng gió, tiếng ồn trong quán cà phê, âm thanh đường phố, tiếng sóng biển, ...(ví dụ $Q = 5$).

Cách xây dựng ma trận phổ tổng quát GSSM từ các mẫu thu thập được mô tả trong hình. Ref fig: c3-GSSM.

3.3 Các ràng buộc thừa Block sparsity và Copmponent sparsity

Các mô hình phổ tổng quát $\mathbf{W}_S$ và $\mathbf{W}_N$ được xây dựng theo công thức (3.3) và (3.4) sẽ có kích thước lớn khi số lượng mẫu huấn luyện cho các nguồn tăng lên. Trong ma trận lớn đó thường chỉ có một phần kích thước nhỏ chứa các đặc trưng phổ giống với phổ tín hiệu cần tách. Chính vì lý do đó, trong pha tách các nguồn âm từ ma trận phổ của tín hiệu mixture, ràng buộc sparsity constraint được sử dụng với mục đích tìm ra những tập con của ma trận phổ tổng quát $\mathbf{W} = [\mathbf{W}_S, \mathbf{W}_N]$ có chứa các đặc trưng phổ của tín hiệu nguồn cần tách [25, 38, 80]. Nói cách khác, ma trận phổ của tín hiệu mixture $\mathbf{V} = |\mathbf{X}|^2$ sẽ được phân tách dựa trên việc tối ưu hóa hàm giá sau đây:

$$\min_{\mathbf{H} \geq 0} D(\mathbf{V} \| \mathbf{W}\mathbf{H}) + \lambda \Omega(\mathbf{H}) \quad (3.5)$$



Hình 3.1: Cách xây dựng mô hình phổ tổng quát GSSM.

ở đó $\Omega(\mathbf{H})$ (được gọi là penalty function) là hàm thể hiện sparsity constraint trên ma trận $\mathbf{H}$, λ là tham số thể hiện sự ảnh hưởng nhiều hay ít của sparsity constraint (được gọi là trade-off parameter). Khi $\lambda = 0$, $\mathbf{H}$ không bị ảnh hưởng bởi sparsity constraint và hình ảnh của ma trận $\mathbf{H}$ như trong hình Fig. 3.2a.

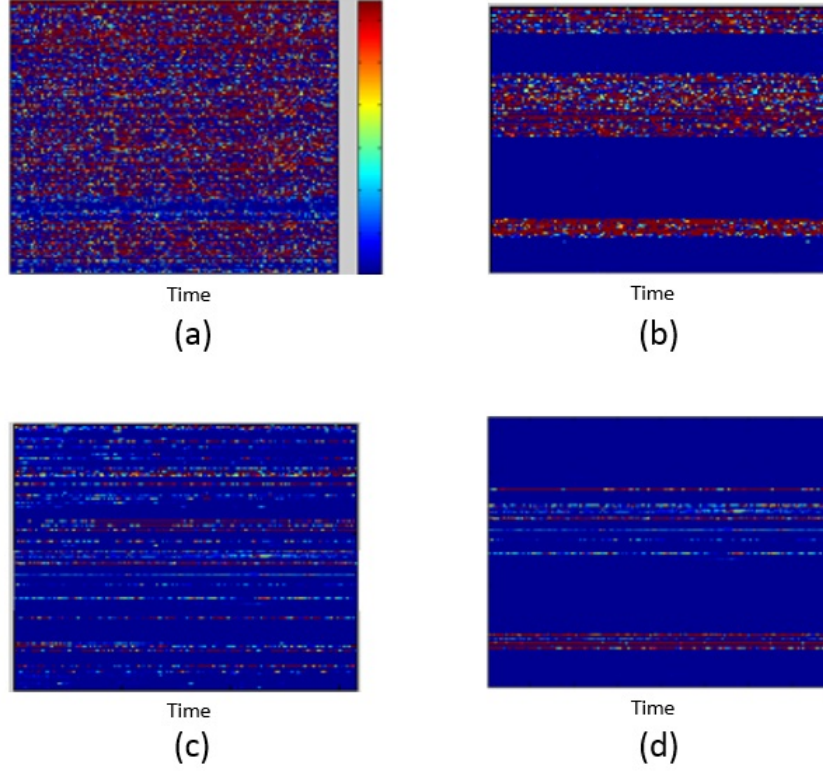
Các nghiên cứu được công bố gần đây trong lĩnh vực tách nguồn âm thanh đã xem xét hai penalty functions tương ứng với hai nhóm ràng buộc thừa sau:

3.3.1 Block sparsity

Để loại bỏ những mẫu huấn luyện không quan trọng là những phần không chứa các đặc trưng phổ tương tự như những đặc trưng của nguồn cần tách, hàm ràng buộc thừa tương ứng với block sparsity đã được đề xuất như sau [63]:

$$\Omega_1(\mathbf{H}) = \sum_{g=1}^G \log(\epsilon + \|\mathbf{H}_{(g)}\|_1), \quad (3.6)$$

với $\mathbf{H}_{(g)}$ là tập con của $\mathbf{H}$ và là ma trận activation của block thứ g , G là tổng số các block, ϵ là một hằng số dương nhỏ. Trong trường hợp này, mỗi block đại diện cho một mẫu huấn luyện và G là tổng số mẫu huấn luyện được sử dụng ($G = P + Q$). Hàm penalty sẽ kích hoạt những block chứa các đặc trưng tương tự với nguồn cần tách và các block còn lại sẽ hội tụ về 0 như quan sát thấy trong Hình. 3.2b.



Hình 3.2: Hình ảnh của ma trận activation $\mathbf{H}$: (a) NMF không dùng sparsity constraint, (b) NMF với block sparsity (3.6), (c) NMF với component sparsity (3.7), và (d) Kết hợp block sparsity và component sparsity (3.8).

3.3.2 Component sparsity

Hàm ràng buộc thứ hai đã được nghiên cứu và đề xuất tương ứng với component sparsity:

$$\Omega_2(\mathbf{H}) = \sum_{k=1}^K \log(\epsilon + \|\mathbf{h}_k\|_1), \quad (3.7)$$

ở đó $\mathbf{h}_k$ là dòng thứ k của ma trận $\mathbf{H}$. Theo tài liệu [11], thay vì kích hoạt cả block, hàm penalty $\Omega_2(\mathbf{H})$ chỉ lựa chọn kích hoạt hoặc không kích hoạt từng component tương ứng với từng véc tơ "dòng" trong $\mathbf{W}$. Hình ảnh của ma trận $\mathbf{H}$ khi áp dụng component sparsity như trong Hình. 3.2c).

3.4 Đề xuất công thức ràng buộc thưa kết hợp

Xuất phát từ nhận định thường chỉ có một phần của mô hình phổ tổng quát GSSM có chứa các đặc trưng của tín hiệu cần tách và các đặc trưng quan trọng đó thường nằm

rải rác trong các mẫu huấn luyện khác nhau chứ không tập trung về một vài mẫu cụ thể. Như vậy block sparsity có thể bỏ qua những mẫu có sự tương đồng tương đối ít với tín hiệu cần tách, trong khi đó component sparsity lại loại bỏ tương đối ít và thường giữ lại cả những đặc tính không mấy tương đồng với tín hiệu cần tách. Ngoài ra, khi GSSM có kích thước lớn (khi số lượng mẫu huấn luyện cho các nguồn lớn và số nguồn tăng) thì tốc độ hội tụ khi sử dụng component sparsity sẽ rất chậm vì đòi hỏi thuật toán phải duyệt và xử lý nhiều lần toàn bộ ma trận GSSM lớn. Từ những phân tích đó, để nâng cao hiệu quả của các hàm penalty (3.6) và (3.7), nhóm nghiên cứu đã đề xuất phương pháp kết hợp hai sparsity được trình bày trong phần 3.3 và áp dụng để giải quyết vấn đề được đặt ra trong chuyên đề là nâng cao chất lượng tiếng nói trong điều kiện thu âm có nhiều môi trường.

Công thức kết hợp được đề xuất như sau [10]:

$$\Omega_{\text{new}}(\mathbf{H}) = \alpha \sum_{g=1}^G \log(\epsilon + \|\mathbf{H}_{(g)}\|_1) + (1 - \alpha) \sum_{k=1}^K \log(\epsilon + \|\mathbf{h}_k\|_1), \quad (3.8)$$

với α là trọng số thể hiện sự đóng góp của mỗi thành phần.

Có thể xem (3.8) là sự tổng quát hóa của (3.6) và (3.7), khi $\alpha = 1$, (3.8) tương đương với (3.6) và khi $\alpha = 0$, (3.8) tương đương với (3.7). Hình. 3.2d là hình ảnh của ma trận kích hoạt $\mathbf{H}$ sau khi áp dụng hàm ràng buộc thừa (3.8). Có thể quan sát thấy khi đó một vài block hội tụ về 0 do ảnh hưởng của thành phần thứ nhất trong (3.8). Với những block còn lại, một vài dòng (components) cũng hội tụ về 0 do ảnh hưởng của thành phần thứ hai trong (3.8).

