Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất
Trang điện tử: http://tapchi.humg.edu.vn

Nâng cao chất lượng tiếng nói từ tín hiệu thu âm đơn kênh chứa nhiễu môi trường ở mức cao

Dương Thị Hiền Thanh [1,*], Trần Thanh Huân [2], Nguyễn Thu Hằng [1], Phạm Quang Hiển [1], Vũ Thị Kim Liên [1]

[1] *Khoa Công nghệ Thông tin, Trường Đại học Mỏ - Địa chất, Việt Nam*
[2] *Trường Đại học Công nghiệp Hà Nội, Việt Nam*

THÔNG TIN BÀI BÁO

Quá trình:
Nhận bài 15/7/2017
Chấp nhận 20/8/2017
Đăng online 30/10/2017

Từ khóa:
Nâng cao chất lượng tiếng nói
Tách nguồn âm thanh
NMF
Mô hình phổ tổng quát
Ràng buộc thưa

TÓM TẮT

Trong lĩnh vực nghiên cứu về xử lý tiếng nói (Speech signal processing) hiện nay, vấn đề nâng cao chất lượng tiếng nói mong muốn trong điều kiện môi trường thu âm có nhiều tiếng ồn và nhiễu vẫn còn rất nhiều khó khăn thách thức, đặc biệt là đối với trường hợp thu âm đơn kênh (single-channel) và khi tín hiệu nhiễu nền ở mức cao. Tiếp cận theo hướng áp dụng kỹ thuật tách nguồn âm (source separation) để nâng cao chất lượng tín hiệu tiếng nói mong muốn, bài báo đề xuất giải pháp kết hợp mô hình thừa số hóa ma trận không âm (Nonnegative Matrix Factorization - NMF) với các ràng buộc thưa (sparsity constraint) để tách tín hiệu tiếng nói từ tín hiệu thu âm đơn kênh chứa nhiễu môi trường ở mức cao trong trường hợp không có dữ liệu huấn luyện cho tín hiệu cần tách. Thí nghiệm đã cho thấy thuật toán đề xuất cho kết quả tốt hơn so với các thuật toán được công bố trước đó.

1. Mở đầu

Trong môi trường thu âm thực tế, tiếng nói mong muốn (desired speech) thường bị trộn lẫn với nhiều âm thanh khác như tiếng ồn môi trường, tiếng nhạc, tiếng xe cộ, hay các giọng nói không mong muốn khác,... Con người với khả năng thính giác bình thường qua hai tai có thể dễ dàng định vị và phân tách nguồn âm thanh mong muốn để hiểu được các nội dung đối thoại. Tuy nhiên với máy học (machine) thì công việc này lại trở nên vô cùng khó khăn. Các đánh giá khoa học uy tín trong những năm gần đây (Emmanuel Vincent et al. 2013; Kinoshita et al. 2013; Liutkus et al. 2017) cũng đã cho thấy tỷ lệ nhận dạng thành công của các hệ thống nhận dạng giọng nói hoạt động trong trong môi trường thực tế có tiếng vọng và nhiễu vẫn còn rất thấp, đặc biệt là trong trường hợp thu âm đơn kênh. Vì vậy, nâng cao chất lượng tiếng nói (speech enhancement) là một vấn đề nghiên cứu quan trọng, đầy thách thức trong những năm gần đây và được ứng dụng rộng rãi trong thực tế cuộc sống như: tương tác người máy; thông tin liên lạc; truyền thông, truyền hình; xử lý âm thanh hậu kỳ trong giải trí,

*Tác giả liên hệ
E-mail: duongthihienthanh@humg.edu.vn

phim ảnh; các hệ thống y tế hỗ trợ theo dõi và chăm sóc người bệnh;… Đó là những kỹ thuật nhằm loại bỏ tiếng ồn môi trường và những âm thanh không mong muốn khác (gọi chung là nhiễu) và làm cho tiếng nói mong muốn trở nên "tốt" hơn (Benesty, Makino, and Chen 2005).

Một số phương pháp nâng cao tiếng nói đã được công bố gần đây (Cohen 2002; Gibak Kim and Loizou 2010) thực hiện ước lượng phổ của tín hiệu nhiễu, sau đó dùng phương pháp trừ phổ hoặc lọc tín hiệu (như Wiener filtering) để loại nhiễu. Một số nhóm nghiên cứu khác phát triển các giải thuật học có giám sát (supervised), bán giám sát (semi-supervised), hay các kỹ thuật học sâu (deep learning) (Sun and Mysore, 2013; Chen Ma and Ding, 2017) để nâng cao chất lượng tiếng nói. Các nghiên cứu trên đều sử dụng dữ liệu huấn luyện để học các đặc tính của tín hiệu tiếng nói và/hoặc tín hiệu nhiễu, sau đó dùng kết quả của bước học để lọc, tách tín hiệu tiếng nói mong muốn. Như vậy, trong trường hợp không có dữ liệu huấn luyện tốt thì những phương pháp này sẽ không thể áp dụng được.