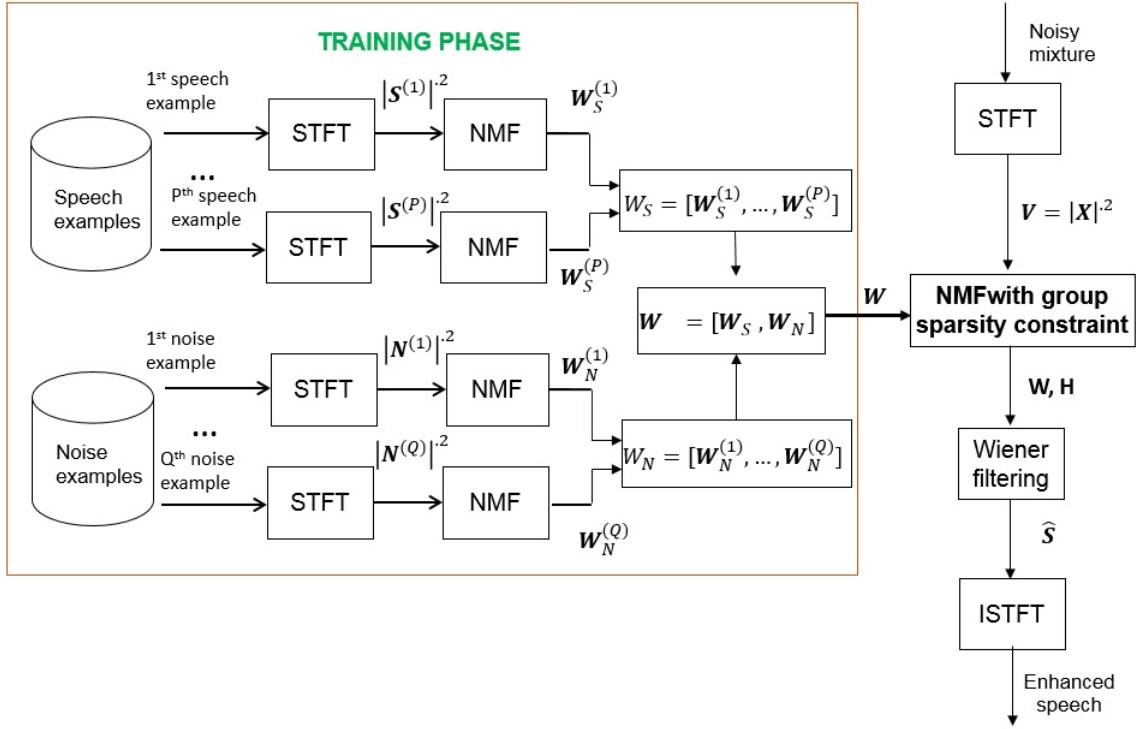
3.5 Xây dựng mô hình thuật toán

Để tối ưu hóa các tham số NMF trong công thức (3.5) với hàm ràng buộc thừa (3.8) đề xuất, chúng tôi dùng công thức cập nhật tham số MU-rules trong các bước lặp.

3.6 Thuật toán đề xuất cho trường hợp Weakly-informed

Trong phần này chúng tôi xây dựng thuật toán cho tình huống nâng cao chất lượng tiếng nói khi không có dữ liệu huấn luyện chính xác cho cả tiếng nói và nhiễu như đã nêu trong phần 2.1:

Trong tình huống không có dữ liệu huấn luyện chính xác cho cả speech và noise, khi đó ở bước huấn luyện, mô hình phổ tổng quát của speech và noise cần được xây dựng bằng cách áp dụng mô hình NMF trên tập dữ liệu huấn luyện cùng loại thu thập được (giả sử gồm P mẫu speech và Q mẫu noise khác nhau). Trong bước phân tách, thuật toán NMF kết hợp với ràng buộc thưa đề xuất được sử dụng để ước lượng ma trận kích hoạt H . Mô hình thuật toán được thể hiện trong Hình 3.3.



Hình 3.3: Mô hình thuật toán UnSupervised-NMF phân tách tiếng nói và nhiễu sử dụng kết hợp mô hình NMF với nhóm ràng buộc thưa đề xuất.

Thuật toán được mô tả chi tiết trong Algorithm 3, ở đó $Y_{(g)}$ là ma trận có cùng kích thước với $H_{(g)}$, z_k là véc tơ dòng có cùng kích thước với h_k .

3.7 Thuật toán đề xuất cho trường hợp Semi-supervised

Trong nhiều trường hợp thực tế, tín hiệu speech cần tách là xác định được (Ví dụ như trong điều khiển rô bốt, nhà thông minh,... đã xác định người dùng). Khi đó đã có dữ liệu huấn luyện chính xác cho tín hiệu speech, công thức (3.1) và (3.3) được áp dụng trong trường hợp này với $P = 1$. Ở bước huấn luyện chỉ cần xây dựng mô hình phổ tổng quát cho tín hiệu noise dựa trên tập gồm Q dữ liệu huấn luyện cùng loại thu

Algorithm 3 Unsupervised-NMF with mixed sparsity-inducing penalty

Require: $\mathbf{V}, \mathbf{W}, \lambda, \alpha$

Ensure: $\mathbf{H}$

Khởi tạo $\mathbf{H}$ với các giá trị không âm ngẫu nhiên

$$\hat{\mathbf{V}} = \mathbf{W}\mathbf{H}$$

repeat

// Tính toán thành phần block sparsity-inducing penalty

for $g = 1, \dots, G$ **do**

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

end for

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

// Tính toán thành phần component sparsity-inducing penalty

for $k = 1, \dots, K$ **do**

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

end for

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

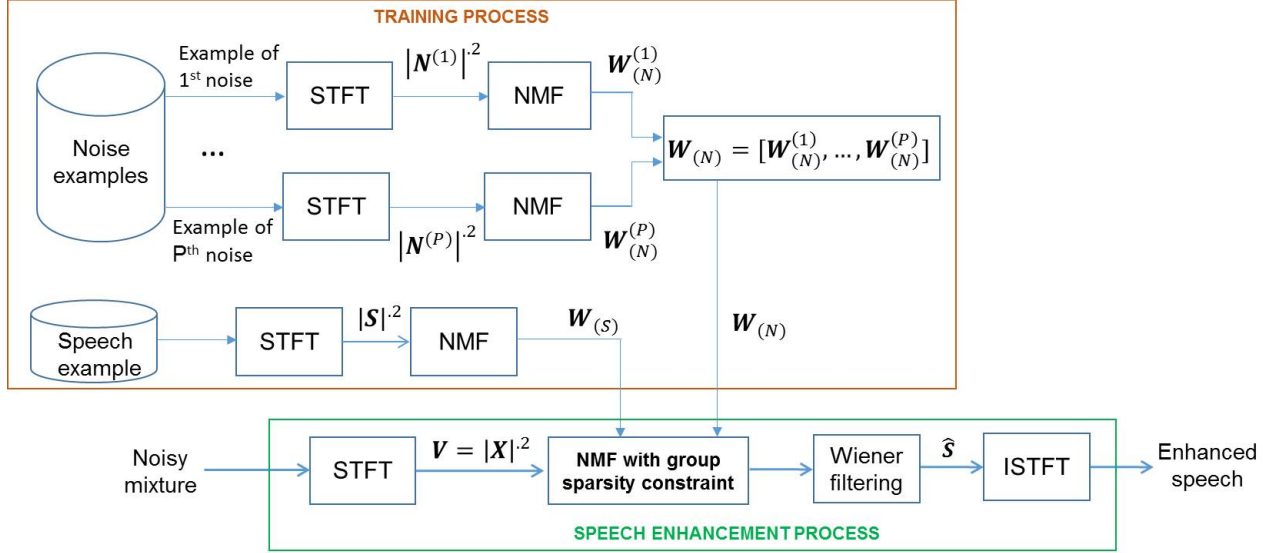
// Cập nhật ma trận activation $\mathbf{H}$

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

$$\hat{\mathbf{V}} \leftarrow \mathbf{W}\mathbf{H}$$

until Thỏa mãn điều kiện hội tụ

thập được theo công thức (3.2) và (3.4). Trong bước tách thuật toán NMF kết hợp với công thức ràng buộc thừa đề xuất chỉ được sử dụng để ước lượng phần ma trận activate $\mathbf{H}_{(N)}$ tương ứng với tín hiệu noise. Ma trận kích hoạt của tín hiệu tiếng nói được cập nhật theo công thức MU-rules thông thường 2.8 trong chương 2. Mô hình thuật toán được thể hiện trong Hình 3.4.



Hình 3.4: Mô hình tách tiếng nói và nhiễu theo thuật toán SemiSupervised-NMF kết hợp với mixed sparsity constraints.

Thuật toán được mô tả chi tiết trong Algorithm 4, ở đó $\mathbf{Y}_{(g)}$ là ma trận có cùng kích thước với $\mathbf{H}_{(g)}$, $\mathbf{z}_k$ là véc tơ dòng có cùng kích thước với $\mathbf{h}_k$.

3.8 Kết luận

Trong chương này, chúng tôi đã tập trung nghiên cứu cách áp dụng NMF kết hợp với các nhóm ràng buộc thừa trong nâng cao chất lượng tiếng nói. Xây dựng mô hình thuật toán giải quyết tình huống tách các nguồn âm thanh từ tín hiệu trộn không có dữ liệu huấn luyện. Nhóm nghiên cứu đã đề xuất công thức kết hợp hai ràng buộc thừa được công bố gần đây là block sparsity và component sparsity trong quá trình ước lượng ma trận kích hoạt của các nguồn âm cần tách, đồng thời xây dựng thuật toán cho hai trường hợp: không có dữ liệu huấn luyện và có dữ liệu huấn luyện cho một loại tín hiệu. Trong chương sau, chúng tôi sẽ mô tả bộ dữ liệu thí nghiệm, cách đánh giá kết quả cũng như đưa ra các kết quả thí nghiệm của các thuật toán đã xây dựng.

Algorithm 4 SemiSupervised-NMF with mixed sparsity-inducing penalty

Require: $\mathbf{V}, \mathbf{W}_{(N)}, \mathbf{W}_{(S)}, \lambda, \alpha$

Ensure: $\mathbf{H}_{(S)}, \mathbf{H}_{(N)}$

Khởi tạo $\mathbf{H}_{(S)}, \mathbf{H}_{(N)}$ với các giá trị không âm ngẫu nhiên, $\mathbf{H} = [\mathbf{H}_{(S)}^T, \mathbf{H}_{(N)}^T]^T$

$\hat{\mathbf{V}} = \mathbf{W}\mathbf{H}$, với $\mathbf{W} = [\mathbf{W}_{(S)}, \mathbf{W}_{(N)}]$ cố định

repeat

 // Tính toán thành phần block sparsity-inducing penalty

for $g = 1, \dots, G$ **do**

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

end for

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

 // Tính toán thành phần component sparsity-inducing penalty

for $k = 1, \dots, K$ **do**

$$\mathbf{z}_k \leftarrow \frac{1}{\epsilon + \|\mathbf{h}_{(N)}^k\|}$$

end for

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

 // Cập nhật ma trận activation $\mathbf{H}$

$$\mathbf{H}_{(S)} \leftarrow \mathbf{H}_{(S)} \odot \frac{\mathbf{w}_{(S)}^T (\hat{\mathbf{V}}^{\cdot -2} \odot \mathbf{V})}{\mathbf{w}_{(S)}^T (\hat{\mathbf{V}}^{\cdot -1})}$$

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

$$\hat{\mathbf{V}} \leftarrow \mathbf{W}\mathbf{H}$$

until thỏa mãn điều kiện hội tụ

CHƯƠNG 4

THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

4.1 Phương pháp đánh giá kết quả phân tách âm thanh

Vấn đề làm thế nào để đánh giá hiệu quả của các phương pháp tách nguồn âm thanh đã được nghiên cứu từ lâu trên thế giới. Một số nghiên cứu đã đưa ra các tiêu chí đánh giá kết quả của mô hình tách nguồn âm trên cả hai phương diện khách quan (objective quality) [21, 77] and subjective quality [12, 20, 78] và chủ quan (subjective quality) [12, 20, 78]. Chúng tôi quan tâm đến nhóm tiêu chí đánh giá khách quan, do nó có thể dùng cho mọi trường hợp của tín hiệu mixture cũng như mọi thuật toán tách nguồn khác nhau. Các tiêu chí này đánh giá kết quả nguồn tách mà không phụ thuộc vào thuật toán tách nguồn cũng như các tham số của bước tách. Nhóm tiêu chí đó là Energy-based criteria hay còn gọi là Energy ratio criteria (SDR, SIR, SAR, ISR). Chúng được sử dụng phổ biến nhất trong cộng đồng nghiên cứu về xử lý âm thanh trên thế giới hiện nay [2, 31, 54, 56, 72, 73, 75, 79].

Theo phương pháp đánh giá đó, sự biến dạng của tín hiệu ước lượng được qua thuật toán tách nguồn còn gọi là độ sai khác so với tín hiệu nguồn đúng được phân tách ra thành ba thành phần: biến dạng do không gian truyền, biến dạng do sự ảnh hưởng của các nguồn khác, biến dạng do sự can thiệp của các phép biến đổi và thuật toán.

Các tiêu chí đánh giá được xây dựng từ việc coi tín hiệu của mỗi nguồn âm tại đầu thu (source image) ước lượng được qua thuật toán $\hat{c}_{ij}(t)$ được phân tách thành bốn thành phần như sau [79]:

$$\hat{c}_{ij}(t) = c_{ij}(t) + e_{ij}^{spat}(t) + e_{ij}^{inter}(t) + e_{ij}^{artif}(t), \quad (4.1)$$

với $c_{ij}(t)$ là tín hiệu nguồn đúng (true source image) của nguồn thứ j thu được tại microphone i , $e_{ij}^{spat}(t)$, $e_{ij}^{inter}(t)$, và $e_{ij}^{artif}(t)$ là các thành phần lỗi riêng biệt đại diện cho sự ảnh hưởng của không gian truyền âm (spatial or filtering), nhiễu và các sự can thiệp khác (interference and artifacts). Phân tách này xuất phát từ sự khác biệt giữa tín hiệu từ các nguồn mục tiêu, nhiễu từ các nguồn khác và nhiễu xuất hiện do các phép biến đổi và các thuật toán tương ứng với tín hiệu $c_{ij}(t)$, $e_{ij}^{spat}(t)$, $e_{ij}^{inter}(t)$, và $e_{ij}^{artif}(t)$.

Sự ảnh hưởng của không gian truyền được biểu diễn như là một bộ lọc của nguồn âm (the distortion filter), được tính bằng bình phương phép chiếu least-squares của nguồn ước lượng được (estimated source image) vào không gian tín hiệu tương ứng như sau [79]:

$$e_{ij}^{spat}(t) = P_j^L(\hat{c}_{ij})(t) - c_{ij}(t), \quad (4.2)$$

$$e_{ij}^{inter}(t) = P_{all}^L(\hat{c}_{ij})(t) - P_j^L(\hat{c}_{ij})(t), \quad (4.3)$$

$$e_{ij}^{artif}(t) = \hat{c}_{ij}(t) - P_{all}^L(\hat{c}_{ij})(t), \quad (4.4)$$

trong đó P_j^L là phép chiếu least-squares vào không gian con mở rộng bởi $c_{kj}(t-\tau)$, $1 \leq k \leq I$, $0 \leq \tau \leq L-1$, P_{all}^L là phép chiếu least-squares vào không gian con mở rộng bởi $c_{kl}(t-\tau)$, $1 \leq k \leq I$, $1 \leq l \leq J$, $0 \leq \tau \leq L-1$, và L là kích thước bộ lọc ($L = 32$ ms) [79].

Độ biến dạng của tín hiệu do không gian truyền âm, do sự tác động từ các nguồn âm khác và do các phép biến đổi của thuật toán được đo bằng đơn vị là dB qua ba tham số sau [79]:

- Tỷ lệ biến dạng bởi nhiễu giao thoa từ các nguồn âm khác SIR - Signal to Interference Ratio:

$$SIR = 10 \log_{10} \frac{\sum_{i=1}^I \sum_t (c_{ij}(t) + e_{ij}^{spat}(t))^2}{\sum_{i=1}^I \sum_t e_{ij}^{inter}(t)^2} \quad (4.5)$$

The suppression of interfering sources in the separation is objectified by this measure.

- Tỷ lệ các tín hiệu nhiễu sinh ra bởi các phép biến đổi của thuật toán SAR - Signal to Artifacts Ratio

$$SAR = 10 \log_{10} \frac{\sum_{i=1}^I \sum_t (c_{ij}(t) + e_{ij}^{spat}(t) + e_{ij}^{inter}(t))^2}{\sum_{i=1}^I \sum_t e_{ij}^{artif}(t)^2} \quad (4.6)$$

- Tỷ lệ biến dạng không gian, độ đo này được dùng cho các thuật toán khai thác các thông tin về không gian thu âm (spatial cues): ISR - Source Image to Spatial distortion Ratio:

$$ISR = 10 \log_{10} \frac{\sum_{i=1}^I \sum_t c_{ij}(t)^2}{\sum_{i=1}^I \sum_t e_{ij}^{spat}(t)^2} \quad (4.7)$$

Tổng các sai số được tính bằng tỷ lệ *Signal to Distortion Ratio* (SDR), đây là độ đo quan trọng nhất đánh giá tổng thể chất lượng tín hiệu tách. Công thức tính SDR như sau:

- Signal to Distortion Ratio:

$$SDR = 10 \log_{10} \frac{\sum_{i=1}^I \sum_t c_{ij}(t)^2}{\sum_{i=1}^I \sum_t (e_{ij}^{spat}(t) + e_{ij}^{inter}(t) + e_{ij}^{artif}(t))^2} \quad (4.8)$$

Bộ tool tính các tham số nói trên đã được xây dựng bằng Matlab và được công bố cho phép sử dụng rộng rãi miễn phí [17]¹. Chúng là bộ tiêu chí đánh giá kết quả tách nguồn âm thanh được sử dụng phổ biến nhất trong cộng đồng nghiên cứu hiện nay.

4.2 Dữ liệu thử nghiệm

Để xác thực hiệu quả của các thuật toán phân tách tiếng nói đề xuất, chúng tôi sử dụng ba bộ dữ liệu âm thanh từ các nguồn dữ liệu được công bố là DEMAND (Diverse Environments Multichannel Acoustic Noise Database)² và SiSEC (International Signal Separation and Evaluation Campaign)³. Những bộ dữ liệu này được thiết kế chuyên nghiệp và được sử dụng phổ biến trong cộng đồng nghiên cứu về xử lý âm thanh.

Bộ dữ liệu đầu tiên gồm các tín hiệu được tạo bằng cách trộn tiếng nói và tiếng ồn từ những file âm thanh riêng rẽ đã được thu âm trước đó. Hai bộ dữ liệu còn lại là tín hiệu được thu âm trong môi trường thực có độ phản xạ âm cao. Chi tiết về các bộ dữ liệu được mô tả như sau:

4.2.1 Bộ dữ liệu Synthetic

- Dữ liệu huấn luyện: Chúng tôi dùng 3 file tiếng nói (speech) gồm hai giọng nữ và một giọng nam⁴, mỗi file dài 10 giây để học đặc trưng của tiếng nói. Tập dữ liệu huấn luyện cho tiếng ồn (noise) gồm ba file tiếng ồn từ các môi trường khác nhau: trong nhà bếp, siêu thị, ngoài cánh đồng⁵, mỗi file dài từ 5 đến 10 giây.
- Dữ liệu thử nghiệm: Tập dữ liệu thử nghiệm gồm 12 file là tín hiệu trộn của speech và noise có tỷ lệ tín hiệu/nhiều $SNR = 0dB$, kích thước mỗi file là 10

¹http://bass-db.gforge.inria.fr/bss_eval/

²<http://parole.loria.fr/DEMAND/>

³<http://sisec.wiki.irisa.fr>.