Nhằm giải quyết bài toán nâng cao chất lượng tín hiệu tiếng nói trong trường hợp thu âm đơn kênh, không xác định được người nói là ai và âm thanh nhiễu môi trường ở dạng nào (không có dữ liệu huấn luyện), bài báo tiếp cận theo hướng sử dụng kỹ thuật tách các nguồn âm thanh bị trộn lẫn (audio source separation) với giả thiết coi tiếng nói mong muốn và âm thanh nhiễu là hai nguồn âm cần được tách rời. Cũng theo hướng tiếp cận này, công bố gần đây của Sun và nnk (Sun and Mysore 2013) đã đề xuất giải pháp sử dụng mô hình thừa số hóa ma trận không âm (Nonnegative Matrix Factorization - NMF) (Lee and Seung 2001) để xây dựng mô hình phổ tổng quát cho tín hiệu tiếng nói từ một số giọng nói khác. Nghiên cứu của El Badawy và nnk (El Badawy, Ozerov, and Duong 2015) sử dụng mô hình phổ tổng quát để tách một nguồn âm thanh bất kỳ với sự hướng dẫn từ các mẫu âm thanh cùng loại được thu thập từ một công cụ tìm kiếm (chẳng hạn như google search). Tiếp tục khai thác ý tưởng đó, trong bài báo này, chúng tôi đề xuất thuật toán nâng cao chất lượng tiếng nói theo hai bước:

- Xây dựng mô hình phổ tổng quát cho cả tiếng nói và nhiễu nền một cách độc lập từ một số file âm thanh mẫu cùng loại.

- Sử dụng kết hợp mô hình NMF với các ràng buộc thưa (sparsity constraint) để khai thác hai mô hình phổ tổng quát, hướng dẫn quá trình phân tách tiếng nói và nhiễu từ hỗn hợp mixture.

Trong thuật toán, chúng tôi xây dựng mô hình phổ tổng quát cho cả tín hiệu tiếng nói và nhiễu, tức là chỉ cần thu thập các tín hiệu cùng loại với tín hiệu cần tách để xây dựng mô hình phổ cho bước học mà không cần phải có dữ liệu huấn luyện chính xác. Ngoài ra chúng tôi sử dụng đồng thời hai nhóm ràng buộc thưa là block sparsity và component sparsity theo công thức chúng tôi đã công bố năm 2015 (Duong et al., 2015) kết hợp với mô hình NMF nhằm nâng cao hiệu quả ước lượng các tín hiệu cần tách.

2. Áp dụng mô hình NMF trong nâng cao chất lượng tiếng nói

Để nâng cao chất lượng tín hiệu tiếng nói mong muốn từ tín hiệu thu âm đơn kênh chứa nhiễu môi trường (gọi là mixture), chúng tôi coi mixture là tín hiệu bị trộn lẫn bởi hai nguồn âm: tiếng nói mong muốn (speech) và nhiễu (noise), trong đó noise có thể bao gồm tiếng ồn môi trường, tiếng vọng và các âm thanh không mong muốn khác. Mục đích của bài toán là phân tách hai tín hiệu speech và noise từ mixture ban đầu.

NMF là mô hình được dùng khá phổ biến trong lĩnh vực xử lý âm thanh nói chung và trong tách nguồn âm nói riêng (Smaragdis, Raj, and Shashanka 2007; Sun and Mysore 2013). Một cách tổng quát, tín hiệu âm thanh được biến đổi từ miền thời gian (time domain) sang miền thời gian-tần số (time-frequency domain) qua phép biến đổi Fourier (STFT). Sau quá trình xử lý, ước lượng các đặc trưng phổ và quá trình phân tách, tín hiệu lại được biến đổi về miền thời gian qua phép biến đổi Fourier ngược (ISTFT).

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

$$X = S + N \quad (1)$$

Gọi $V = |X|^{\cdot 2}$ là ma trận năng lượng phổ của tín hiệu mixture, với $|X|^{\cdot n}$ là ma trận có các phần tử là $[X]_{il}^{n}$, mô hình NMF sẽ phân tách ma trận không âm V kích thước $F \times M$ thành hai ma trận

không âm W ∈ $\mathbb{R}^{F\times K}$ và H ∈ $\mathbb{R}^{K\times M}m$ trong đó W là ma trận đặc trưng phổ (spectral matrix) và H là ma trận ma trận kích hoạt (time activations) của tín hiệu. K là số thành phần đặc trưng phổ của tín hiệu, thường được chọn nhỏ hơn F và M. Để ước lượng các ma trận W và H, hàm giá thể hiện độ sai khác giữa V và WH sẽ được cực tiểu hóa theo một độ đo thích hợp:

$$\min_{H\geq 0, W\geq 0} D(V\|WH), \quad (2)$$

Với $D(V\|WH) = \sum_{f=1}^{F}\sum_{m=1}^{M} d_{IS}\left(V_{fm}\|\widehat{V_{fm}}\right)$, $\hat{V} = WH$, f và m lần lượt là chỉ số tần số và chỉ số khung thời gian, $d_{IS}(x\|y) = \frac{x}{y} - log\frac{x}{y} - 1$ là độ đo IS-divergence được sử dụng phổ biến với dữ liệu âm thanh (Févotte, Bertin, and Durrieu 2009).

Các tham số W, H được khởi tạo giá trị không âm ngẫu nhiên và được cập nhật trong quá trình lặp theo quy tắc cập nhật nổi tiếng MU-rules (Févotte, Bertin, and Durrieu 2009) theo công thức:

$$H \leftarrow H\odot \frac{W^T((WH)^{\cdot(\beta-2)}\odot V)}{W^T(WH)^{\cdot(\beta-1)}}, \quad (3)$$

$$W \leftarrow W\odot \frac{((WH)^{\cdot(\beta-2)}\odot V)H^T}{(WH)^{\cdot(\beta-1)}H^T}, \quad (4)$$

với A^T là ma trận chuyển vị của ma trận A, $\odot$ là phép toán element-wise Hadamard, β=0 đối với IS-divergence.

Ký hiệu $W_{(S)}$ và $W_{(N)}$ lần lượt là ma trận đặc trưng phổ của tín hiệu tiếng nói và nhiễu. Trong trường hợp có dữ liệu huấn luyện, $W_{(S)}$ và $W_{(N)}$ được ước lượng qua bước huấn luyện từ các dữ liệu mẫu tương ứng. Sau đó ma trận đặc trưng phổ của hai nguồn được xây dựng bằng công thức:

$$W = [W_{(S)}, W_{(N)}] \quad (5)$$

NMF sẽ cố định ma trận đặc trưng phổ của tín hiệu W có được sau pha huấn luyện và ước lượng H bằng công thức cập nhật MU-rules (3), H gồm hai thành phần $H_{(S)}$ và $H_{(N)}$ là ma trận kích hoạt (time activations) tương ứng của speech và noise:

$$H = [H_{(S)}^T, H_{(N)}^T] \quad (6)$$

Sau khi ước lượng các thành phần W và H, tín hiệu speech và noise được xác định bởi công thức Wiener filtering:

$$\hat{S} = \frac{W_{(S)}H_{(S)}}{WH}\odot X, \quad (7)$$

$$\hat{N} = \frac{W_{(N)}H_{(N)}}{WH}\odot X \quad (8)$$

Chi tiết các bước của thuật toán được mô tả trong Algorithm 1.

Algorithm 1 Supervised-NMF source separation

Require: Tín hiệu mixture, Tập dữ liệu huấn luyện.

Ensure: Các nguồn cần tách.

BƯỚC HUẤN LUYỆN (TRAINING STEP)

Khởi tạo ngẫu nhiên ma trận không âm H, W.

Ước lượng W từ tập dữ liệu huấn luyện dựa trên NMF và công thức cập nhật tham số MU-rules (4) và (5).

BƯỚC TÁCH NGUỒN (SEPATATION STEP)

Cố định W và khởi tạo ngẫu nhiên ma trận H không âm.