⁴Speech files are from the International Signal Separation and Evaluation Campaign (SiSEC): <http://sisec.wiki.irisa.fr>.

⁵Some noise files are from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND): <http://parole.loria.fr/DEMAND/>.

giây. Các giọng nói và âm thanh nhiễu trong tập thử nghiệm là khác nhau và không trùng lặp với tập dữ liệu huấn luyện. Một số file thử nghiệm có sự pha trộn của 2 âm thanh nhiễu khác nhau như: traffic + wind sound, ocean waves + birdsong, restaurant + guitar, forest birds + car, square + music,... Các file âm thanh được lấy mẫu theo tần số 16000 Hz.

4.2.1.1 Bộ dữ liệu SiSEC-MUS

- Dữ liệu huấn luyện: Chúng tôi dùng 3 tập huấn luyện cho giọng hát (vocal) và âm nhạc (music). Giọng hát được huấn luyện bởi 3 file khác nhau có kích thước 10 giây, gồm 1 giọng nam và 2 giọng nữ ⁶. Dữ liệu huấn luyện âm nhạc gồm 3 file nhạc bass, 3 file tiếng trống (drums) và 3 file các nhạc cụ khác (guitar, piano, trumpet), mỗi file có độ dài từ 10 đến 15 giây.
- Dữ liệu thử nghiệm: Gồm 5 đoạn bài hát được mô tả trong Bảng 4.1 ⁷.

Bảng 4.1: Dữ liệu thử nghiệm SiSEC-MUS.

TT.	Tên bài hát	Độ dài (giây)
1	Bearlin - Roads	14
2	Tamy - Que pena Tanto faz	15
3	Another dreamer - The ones we love	25
4	Fort Minor - Remember the name	25
5	Ultimate nz tour	19

4.2.2 Bộ dữ liệu SiSEC-BNG

- Dữ liệu huấn luyện: gồm các file tiếng nói và nhiễu như mô tả trong phần 4.2.1.
- Dữ liệu thử nghiệm: Để thử nghiệm thuật toán, chúng tôi sử dụng bộ dữ liệu chuẩn của SiSEC 2016 gồm các file âm thanh được thu bởi 2 microphone, là hỗn hợp của tiếng nói và tiếng ồn môi trường trong thế giới thực (Two-channel mixtures of speech and real-world background noise - BGN) ⁸ [44]. Bộ dữ liệu

⁶Voice examples are from the International Signal Separation and Evaluation Campaign (SiSEC): <https://sisec.inria.fr/sisec-2016/2016-professionally-produced-music-recordings/>.

⁷<https://sisec.inria.fr/sisec-2016/2016-professionally-produced-music-recordings/>.

⁸<https://sisec.inria.fr/sisec-2016/bgn-2016/>

này chứa 29 hỗn hợp âm thanh có độ dài mỗi file là 10 giây, tần số lấy mẫu 16 kHz. Chúng là hỗn hợp của các giọng nam hoặc nữ và tiếng ồn từ các môi trường công cộng khác nhau như: trong quán cà phê (Ca), tại quảng trường (Sq), tại ga tàu điện (Su), ... Lưu ý rằng các môi trường ghi âm có độ vang âm khác nhau, chẳng hạn hỗn hợp ghi âm trong quán cà phê và ga tàu điện ngầm có độ vang cao hơn hỗn hợp thu âm ở quảng trường. Tỷ lệ tín hiệu trên nhiễu (Signal to Noise Ratio - SNR) của các file âm thanh biến đổi trong khoảng từ -17 đến +12 dB. Các file âm thanh được chia thành hai tập:

- *Devset* bao gồm 9 file : 03 file với nhiễu Ca, 04 file nhiễu Sq và 02 file nhiễu Su.
- *Testset* chứa 20 file: 08 file với nhiễu Ca, 08 file nhiễu Sq và 04 file với nhiễu Su.

4.3 Kết quả thực nghiệm và đánh giá

Trong phần này nhóm nghiên cứu thực hiện hai thí nghiệm với hai thuật toán Algorithm 3 và Algorithm 4 đã đề xuất trong chương 3 (phần 3.5) theo công thức ràng buộc thừa (3.8).

4.3.1 Kết quả thử nghiệm thuật toán Weakly-informed

4.3.1.1 So sánh với các thuật toán cơ sở

Kết quả thí nghiệm thuật toán đề xuất Algorithm 3 được so sánh với thuật toán gốc Algorithm 1 và hai thuật toán kết hợp NMF và các nhóm ràng buộc thừa được công bố gần đây là block sparsity (được gọi là thuật toán "**NMF-Block sparsity**") và component sparsity (được gọi là thuật toán "**NMF-Component sparsity**") [11, 63] trên hai bộ dữ liệu Synthetic và SiSEC-MUS.

- Baseline NMF - without training: Thuật toán Algorithm 2 được mô tả trong Chương 2 (Phần 2.3.2) khi không dùng dữ liệu huấn luyện. Ma trận phổ cơ sở của speech/vocal và noise/music được khởi tạo ngẫu nhiên và cập nhật trong từng bước lặp bằng công thức (2.8) và (2.9).
- NMF non-sparsity: Thuật toán Algorithm 3 được mô tả trong phần 3.5 nhưng không kết hợp điều kiện ràng buộc thừa trong quá trình cập nhật các tham số

NMF. Dữ liệu huấn luyện cho speech/vocal và noise/music được dùng như mô tả trong bộ dữ liệu tương ứng.

- NMF - Block sparsity: Thuật toán kết hợp NMF với ràng buộc thưa block sparsity theo công thức (3.6), thuật toán được công bố trong bài báo [68].
- NMF - Component sparsity: Là thuật toán kết hợp NMF và ràng buộc thưa component sparsity theo công thức (3.7) được công bố năm 2015 [11].

Bảng 4.2: Kết quả tách nguồn trên bộ dữ liệu Synthetic và SiSEC-MUS.

Bộ dữ liệu	Thuật toán	Speech/Vocal			Noise/Music		
		SDR	SIR	SAR	SDR	SIR	SAR
Synthetic	Baseline NMF-without training	-0.5	2.4	6.4	0.2	5.5	6.7
	NMF non-sparsity	2.7	6.9	11.7	3.6	14.3	5.2
	NMF-Block sparsity[68] ($\lambda = 25$)	7.4	10.2	16.4	6.9	19.8	8.5
	NMF-Component sparsity[11] ($\lambda = 50$)	7.4	10.9	16.2	7.6	16.3	9.3
	Proposed ($\lambda = 50, \gamma = 0.4$)	7.7	10.8	17.8	7.8	18.7	9.4
SiSEC-MUS	Baseline NMF-without training	-2.3	2.0	5.9	-1.2	4.4	7.8
	NMF non-sparsity	1.3	3.7	7.1	3.8	9.5	11.2
	NMF-Block sparsity[68] ($\lambda = 50$)	2.5	4.9	8.1	6.2	7.7	13.3
	NMF-Component sparsity[11] ($\lambda = 25$)	2.7	5.6	7.3	6.2	7.7	13.5
	Proposed-Weakly informed ($\lambda = 50, \gamma = 0.4$)	3.2	6.2	7.9	6.4	7.9	14.2

Kết quả trong Bảng 4.2 cho thấy:

- Thứ nhất, kết quả của thuật toán "Baseline NMF-without training" là thấp nhất. Điều này cho thấy thuật toán tách nguồn dựa trên mô hình NMF cơ bản cho kết quả rất thấp trong trường hợp không có dữ liệu huấn luyện.

- Thứ hai, các kết quả thu được từ ba thuật toán dựa trên NMF với các ràng buộc thưa tốt hơn nhiều so với thuật toán "NMF non-sparsity". Điều này thể hiện sự dư thừa

trong ma trận từ điển phổ tổng quát khi được xây dựng từ các mẫu cùng loại. Qua đó, vai trò của thành phần ràng buộc thưa được khẳng định một cách rõ ràng.

- Cuối cùng, thuật toán đề xuất cho kết quả phân tách tốt nhất ví các độ đo SDR, SIR và SAR, trên cả 2 bộ dữ liệu. Cụ thể hơn, so với hai thuật toán "NMF-Block sparsity" và "NMF-Component sparsity", thuật toán đề xuất cho kết quả tách tiếng nói cao hơn 0,3 dB SDR trên bộ dữ liệu Synthetic. Tăng lần lượt 0,7 dB và 0,5 dB SDR so với hai thuật toán đó trên bộ dữ liệu SiSEC-MUS. Điều này xác thực tính hiệu quả của sự kết hợp hai ràng buộc thưa mà nhóm nghiên cứu đã đề xuất.