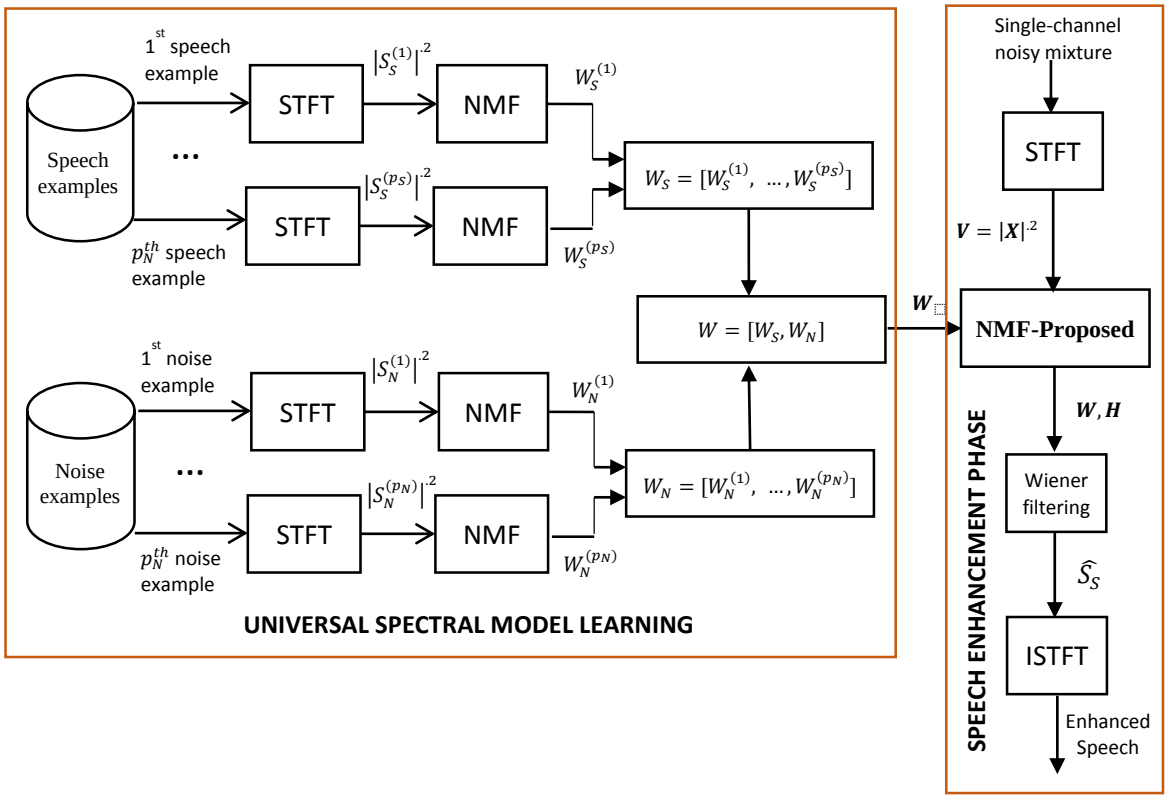
Ước lượng H từ tín hiệu mixture, dựa trên NMF và MU-rules.

Ước lượng các nguồn cần tách theo công thức (8) và (9).

Vấn đề giới hạn của thuật toán này là bắt buộc phải có dữ liệu huấn luyện cho tín hiệu cần tách, có nghĩa là để nâng cao tín hiệu tiếng nói, cần sử dụng giọng của chính người nói và âm thanh nhiễu môi trường cho bước huấn luyện, điều này không phải lúc nào cũng dễ dàng đáp ứng được trong các ứng dụng thực tế, chẳng hạn khi người nói di chuyển đến các môi trường khác nhau và khi không xác định được ai là người nói.

3. Đề xuất thuật toán kết hợp NMF với các ràng buộc thưa

Để giải quyết tình huống không có dữ liệu huấn luyện chính xác cho cả speech và noise, chúng tôi đề xuất trong phần này thuật toán nâng cao chất lượng tín hiệu tiếng nói bằng cách xây dựng mô hình phổ tổng quát cho cả hai tín hiệu speech và noise từ một tập dữ liệu huấn luyện cùng loại thu thập được. Ngoài ra chúng tôi sử dụng công thức kết hợp hai loại sparsity constraint trong mô hình NMF cho bước ước lượng tín hiệu speech và noise cần tách. Sơ đồ



Hình 1. Sơ đồ thuật toán nâng cao chất lượng tiếng nói đề xuất.

thuật toán được thể hiện trong Hình 1.

3.1. Xây dựng mô hình phổ tổng quát

Từ tập các tín hiệu cùng loại với tín hiệu cần tách (chẳng hạn với mục đích tách tín hiệu tiếng nói từ tín hiệu thu âm chứa nhiễu môi trường, dễ dàng thu thập được một vài file tiếng nói và một vài file âm thanh nhiễu môi trường bất kỳ để làm tập mẫu huấn luyện), ma trận đặc trưng phổ của từng mẫu huấn luyện được ước lượng bằng cách áp dụng mô hình NMF theo công thức (2), (3) và (4).

Giả sử p_S và p_N lần lượt là số lượng mẫu huấn luyện cho speech và noise, sau khi ma trận đặc trưng phổ của từng mẫu huấn luyện riêng biệt đã được ước lượng, ma trận đặc trưng phổ tổng quát của speech và noise được xây dựng theo công thức sau:

$$W_S = \left[W_S^{(1)}, W_S^{(2)}, \dots, W_S^{(p_S)}\right], \quad (9)$$

$$W_N = \left[W_N^{(1)}, W_N^{(2)}, \dots, W_N^{(p_N)}\right], \quad (10)$$

Công thức kết hợp hai sparsity constraint

Các mô hình phổ tổng quát được xây dựng theo công thức (9) và (10) sẽ có kích thước lớn khi số lượng mẫu huấn luyện cho các nguồn tăng lên (p_S và p_N lớn). Tuy nhiên, thường chỉ có một phần kích thước nhỏ của mô hình phổ tổng quát chứa các đặc trưng phổ giống với phổ tín hiệu cần tách (Virtanen 2007; Lefevre, Bach, and Févotte 2011). Do đó, trong pha tách nguồn âm, ràng buộc sparsity constraint được sử dụng với mục đích tìm ra những tập con của ma trận phổ tổng quát $\boldsymbol{W}_S, \boldsymbol{W}_N$ chứa các đặc trưng phổ của tín hiệu nguồn cần tách. Nói cách khác, ma trận phổ của tín hiệu mixture $\boldsymbol{V} = |\boldsymbol{X}|^{\cdot 2}$ sẽ được phân tách dựa trên việc tối ưu hóa hàm giá sau đây:

$$\min_{H\geq 0, W\geq 0} D(\boldsymbol{V}\|\boldsymbol{W}\boldsymbol{H}) + \lambda\Omega(\boldsymbol{H}) \quad (11)$$

Ở đó $\Omega(\boldsymbol{H})$ (được gọi là penalty function) thể hiện sự ảnh hưởng của ràng buộc thưa đối với $\boldsymbol{H}$. λ là tham số thể hiện mức độ ảnh hưởng: nếu $\lambda = 0$ thì $\boldsymbol{H}$ sẽ không bị ảnh hưởng bởi sparsity constraint, λ càng lớn thì mức ảnh hưởng càng cao.

Các nhóm nghiên cứu trước đây đã đề xuất hai penalty function tương ứng với hai loại ràng buộc thưa như sau:

$$\Omega_1(\boldsymbol{H}) = \sum_{g=1}^{G} log(\epsilon + \left\|\boldsymbol{H}_{(g)}\right\|_1), \quad (12)$$

$$\Omega_2(\boldsymbol{H}) = \sum_{k=1}^{K} log(\epsilon + \left\|\boldsymbol{h}_{(k)}\right\|_1). \quad (13)$$

Công thức (12) sử dụng block sparsity được nhóm nghiên cứu Reynolds và nnk đề xuất năm 2000 (Reynolds, Quatieri, and Dunn 2000). Ở đó ϵ là một hằng số dương đủ nhỏ, $\boldsymbol{H}_{(g)}$ là tập con của $\boldsymbol{H}$ và là ma trận activation của block thứ g. Trong trường hợp này, mỗi block đại diện cho một mẫu huấn luyện và G là tổng số mẫu huấn luyện được sử dụng ($G = p_S + p_N$). 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 giá trị. Công thức (13) sử dụng component sparsity được nhóm nghiên cứu Badawy và nnk đề xuất lần đầu năm 2014 và tiếp tục phát triển cho đến nay (El Badawy, Duong, and Ozerov 2014; Badawy, Duong, and Ozerov 2017) với $\boldsymbol{h}_{(k)}$ là dòng thứ k của $\boldsymbol{H}$. Hàm penalty (13) sẽ kích hoạt những "dòng" chứa các đặc trưng tương tự với nguồn cần tách.

Xuất phát từ nhận định thường chỉ có một phần của mô hình phổ tổng quát đã được huấn luyện từ tập mẫu 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 **W** có kích thước lớn (khi số lượng mẫu huấn luyện $p_S + p_N$ lớn) 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 **W** lớn. Từ những phân tích trên, chúng tôi đã đề xuất sử dụng kết hợp hai sparsity constraint để nâng cao hiệu quả của hàm penalty theo công thức sau (Duong et al. 2015, 2016):

$$\Omega(\boldsymbol{H}) = \alpha \sum_{g=1}^{G} \log\left(\epsilon + \left\|\boldsymbol{H}_{(g)}\right\|_1\right) + (1-\alpha) \sum_{k=1}^{K} \log\left(\epsilon + \left\|\boldsymbol{h}_{(k)}\right\|_1\right), \quad (14)$$

Trong đó α là trọng số thể hiện sự đóng góp của mỗi thành phần. Có thể xem (17) là sự tổng quát hóa của (12) và (13).

Thuật toán đề xuất được mô tả chi tiết trong Algorithm 2, với $\boldsymbol{H}_{(g)}$ là ma trận có cùng kích thước với $\boldsymbol{H}_{(g)}$, $\boldsymbol{z}_k$ là véc tơ dòng có cùng kích thước với với $\boldsymbol{h}_k$.