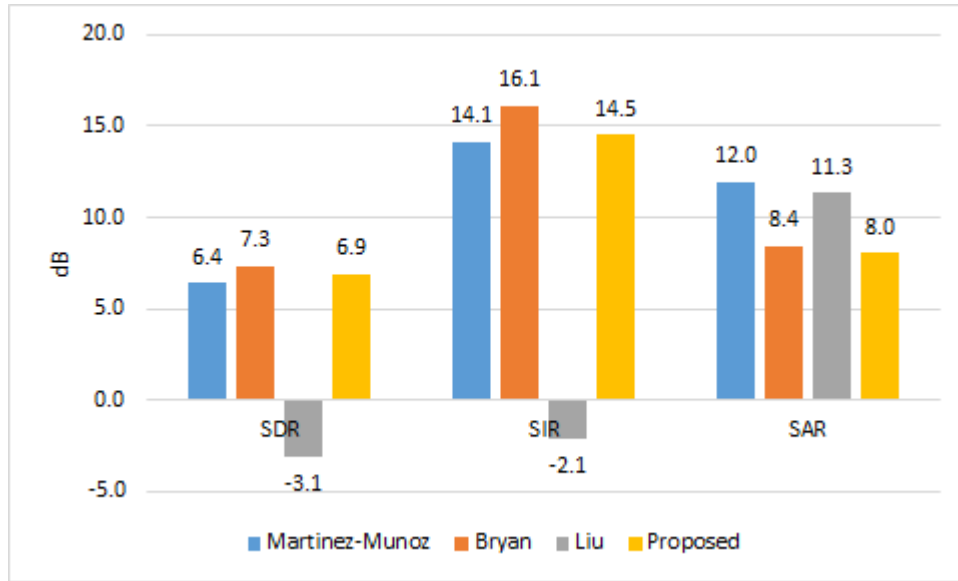
4.3.1.2 Mở rộng so sánh với các thuật tách nguồn âm thanh khác

Bên cạnh đó, kết quả của thuật toán đề xuất Algorithm 3 còn được mở rộng so sánh với các thuật toán cùng loại đã được công bố trên thế giới trong những năm gần đây trên bộ dữ liệu SiSEC-BGN:

- Thuật toán của Martinez-Munoz's [54]: Thuật toán được công bố năm 2013, sử dụng kết hợp NMF với mô hình lọc thích nghi cho từng loại tín hiệu tiếng nói và nhiễu.
- Thuật toán của Bryan's [5]: công bố năm 2013, thuật toán sử dụng mô hình phân tích các thành phần xác suất tiềm ẩn (PLCA) đồng thời sử dụng chuyên gia chú thích trực tiếp lên hình ảnh quang phổ của tín hiệu để đánh dấu vị trí xuất hiện tiếng nói, qua đó hướng dẫn bước phân tách (Phương pháp được mô tả trong Hình 1.8).
- Thuật toán của López's [45]: công bố năm 2015, thuật toán này sử dụng phép trừ phổ, thiết kế ma trận phân tách và bộ lọc tín hiệu dựa trên phương pháp tách nguồn đơn kênh.
- Thuật toán của Liu's [44]: công bố năm 2016, sử dụng kỹ thuật phân cụm theo thời gian đến microphone khác nhau của các nguồn âm (TDOA) và phép biến đổi pha tương quan chéo tổng quát (GCC-PHAT).

Bảng 4.3: Kết quả tách tiếng nói trên bộ dữ liệu SiSEC-BGN.

Thuật toán		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* [5] (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-Weakly informed	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

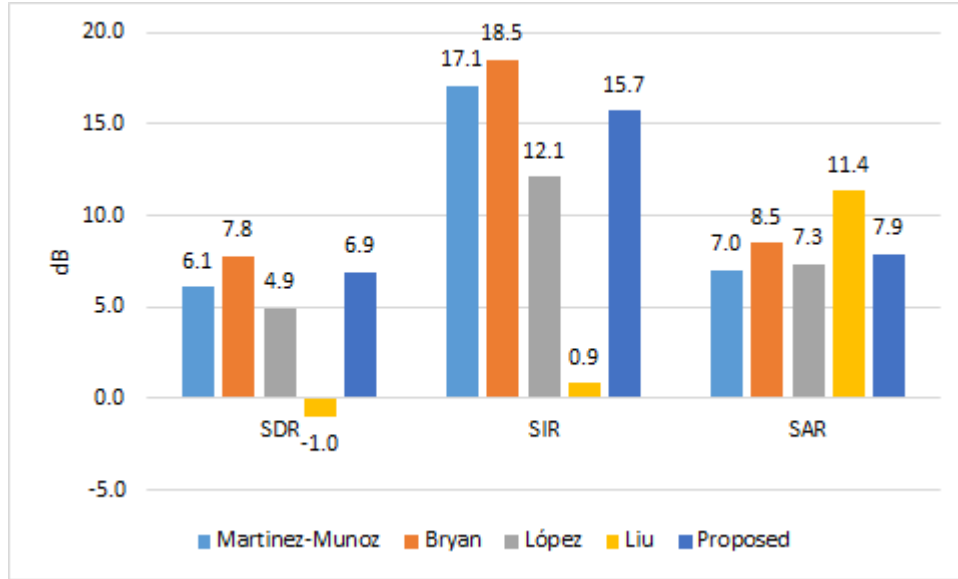


Hình 4.1: Kết quả phân tách tiếng nói trên bộ dữ liệu SiSEC-BGN-devset.

Bảng 4.3, Hình 4.1 và Hình 4.2 là kết quả phân tách tiếng nói của thuật toán đề xuất trên bộ dữ liệu SiSEC-BGN (gồm cả devset và testset) được so sánh với kết quả của các thuật toán đã tham gia SiSEC từ năm 2013 đến nay. Lưu ý rằng kết quả của những thuật toán * được công bố bởi chính tác giả của những thuật toán đó trong [44, 53, 55]. Kết quả cho thấy: thuật toán đề xuất cho kết quả kém hơn thuật toán của Bryan nhưng lại vượt trội hơn so với các thuật toán còn lại trên độ đo quan trọng nhất SDR. Tuy nhiên, thuật toán của Bryan sử dụng chú thích của chuyên gia trên hình ảnh phổ của tín hiệu để hướng dẫn và tinh chỉnh các bước tách nguồn, vì vậy thuật toán này sẽ không đạt được hiệu quả cao nếu không có sự tham gia của một chuyên gia xử lý âm thanh. Kết quả cho thấy giải pháp đề xuất sử dụng mô hình phổ tổng quát GSSM, mô hình NMF và công thức ràng buộc thừa kết hợp đã cho phép phân tách tín hiệu âm thanh đạt kết quả tốt hơn những thuật toán trước đó trong điều kiện không có dữ liệu huấn luyện.

4.3.1.3 Khảo sát sự hội tụ và tính ổn định của thuật toán

Hình 4.3 cho thấy cả ba độ đo SDR, SIR và SAR đều tăng dần khi số vòng lặp tăng và đạt giá trị bão hòa sau khoảng 20 vòng lặp MU. Kết quả phân tích này cho thấy thuật toán đề xuất hội tụ nhanh nên chi phí thời gian chạy thấp. Nhóm nghiên cứu cũng tính toán thời gian chạy thử nghiệm, kết quả cho thấy thời gian tách tín hiệu trộn có kích thước 10 giây là 1.8 giây, và cần 15 giây để tách file tín hiệu dài 60 giây. Đây là kết quả tương đối tốt để có phát triển tích hợp thuật toán trong các hệ thống thời gian thực hay online.



Hình 4.2: Kết quả phân tách tiếng nói trên bộ dữ liệu SiSEC-BGN-testset.

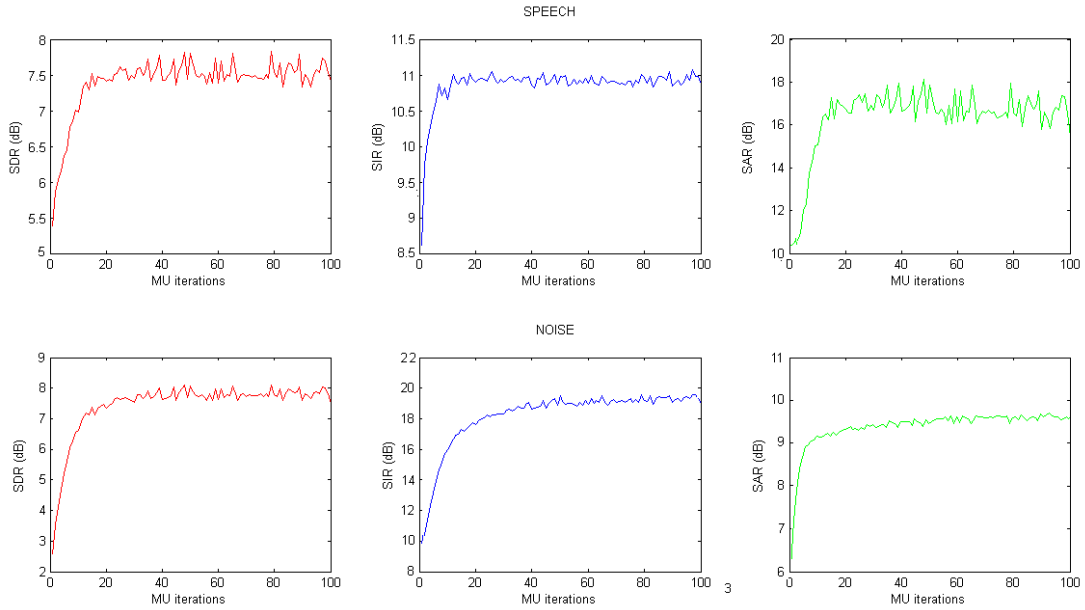
Khảo sát mức độ ảnh hưởng của hai tham số λ và γ đối với kết quả phân tách, Hình 4.4 cho thấy: sự biến đổi chất lượng tín hiệu đầu ra khi thay đổi λ và γ là tương đối điều hòa. Xác định được một vùng giá trị cho kết quả tốt là $10 \leq \lambda \leq 50$ và $0.2 \leq \gamma \leq 0.5$, và thuật toán đạt kết quả tốt nhất với $\lambda = 50$ và $\gamma = 0.4$.

4.3.2 Kết quả thử nghiệm thuật toán Semi-Supervised

Thuật toán Algorithm 4 được xây dựng cho trường hợp có dữ liệu huấn luyện cho tín hiệu tiếng nói nhưng không có dữ liệu huấn luyện cho tiếng ồn. Do đó chúng tôi dùng 01 file tiếng nói có độ dài 5 giây để huấn luyện mô hình phổ cho tiếng nói. Mô hình phổ tổng quát của tín hiệu nhiễu được huấn luyện bằng 3 file nhiễu môi trường như mô tả trong phần 4.2.1.