Algorithm 2 NMF - Proposed
Require: $\boldsymbol{V}, W, \lambda, \alpha$
Ensure: H
Khởi tạo H với các giá trị không âm ngẫu nhiên.
$\hat{V} = WH$
Repeat
//Tính toán thành phần block sparsity
For $g = 1, \dots, G$ do

$$Y_{(g)} \leftarrow \frac{1}{\varepsilon + \left\|H_{(g)}\right\|_1}$$

End for

$$Y = \left[Y_{(1)}^T, \dots, Y_{(G)}^T\right]^T$$

//Tính toán thành phần component sparsity
For $k = 1, \dots, K$ do

$$z_{(k)} \leftarrow \frac{1}{\varepsilon + \left\|h_{(k)}\right\|}$$

End for

$$Z = \left[z_{(1)}^T, \dots, z_{(K)}^T\right]^T$$

//Cập nhật H

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

$$\hat{V} \leftarrow WH$$

Until Thỏa mãn điều kiện hội tụ

4. Thí nghiệm và đánh giá kết quả

4.1. Dữ liệu thử nghiệm

Thí nghiệm sử dụng bộ dữ liệu được cung cấp bởi hai website uy tín thuộc lĩnh vực nghiên cứu là International Signal Separation and Evaluation Campaign (SiSEC) và Diverse Environments Multichannel Acoustic Noise Database (DEMAND), gồm tập dữ liệu huấn luyện và tập dữ liệu thử nghiệm.

- Tập mẫu huấn luyện speech gồm 4 file ".wav" có độ dài 10 giây mỗi file, là giọng của 4 người gồm 2 giọng nam và 2 giọng nữ và không trùng với những giọng nói trong dữ liệu test. Tập mẫu huấn luyện noise gồm 5 file ".wav" có kích thước từ 5 đến 15 giây, là âm thanh của các loại nhiễu môi trường khác nhau: kitchen sound, waterfall, metro, field sound, cafeteria.

- Tập dữ liệu test gồm 20 file thu âm đơn kênh là tín hiệu mixture của speech và nhiễu môi trường với tỷ lệ tín hiệu trên nhiễu (signal-to-distortion ratio) SNR = 0 dB, các file có kích thước từ 5 đến 10 giây. Các loại nhiễu trong dữ liệu test gồm: traffic + wind sound, oceanwaves, cafeteria + guita, forest birds + car, bird song, square,....

4.2. Phương pháp đánh giá kết quả

Kết quả của thuật toán đề xuất được so sánh với kết quả của 2 thuật toán đã được công bố trước đó bởi Reynolds và nnk (Reynolds, Quatieri, and Dunn 2000) và Badawy và nnk (El Badawy, Duong, and Ozerov 2014) trên cùng một bộ dữ liệu thử nghiệm và cùng điều kiện thí nghiệm. Trong đó thuật toán của Reynolds và nnk sử dụng kết hợp NMF với ràng buộc thưa block sparsity theo công thức (12) (gọi tắt là "NMF block sparsity"), thuật toán của Badawy và nnk sử dụng kết hợp NMF với ràng buộc thưa component sparsity theo công thức (13) (gọi tắt là "NMF-component sparsity"). Tín hiệu speech tách được từ ba thuật toán được tính toán các độ đo SDR (Source to Distortion Ratio), SIR (Source to Interference Ratio), SAR (Signal to Artifacts Ratio) với đơn vị đo là dB. Các độ đo càng lớn thì chất lượng của tín hiệu speech tách được là càng tốt. Để tính toán các độ đo đó, chúng tôi dùng bộ Tool được cung cấp và sử dụng phổ biến hiện nay trong cộng đồng nghiên cứu về xử lý âm thanh là BSS-EVAL Tools (E. Vincent, Gribonval, and Fevotte 2006).

4.3. Các tham số của thuật toán

Các file âm thanh được lấy mẫu với tần số 16000Hz; số thành phần đặc trưng phổ của speech và noise lần lượt là 32 và 16; số bước lặp là 100 cho cả hai pha huấn luyện và tách. Các tham số của nhóm ràng buộc thưa được lựa chọn tối ưu sau bước huấn luyện đối với từng thuật toán lần lượt như sau: $\lambda = 25$ với "NMF block sparsity"; $\lambda = 75$ với "NMF-component sparsity"; $\lambda = 120$ và $\alpha = 0.2$ cho thuật toán đề xuất "NMF proposed".

4.4. Kết quả và thảo luận

Kết quả trung bình của 20 tín hiệu speech tách được từ tập dữ liệu test với ba thuật toán khác nhau được thể hiện trong Bảng 1. Có thể quan sát thấy thuật toán đề xuất cho kết quả tốt hơn hai thuật toán trước đó đối với cả ba độ đo SDR, SIR và SAR. Đối với độ đo quan trọng nhất là SDR, thuật toán đề xuất có giá trị trung bình cao hơn 0.3 dB và 0.5 dB so với "NMF-Block sparsity" và "NMF-Component sparsity".

Bảng 1. Độ đo trung bình của tín hiệu tiếng nói sau khi tách.

Thuật toán	SDR (dB)	SIR (dB)	SAR (dB)
NMF block sparsity	6.6	12.5	10.4
NMF component sparsity	6.8	12.5	10.6
NMF proposed	7.1	12.9	10.7

Trong khi "NMF block sparsity" bỏ qua hoặc giữ lại toàn bộ những mẫu có ít đặc trưng tương đồng với tín hiệu cần tách (tùy theo độ lớn hay nhỏ của tham số λ), "NMF component sparsity" lại thường giữ lại quá nhiều những đặc tính không mấy tương đồng với tín hiệu cần tách (như thể hiện trên hình 2). Phương pháp đề xuất đã khắc phục được hai vấn đề trên bằng cách trước hết loại bỏ những "block" không tương đồng với tín hiệu cần tách, với những "block" còn lại sẽ loại bỏ tiếp những "component" tương ứng với những đặc tính không tương đồng. Chính vì vậy, phương pháp đề xuất cho phép ước lượng tốt hơn các đặc trưng của tín hiệu cần tách và cho kết quả tốt hơn đối với bài toán nâng cao chất lượng tín hiệu tiếng nói trong môi trường có mức nhiễu cao.

So với hai thuật toán trước đó, thuật toán đề xuất có thêm tham số α thể hiện tỷ lệ ảnh hưởng của mỗi loại ràng buộc thưa đối với ma trận kích hoạt $\boldsymbol{H}$. Do đó ở bước huấn luyện, sẽ cần thử nghiệm và lựa chọn α tối ưu nhất cho bước test và điều này làm tăng thời gian huấn luyện của thuật toán. Tuy nhiên với cấu hình máy tính PC phổ biến hiện nay và dữ liệu huấn luyện có độ dài khoảng vài phút thì độ chênh lệch thời gian là không đáng kể, chỉ khoảng vài giây.

5. Kết luận

Trong bài báo, chúng tôi đã trình bầy thuật toán áp dụng mô hình NMF để nâng cao chất

lượng tín hiệu tiếng nói từ tín hiệu thu âm đơn kênh chứa nhiễu theo hướng tiếp cận của phương pháp tách các âm thanh bị trộn lẫn. Thuật toán đó đòi hỏi phải có dữ liệu huấn luyện cho các tín hiệu cần tách. Để giải quyết trường hợp không có dữ liệu huấn luyện, chúng tôi đã đề xuất thuật toán kết hợp mô hình NMF với đồng thời hai loại ràng buộc thưa trong quá trình ước lượng tín hiệu cần tách. Kết quả thí nghiệm với tập dữ liệu test chứa các loại nhiễu môi trường khác nhau ở mức cao (SNR = 0 dB) đã cho thấy hiệu quả của thuật toán đề xuất.

Chúng tôi mong muốn sẽ tiếp tục phát triển thuật toán và kết hợp với mô hình không gian (spatial model) để ứng dụng cho trường hợp thu âm đa kênh (multi-channel). Đồng thời mong muốn thử nghiệm hiệu quả của thuật toán đối với hệ thống nhận dạng tiếng nói tự động (Automatic Speech Recognition - ASR).

Tài liệu tham khảo

Badawy, Dalia El, Ngoc Q. K. Duong, and Alexey Ozerov. 2017. On-the-Fly Audio Source Separation-A Novel User-Friendly Framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (2): 261-72. doi:10.1109/TASLP.2016.2632528.

Benesty, Jacob, S Makino, and J Chen. 2005. *Speech Enhancement*. Springer, Berlin.

Chen, Linlin, Xiaohong Ma, and Shuxue Ding. 2017. Single Channel Speech Separation Using Deep Neural Network. *In Advances in Neural Networks - ISNN 2017*. Springer International Publishing. doi:10.1007/978-3-319-59072-1_34.

Cohen, I. 2002. Optimal Speech Enhancement under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator. *IEEE Signal Processing Letters* 9 (4): 113-16. doi:10.1109/97.1001645.

Duong Hien-Thanh T., Quoc-Cuong Nguyen, Cong-Phuong Nguyen, and Ngoc Q. K. Duong. 2016. Single-Channel Speaker-Dependent Speech Enhancement Exploiting Generic Noise Model Learned by Non-Negative Matrix Factorization. In, 1-4. *The 15th IEEE International Conference on Electronics, Information and Communication.* doi:10.1109/ELINFOCOM.2016.7562952.