Kết quả phân tách tiếng nói của thuật toán đề xuất được so sánh với các thuật toán dựa trên mô hình NMF cơ bản và hai thuật toán được công bố trước đó dùng ràng buộc thưa block sparsity và component sparsity:

- Baseline NMF - speech training: Thuật toán Algorithm 2 được mô tả trong Chương 2 (Phần 2.3.2). Mô hình phổ của tín hiệu speech được huấn luyện bằng 01 file tiếng nói có độ dài 5 giây. Ma trận phổ tổng quát của noise được khởi tạo ngẫu nhiên và cập nhật trong từng bước lặp bằng công thức (2.8) và (2.9).
- NMF - Block sparsity: Thuật toán kết hợp NMF với ràng buộc thưa block sparsity theo công thức (3.6), thuật toán được công bố trong bài báo [68].



Hình 4.3: Khảo sát sự hội tụ của thuật toán theo số vòng lặp MU.

- NMF - Component sparsity: Là thuật toán kết hợp NMF và ràng buộc thưa component sparsity theo công thức (3.7) được công bố năm 2015 [11].

Kết quả phân tách tiếng nói trên bộ dữ liệu Synthetic của thuật toán đề xuất và 3 thuật toán trước đó được thể hiện trong Bảng 4.4.

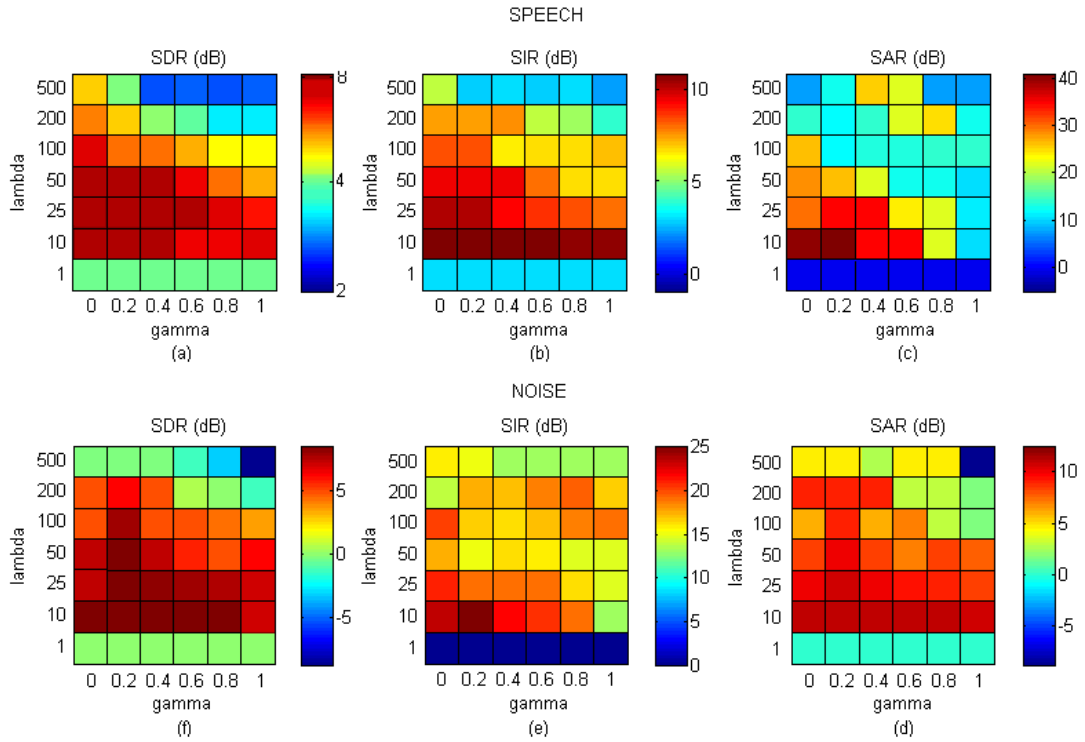
Bảng 4.4: Kết quả phân tách tiếng nói trên bộ dữ liệu Synthetic với trường hợp Semisupervised.

Thuật toán	SDR (dB)	SIR (dB)	SAR (dB)
Baseline NMF - speech training	5.8	9.6	10.4
NMF - Block sparsity [68]	7.9	12.7	11.4
NMF - Component sparsity [11]	8.0	13.1	11.4
Proposed-Semisupervised	8.3	13.3	11.5

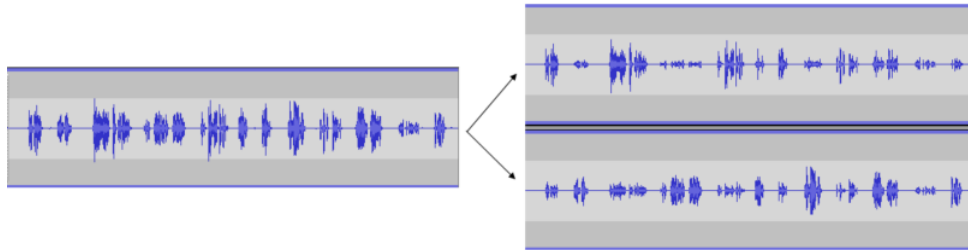
Kết quả mô tả trong bảng 4.4 cho thấy:

- Mặc dù có dữ liệu huấn luyện cho speech nhưng kết quả của thuật toán "Baseline NMF-speech training" vẫn thấp hơn kết quả của ba thuật toán sử dụng ràng buộc thưa trong quá trình phân tách. Điều này khẳng định rõ hiệu quả của nhóm ràng buộc thưa khi kết hợp với mô hình phân tách NMF.

- Kết quả của thuật toán "Proposed-Semisupervised" là tốt nhất trên cả ba độ đo SDR, SIR, và SAR. Cụ thể hơn, thuật toán đề xuất cho kết quả cao hơn lần lượt 0.4



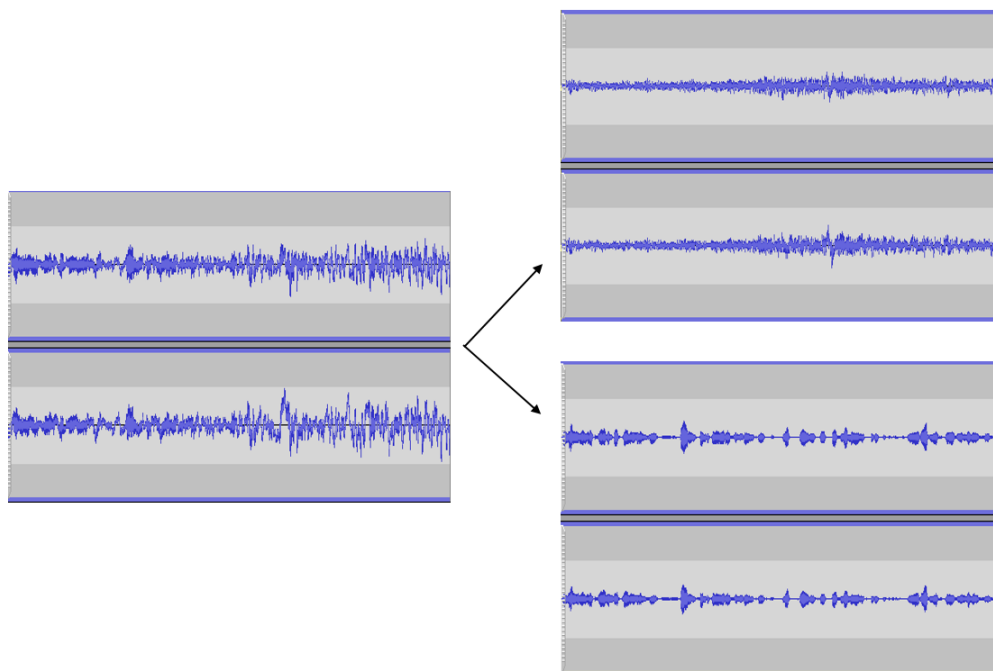
Hình 4.4: Khảo sát sự ảnh hưởng của hai tham số λ và γ với kết quả phân tách.



Hình 4.5: Hình ảnh phổ tín hiệu trước và sau khi tách nguồn đơn kênh.

dB và 0.3 dB SDR so với hai thuật toán "NMF-Block sparsity" và "NMF-Component sparsity". Điều này cho thấy thuật toán đề xuất thể hiện được ưu thế trong cả trường hợp Semisupervised, khi có dữ liệu huấn luyện cho một trong số các âm thanh cần tách.

Hình 4.5 và 4.6 là hình ảnh phổ của tín hiệu trộn và các tín hiệu tách của thuật toán đề xuất cho hai trường hợp tách nguồn đơn kênh và đa kênh. Tín hiệu trộn đơn kênh là giọng của 2 người nói (speech to speech separation). Tín hiệu đa kênh là tiếng nói thu âm tại ga tàu điện ngầm, âm thanh nhiễu là tiếng tàu vào ga.



Hình 4.6: Hình ảnh phổ tín hiệu trước và sau khi tách nguồn đa kênh.

4.4 Kết luận

Chương này đã giới thiệu bộ tiêu chí đánh giá kết quả phân tách âm thanh mà nhóm nghiên cứu sử dụng để đánh giá kết quả thuật toán đề xuất. Ngoài ra, cách xây dựng bộ dữ liệu thử nghiệm cho bài toán cũng được mô tả với các tham số chi tiết. Kết quả thí nghiệm và các đánh giá, so sánh đã cho thấy hiệu quả của thuật toán đề xuất là tốt hơn so với các thuật toán cùng loại khi thực hiện với cùng bộ dữ liệu thử nghiệm.

KẾT LUẬN VÀ KIẾN NGHỊ

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 đã đề xuất thuật toán phân tách các nguồn âm từ tín hiệu thu âm gồm nhiều nguồn âm thanh bị trộn lẫn (mixture) theo hướng tiếp cận sử dụng thông tin hướng dẫn ít (weakly-informed approach). Thuật toán sử dụng mô hình thừa số hóa ma trận không âm (Nonnegative Matrix Factorization - NMF) để ước lượng đặc trưng phổ của các nguồn âm cần tách từ ma trận phổ tổng quát GSSM, trong đó GSSM được huấn luyện từ một vài ví dụ mẫu cùng loại với âm thanh cần phân tách. Đề tài đề xuất một công thức ràng buộc thừa mới cho bước ước lượng các tham số của NMF. Hàm ràng buộc thừa mới đề xuất Mixed sparsity là sự kết hợp tuyến tính của hai hàm ràng buộc thừa đã được công bố trước đó là Block sparsity và Component sparsity. Chúng tôi tính toán công thức cập nhật tham số theo hàm ràng buộc thừa mới đề xuất và xây dựng thuật toán tách nguồn đơn kênh cho hai trường hợp: (1) Weakly-informed: không có dữ liệu huấn luyện cho các nguồn cần tách và (2) Semi-supervised: có dữ liệu huấn luyện cho tín hiệu tiếng nói nhưng không xác định nhiều.

Hiệu suất phân tách của các thuật toán đề xuất cũng như khả năng hội tụ và tính ổn định của thuật toán được kiểm chứng bằng kết quả thí nghiệm trên bộ dữ liệu SiSEC (International Signal Separation Evaluation Campaign), là bộ dữ liệu được công bố và sử dụng phổ biến trong cộng đồng xử lý âm thanh. Kết quả thí nghiệm cũng đã được đánh giá, so sánh với các nghiên cứu khác trên thế giới và cho thấy hiệu quả cải thiện chất lượng tín hiệu tách của thuật toán đề xuất.

Kiến nghị

Từ kết quả nghiên cứu hiện nay, chúng tôi mong muốn được tiếp tục mở rộng nghiên cứu theo các nội dung sau đây:

- 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 [52].
- 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 với kỹ thuật phát hiện đối tượng ảnh, xây dựng hệ thống phát hiện và phân tách đồng thời âm thanh, hình ảnh từ dữ liệu video. Dùng thông tin hình ảnh để hướng dẫn phân tách âm thanh và ngược lại, dùng thông tin về âm thanh phân tách được để phát hiện, phân tách và gán nhãn đối tượng ảnh.
- 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] Ainhoren, Y., Engelberg, S., and Friedman, S. (2008). The Cocktail Party Problem. *IEEE Instrumentation Measurement Magazine*, 11(3):44–48.
- [2] Araki, S., Nesta, F., Vincent, E., Koldovský, Z., Nolte, G., Ziehe, A., and Benichoux, A. (2012). The 2011 Signal Separation Evaluation Campaign (SiSEC2011): Audio Source Separation. In *Latent Variable Analysis and Signal Separation*, pages 414–422, Berlin, Heidelberg. Springer Berlin Heidelberg.
- [3] Barkera, J., Vincent, E., Maa, N., Christensena, H., and Greena, P. (2013). The pascal chime speech separation and recognition challenge. *Computer Speech and Language*, 27(3):621–633.
- [4] Benesty, J., Makino, S., and Chen, J. (2005). *Speech Enhancement*. Springer.
- [5] Bryan, N. and Mysore, G. (2013). An Efficient Posterior Regularized Latent Variable Model for Interactive Sound Source Separation. In *Proc. The 30th International Conference on Machine Learning (ICML)*, pages 208–216.
- [6] C. Blandin, A. O. and Vincent, E. (1989). Multi-source tdoa estimation in reverberant audio using angular spectra and clustering. *IEEE Signal Processing*, 92:1950–1960.
- [7] Cardoso, J.-F. (1989). Blind signal separation: Statistical principles. *IEEE Signal Processing Letters*, 86:2009–2025.
- [8] Chen, L., Ma, X., and Ding, S. (2017). Single Channel Speech Separation Using Deep Neural Network. In Cong, F., Leung, A., and Wei, Q., editors, *Advances in Neural Networks - ISNN 2017*, volume 10261, pages 285–292. Springer International Publishing, Cham.
- [9] Cohen, I. and Berdugo, B. (2001). Speech enhancement for non-stationary noise environments. *Signal Processing*, 81(11):2403 – 2418.
- [10] Duong, H.-T. T., Nguyen, Q.-C., Nguyen, C.-P., Tran, T.-H., and Duong, N. Q. K. (2015). Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint. pages 1–5. ACM Press.

- [11] El Badawy, D., Ozerov, A., and Duong, N. Q. K. (2015). Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*.
- [12] Emiya, V., Vincent, E., Harlander, N., and Hohmann, V. (2011). Subjective and Objective Quality Assessment of Audio Source Separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:2046–2057.
- [13] Essid, S. and Ozerov, A. (2014). A Tutorial on Nonnegative Matrix Factorisation with Applications to Audiovisual Content Analysis. In *IEEE International Conference on Multimedia and Expo (ICME)*.
- [14] Fan, H.-T., Hung, J.-w., Lu, X., Wang, S.-S., and Tsao, Y. (2014). Speech Enhancement Using Segmental Nonnegative Matrix Factorization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4483–4487.
- [15] 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.
- [16] Févotte, C., Bertin, N., and Durrieu, J. L. (2009). Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis. *Neural Computation*, 21(3):793–830.
- [17] Févotte, C., Gribonval, R., and Vincent, E. (2005). BSS_EVAL Toolbox User Guide – Revision 2.0. Technical report. Developed with the support of the French GdR-ISIS/CNRS Workgroup “Resources for Audio Source Separation”.
- [18] Févotte, C. and Idier, J. (2011). Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456.
- [19] Févotte, C., Vincent, E., and Ozerov, A. (2017). Single-channel Audio Source Separation with NMF: Divergences, Constraints and Algorithms. In *Audio Source Separation*. Springer.
- [20] Fox, B., Sabin, A., Pardo, B., and Zopf, A. (2007). Modeling Perceptual Similarity of Audio Signals for Blind Source Separation Evaluation. In *Independent Component Analysis and Signal Separation - 7th International Conference, ICA*

- 2007, *Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 454–461.
- [21] Gribonval, R., Vincent, E., Févotte, C., and Benaroya, L. (2003). Proposals for Performance Measurement in Source Separation. In *4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 763–768.
 - [22] Gustafsson, T., Rao, B., and Trivedi, M. (2003). Source Localization in Reverberant Environments: Modeling and Statistical Analysis. *IEEE Transactions on Speech and Audio Processing*, 11(6):791–803.
 - [23] Houda, A. and Otman, C. (2015). Article: Blind Audio Source Separation: State-of-Art. *International Journal of Computer Applications*, 130(4):1–6. Published by Foundation of Computer Science (FCS), NY, USA.
 - [24] Huang, A. (2013). NMF Face Recognition Method Based on Alpha Divergence. In Zhong, Z., editor, *Proceedings of the International Conference on Information Engineering and Applications (IEA) 2012*, volume 217, pages 477–483. Springer London.
 - [25] Hurmalainen, A., Saeidi, R., and Virtanen, T. (2012). Group sparsity for speaker identity discrimination in factorisation-based speech recognition. In *Interspeech*, pages 17–20.
 - [26] Ito, N., Araki, S., and Nakatani, T. (2013). Permutation-free Convolutional Blind Source Separation via Full-band Clustering Based on Frequency-independent Source Presence Priors. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3238–3242.
 - [27] Ito, N., Ono, N., Vincent, E., and Sagayama, S. (2010). Designing the wiener post-filter for diffuse noise suppression using imaginary parts of inter-channel cross-spectra. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2818–2821.
 - [28] Jeter, M. and Pye, W. (1981). A Note on Nonnegative Rank Factorizations. *Linear Algebra and its Applications*, 38:171–173.
 - [29] Joachim Thiemann, E. V. (2013). An experimental comparison of source separation and beamforming techniques for microphone array signal enhancement. In

- Proc. The 23rd IEEE International Work-shop on Machine Learning for Signal Processing (MLSP)*, pages 1–4.
- [30] Kim, G. and Loizou, P. (2010). Improving speech intelligibility in noise using environment-optimized algorithms. *IEEE Trans. Audio, Speech, Language Processing*, 18(8):2080–2090.
 - [31] Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., and Maas, R. (2013). The Reverb Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4, NY, USA.
 - [32] Kitamura, D., Ono, N., Sawada, H., Kameoka, H., and Saruwatari, H. (2016a). Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization. *IEEE/ACM Trans. on Audio, Speech and Language Processing*, 24(9):1622–1637.
 - [33] Kitamura, D., Ono, N., Sawada, H., Kameoka, H., and Saruwatari, H. (2016b). Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1626–1641.
 - [34] Kleinschmidt, T. F. (2010). *Robust speech recognition using speech enhancement*. Ph.D. thesis, Queensland University of Technology.
 - [35] Le Magoarou, L., Ozerov, A., and Duong, N. Q. K. (2013). Text-informed Audio Source Separation using Nonnegative Matrix Partial Co-factorization. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.
 - [36] Lee, D. D. and Seung, H. S. (1999). Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature*, 401 6755:788–91.
 - [37] 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.
 - [38] 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.