Duong Hien-Thanh T., Quoc-Cuong Nguyen, Cong-Phuong Nguyen, Thanh-Huan Tran, and Ngoc Q. K. Duong. 2015. Speech Enhancement Based on Nonnegative Matrix Factorization with Mixed Group Sparsity Constraint. In, 247-251. *The Sixth ACM International Symposium on Information and Communication Technology.* doi:10.1145/2833258.2833276.

El Badawy, Dalia, Ngoc Q. K. Duong, and Alexey Ozerov. 2014. On-the-Fly Audio Source Separation. In , 1-6. *IEEE.* doi:10.1109/MLSP.2014.6958922.

El Badawy, Dalia, Alexey Ozerov, and Ngoc Q. K. Duong. 2015. Relative Group Sparsity for Non-Negative Matrix Factorization with Application to on-the-Fly Audio Source Separation. In , 256-60. *IEEE.* doi:10.1109/ICASSP.2015.7177971.

Févotte, Cédric, Nancy Bertin, and Jean-Louis Durrieu. 2009. Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis. *Neural Computation* 21 (3): 793-830. doi:10.1162/neco.2008.04-08-771.

Gibak Kim, and P C Loizou. 2010. Improving Speech Intelligibility in Noise Using Environment-Optimized Algorithms. *IEEE Transactions on Audio, Speech, and Language Processing* 18 (8): 2080-90. doi:10.1109/TASL.2010.2041116.

Kinoshita, Keisuke, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Armin Sehr, Walter Kellermann, and Roland Maas. 2013. The Reverb Challenge: Acommon Evaluation Framework for Dereverberation and Recognition of Reverberant Speech. In , 1-4. *IEEE*. doi:10.1109/WASPAA.2013.6701894.

Lee, Daniel D., and H. Sebastian Seung. 2001. Algorithms for Non-Negative Matrix Factorization. In *Advances in Neural Information Processing Systems*, 556-62.

Lefevre, Augustin, Francis Bach, and Cédric Févotte. 2011. Itakura-Saito Nonnegative Matrix Factorization with Group Sparsity.

In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference On, 21-24. IEEE.

Liutkus, Antoine, Fabian-Robert Stöter, Zafar Rafii, Daichi Kitamura, Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave. 2017. The 2016 Signal Separation Evaluation Campaign. In *Latent Variable Analysis and Signal Separation*, 323-32. Cham: Springer International Publishing. doi:10.1007/978-3-319-53547-0_31.

Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn. 2000. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing* 10 (1-3): 19-41. doi:10.1006/dspr.1999.0361.

Smaragdis, Paris, Bhiksha Raj, and Madhusudana Shashanka. 2007. Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures*. In Independent Component Analysis and Signal Separation*, 4666:414-21. Springer Berlin Heidelberg. doi:10.1007/978-3-540-74494-8_52.

Sun, Dennis L., and Gautham J. Mysore. 2013. Universal Speech Models for Speaker Independent Single Channel Source Separation. In , 141-45. *IEEE*. doi:10.1109/ICASSP.2013.6637625.

Vincent, E., R. Gribonval, and C. Fevotte. 2006. Performance Measurement in Blind Audio Source Separation. *IEEE Transactions on Audio, Speech and Language Processing* 14 (4): 1462-69. doi:10.1109/TSA.2005.858005.

Vincent, Emmanuel, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni. 2013. The Second 'Chime' Speech Separation and Recognition Challenge: Datasets, Tasks and Baselines. In , 126-30. *IEEE*. doi:10.1109/ICASSP.2013.6637622.

Virtanen, Tuomas. 2007. Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria. *IEEE Transactions on Audio, Speech and Language Processing* 15 (3): 1066-74. doi:10.1109/TASL.2006.885253.

ABSTRACT

Single-channel speech enhancement method for high-level background noise mixture

Thanh Hien Thi Duong [1], Huan Thanh Tran [2], Hang Thu Nguyen [1], Hien Quang Pham [1], Liên Kim Thi Vu [1]

[1] Faculty of Information Technology, Hanoi University of Mining and Geology, Vietnam.
[2] Hanoi University of Industry, Hanoi City, Vietnam.

This paper focuses on using the single-channel source separation techniques to improve the quality of the desired speech in the real-world environment where the speech signal is corrupted by high-level background noise, and especially when there is no source-specific training data. We propose a solution combining the Nonnegative Matrix Factorization model (NMF) with mixed group sparsity constraints to separate the speech signal from the single - channel audio signal with high ambient noise. Experiment result over mixtures containing different real-world noises confirms the effectiveness of the proposed algorithm.

Keywords: speech enhancement, audio source separation, nonnegative matrix factorization, universal spectral model, group sparsity.