- [39] Li, Y. and Ngom, A. (2013). The Non-negative Matrix Factorization Toolbox for Biological Data Mining. *Source Code for Biology and Medicine*, 8(1):1–10.
- [40] Linh-Trung, N., Aissa-El-Bey, A., Abel-Meraim, K., and Belouchrani, A. (2005). Underdetermined Blind Source Separation of Non-disjoint Nonstationary Sources in the Time-Frequency Domain. In *Proceedings of the Eighth International Symposium on Signal Processing and Its Applications, 2005.*, volume 1, pages 46–49.
- [41] Liutkus, A., Badeau, R., and Richard, G. (2011). Gaussian Processes for Underdetermined Source Separation. *IEEE Transactions on Signal Processing*, 59(7):3155–3167.
- [42] Liutkus, A., Durrieu, J. L., Daudet, L., and Richard, G. (2013). An Overview of Informed Audio Source Separation. In *Proc. IEEE Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4.
- [43] Liutkus, A., Stöter, F. R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., and Fontecave, J. (2017a). The 2016 Signal Separation Evaluation Campaign. In *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*, pages 323–332.
- [44] Liutkus, A., Stöter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., and Fontecave, J. (2017b). The 2016 Signal Separation Evaluation Campaign. In *Latent Variable Analysis and Signal Separation*, volume 10169, pages 323–332. Springer International Publishing, Cham.
- [45] López, A. R., Ono, N., Remes, U., Palomäki, K., and Kurimo, M. (2015). Designing Multichannel Source Separation based on Single-channel Source Separation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 469–473.
- [46] Makino, S., Lee, T.-W., and Sawada, H. (2007). *Blind Speech Separation*. Springer.
- [47] McCowan, I. and Bourslard, H. (2003). Microphone Array Post-filter based on Noise Field Coherence. *IEEE Transactions on Speech and Audio Processing*, 11(6):709–716.
- [48] Mohammadiha, N., Smaragdis, P., and Leijon, A. (2013). Supervised and Un-supervised Speech Enhancement Using Nonnegative Matrix Factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2140–2151.

- [49] Naik, G. R. and Wang, W., editors (2014). *Blind Source Separation: Advances in Theory, Algorithms and Applications*. Signals and Communication Technology. Springer, Berlin.
- [50] Nikunen, J. and Virtanen, T. (1998). Multidimensional independent component analysis. pages 1941–1944.
- [51] Nikunen, J. and Virtanen, T. (2014). Direction of arrival based spatial covariance model for blind sound source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22:727–739.
- [52] 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.
- [53] Ono, N., Koldovský, Z., Miyabe, S., and Ito, N. (2013a). The 2013 Signal Separation Evaluation Campaign. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.
- [54] Ono, N., Koldovský, Z., Miyabe, S., and Ito, N. (2013b). The 2013 Signal Separation Evaluation Campaign. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.
- [55] Ono, N., Rafii, Z., Kitamura, D., Ito, N., and Liutkus, A. (2015a). The 2015 Signal Separation Evaluation Campaign. In *Latent Variable Analysis and Signal Separation*, volume 9237, pages 387–395. Springer International Publishing, Cham.
- [56] Ono, N., Rafii, Z., Kitamura, D., Ito, N., and Liutkus, A. (2015b). The 2015 Signal Separation Evaluation Campaign. In *Latent Variable Analysis and Signal Separation (LVAICA)*, volume 9237, pages 387–395. Springer.
- [57] Ozerov, A., Fevotte, C., Blouet, R., and Durrieu, J.-L. (2011). Multichannel Non-negative Tensor Factorization with Structured Constraints for User-guided Audio Source Separation. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 257–260.
- [58] Ozerov, A., Févotte, C., and Vincent, E. (2017). An Introduction to Multichannel NMF for Audio Source Separation. In *Audio Source Separation*, Signals and Communication Technology. Springer.

- [59] Ozerov, A., Vincent, E., and Bimbot, F. (2012). A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1118–1133.
- [60] Paatero, P. (1997). Least Squares Formulation of Robust Non-negative Factor Analysis. *Chemometrics and Intelligent Laboratory Systems*, 37(1):23–35.
- [61] Paatero, P. and Tapper, U. (1994). Positive Matrix Factorization: A Non-negative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics*, 5(2):111–126.
- [62] Quang, T. T., Huy, T. Q., and Phuong, N. H. (2011). Blind Source Separation applied to Sound in Various Conditions. *Journal of Science and Technology Development*, 14:34–42.
- [63] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41.
- [64] Sandler, R. and Lindenbaum, M. (2011). Nonnegative Matrix Factorization with Earth Mover’s Distance Metric for Image Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1590–1602.
- [65] Sawada, H., Araki, S., and Makino, S. (2011). Underdetermined Convolutional Blind Source Separation via Frequency Bin-wise Clustering and Permutation Alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):516–527.
- [66] Smaragdis, P., Raj, B., and Shashanka, M. (2007). Supervised and semi-supervised separation of sounds from single-channel mixtures. In *Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pages 414–421.
- [67] Sprechmann, P., Bronstein, A. M., and Sapiro, G. (2015). Supervised Non-negative Matrix Factorization for Audio Source Separation. In *Excursions in Harmonic Analysis, Volume 4*, pages 407–420. Springer International Publishing, Cham.
- [68] Sun, D. L. and Mysore, G. J. (2013). Universal speech models for speaker independent single channel source separation. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 141–145.

- [69] Thuy-Duong, N. T., Linh-Trung, N., Tran-Duc, T., and Boashash, B. (2013). Separation of Nonstationary EEG Epileptic Seizures Using Time-Frequency-Based Blind Signal Processing Techniques. In *Proceedings of the 4th International Conference on Biomedical Engineering in Vietnam*, pages 317–323.
- [70] Tien Dung Tran, Dang Khoa Nguyen, Q. C. N. and Nguyen, H. B. (2012). Binaural speech enhancement system combining dereverberation and spatial masking-based noise removal for robust speech recognition. In *Proc. IEEE Int. Conf. on Communications and Electronics (ICCE)*, page 345–350.
- [71] Tien Dung Tran, Dang Khoa Nguyen, T. A. X. T. Q. C. N. and Nguyen, H. B. (2011). Speech enhancement using combination of dereverberation and noise reduction for robust speech recognition. In *Proc. ACM Int. Sym. on Information and Communication Technology (SoICT)*, pages 151–154.
- [72] Vincent, E., Araki, S., and Bofill, P. (2009). The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation. In *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pages 734–741.
- [73] Vincent, E., Araki, S., Theis, F., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, V., Lutter, D., and Duong, N. Q. (2012a). The Signal Separation Evaluation Campaign (2007–2010): Achievements and Remaining Challenges. *Signal Processing*, 92(8):1928–1936.
- [74] Vincent, E., Araki, S., Theis, F., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, V., Lutter, D., and Duong, N. Q. K. (2012b). The Signal Separation Campaign (2007–2010): Achievements and remaining challenges. *Signal Processing*, 92:1928–1936.
- [75] Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., and Matassoni, M. (2013). The Second 'CHiME' Speech Separation and Recognition Challenge" Datasets, Tasks and Baselines. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 126–130.
- [76] Vincent, E., Bertin, N., Gribonval, R., and Bimbot, F. (2014). From Blind to Guided Audio Source Separation: How Models and Side Information can Improve the Separation of Sound. *IEEE Signal Processing Magazine*, 31(3):107–115.

- [77] Vincent, E., Gribonval, R., and Fevotte, C. (2006a). Performance Measurement in Blind Audio Source Separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469.
- [78] Vincent, E., Jafari, M. G., and Plumbley, M. D. (2006b). Preliminary Guidelines for Subjective Evaluation of Audio Source Separation Algorithms. In *UK ICA Research Network Workshop*, Southampton, United Kingdom.
- [79] Vincent, E., Sawada, H., Bofill, P., Makino, S., and Rosca, J. P. (2007). First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. In *Independent Component Analysis and Signal Separation*, pages 552–559. Springer Berlin Heidelberg.
- [80] Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparse criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3).
- [81] Vuong-Hoang, N., Nguyen-Quoc, T., and Tran-Hoai, L. (2010). Blind Speech Separation in Convolutional Mixtures using Non-Gaussianity Maximization and Inverse Filters. In *International Conference on Communications and Electronics 2010*, pages 190–194.
- [82] Wang, B. and Plumbley, M. D. (2006). Investigating single-channel audio source separation methods based on non-negative matrix factorization. In *ICA Research Network International Workshop*, pages 17–20.
- [83] Wang, D. (2017). Deep Learning Reinvents the Hearing Aid. *IEEE Spectrum*, 54(3):32–37.
- [84] Wang, D. and Brown, G. J., editors (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. IEEE Press ; Wiley Interscience.
- [85] Wang, Y.-X. and Zhang, Y.-J. (2013). Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353.
- [86] Wood, S. U. N., Rouat, J., Dupont, S., and Pironkov, G. (2017). Blind Speech Separation and Enhancement With GCC-NMF. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):745–755.

- [87] Zhang, Z.-Y. (2012). Nonnegative Matrix Factorization: Models, Algorithms and Applications. In *Data Mining: Foundations and Intelligent Paradigms*, volume 24, pages 99–134. Springer Berlin Heidelberg.

DANH SÁCH BÀI BÁO

1. **Thanh Thi Hien Duong**, Ngoc Q. K. Duong, Phuong Cong Nguyen, and Cuong Quoc Nguyen. *Multichannel source separation exploiting NMF-based generic source spectral model in Gaussian modeling framework*. In Latent Variable Analysis and Signal Separation, vol. 10891, pp. 547-557. Springer. Jul. 2018. ISSN 0302-9743, Electronic ISSN 1611-3349. DOI 10.1007/978-3-319-93764-9_50.
2. **Trần Thanh Huân**, Dương Thị Hiền Thanh, Nguyễn Thu Hằng. *Weakly-informed Audio Source Separation based on Nonnegative Matrix Factorization*. Hội nghị toàn quốc Khoa học trái đất và tài nguyên với phát triển bền vững (ERSD 2018). Hà Nội, 2018.