TÁCH NGUỒN MÙ (BSS) ÁP DỤNG CHO ÂM THANH TRONG MỘT SỐ ĐIỀU KIỆN KHÁC NHAU

Trương Tấn Quang, Trần Quang Huy, Nguyễn Hữu Phương

Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM

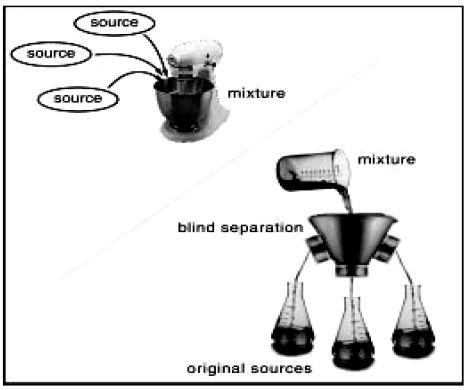
(Bài nhận ngày 21 tháng 03 năm 2011, hoàn chỉnh sửa chữa ngày 23 tháng 04 năm 2012)

TÓM TẮT: *Tai ta thường đồng thời tiếp nhận nhiều nguồn âm (tiếng nói, âm nhạc, nhiễu…) khác nhau nhưng ta vẫn có thể lắng nghe nguồn âm chủ định. Một hệ thống nhận dạng tiếng cần đạt đến khả năng thông minh như vậy. Bài toán là từ nhiều tín hiệu đã trộn lẫn ta muốn khôi phục các tín hiệu nguồn riêng rẽ. Đây là bài toán tách nguồn mù (Blind Source Separation - BSS). Trong hơn chục năm qua, người ta đã phát triển một phương pháp mới giúp giải bài toán nêu trên rất hiệu quả, đó là phân tích thành phần độc lập (Independent Component Analysis – ICA). Có nhiều thuật toán ICA cho các ứng dụng khác nhau. Báo cáo trình bày ứng dụng ICA cho tách âm trường hợp số nguồn nhiều hơn số trộn (dưới xác định). Chúng tôi thực nghiệm trên nhiều loại tín hiệu. Kết quả rất tốt.*

Từ khóa*: tách mguồn mù, phân tích thành phần độc lập, dưới xác định.*

MỞ ĐẦU

Bài toán phân tách nguồn mù BSS (Blind Source Separation) đang được quan tâm nghiên cứu và ứng dụng trong nhiều lĩnh vực xử lý tín hiệu khác nhau: tách âm, nhận dạng, tín hiệu y sinh [1][2][3]. Bài toán BSS cho phép ước lượng lại các tín hiệu nguồn nguyên bản mà chỉ dựa vào những dữ liệu hỗn hợp thu được tại các cảm biến khảo sát và đặc trưng của kênh truyền cũng như các tín hiệu nguồn gần như không biết (Hình 1).



Hình 1. Mục đích của phân tách nguồn mù là chỉ sử dụng tín hiệu hỗn hợp lai trộn để tìm lại tín hiệu gốc

Tổng quát bài toán phân tách nguồn mù - BSS được phát biểu như sau: cho M hỗn hợp lai trộn tuyến tính từ N nguồn tạo qua ma trận lai trộn M x N không biết trước **A**. Bài toán phân tách nguồn mù BSS yêu cầu phân tích cấu trúc dữ liệu khảo sát và tách các nguồn gốc từ hỗn hợp lai trộn này. Khi M ≥ N, có thể thực hiện bằng cách xây dựng ma trận giải lai trộn **W**, với $\mathbf{W}=\mathbf{A}^{-1}$. Để đảm bảo phân tách được, các điều kiện cần tuân theo định lý Darmois [1]: các nguồn là phi Gauss và độc lập thống kê. Số chiều của quá trình lai trộn ảnh hưởng đến tính phức tạp của bài toán. Nếu M= N, ma trận lai trộn **A** được xem là *xác định chẵn (Even-determined)* hay *xác định* (*Determined*), các tín hiệu nguồn được phân tách qua biến đổi tuyến tính. Nếu M > N, ma trận **A** được xem là *trên xác định (Over-determined)* , có thể ước lượng các nguồn qua tối ưu bình phương tối thiểu hoặc biến đổi tuyến tính giả nghịch đảo ma trận. Nếu M < N quá trình lai trộn được xem như *dưới xác định (Under-determined)* và hệ quả là khôi phục các tín hiệu gốc phức tạp hơn và luôn được thực hiện qua kỹ thuật phi tuyến [2].

Những giả thiết về môi trường xung quanh các cảm biến khảo sát cũng đồng thời ảnh hưởng đến tính phức tạp của bài toán. Phân tách mù tín hiệu âm thường được liên hệ đến ví dụ bài toán Cocktail Party [4], tức là phân tách các tiếng độc lập từ vô vàn tiếng nói trong môi trường âm không kiểm soát được. Các cảm biến khảo sát còn bị lẫn lộn với bởi các rung động tín hiệu, dẫn đến ước lượng ma trận giải lai trộn cần nhận biết nguồn đến từ nhiều hướng khác nhau tại nhiều thời điểm khác nhau của cùng một nguồn phát. Tổng quát, bài toán phân tách mù xuất phát từ thực tế rất phức tạp và khó khăn, do đó yêu cầu giới hạn các giả thiết thực tế nhằm giúp bài toán có thể xử lý được. Có ba dạng giả thiết cơ bản về môi trường. Cơ bản nhất là trường hợp *lai trộn tức thời (Instantaneous)*, trong đó các tín hiệu đến các cảm biến tức thời, chỉ sai khác biên độ. Mở rộng giả thiết này là xem xét có trễ giữa các cảm biến được biết là trường hợp *lai trộn có trễ (Anechoic)*. Tiếp tục mở rộng bằng cách xem có sự phản xạ nhiều đường tín hiệu giữa mỗi nguồn phát và mỗi cảm biến cho trường hợp *lai trộn có dội (Echoic)*, đôi khi còn được xem là lai trộn có chập. Mỗi trường hợp có thể mở rộng, kết hợp tuyến tính với nhiễu cộng, mà thường giả sử là nhiễu trắng, Gauss.

Trong báo cáo này, chúng tôi thực hiện tách âm với mô hình lai trộn tức thời - dưới xác định, có số tín hiệu nguồn cần ước lượng lớn hơn số tín hiệu thu được tại các cảm biến (N>M).

PHƯƠNG PHÁP

Phân tích thành phần độc lập IOA

Để định nghĩa ICA, ta có thể sử dụng mô hình các biến thống kê. Khảo sát n biến ngẫu nhiên $x_1(t)$, …, $x_n(t)$ là tổ hợp tuyến tính của n biến ngẫu nhiên $s_1(t)$, …, $s_n(t)$:

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \ldots + a_{in}s_n, \quad \text{với mọi } i = 1, \ldots, n \qquad (1)$$

với a_{ij}, i, j = 1, …, n là các hệ số thực. Mô hình này mô tả quá trình phát sinh lai trộn của các thành phần s_j. Các thành phần độc lập s_j

(thường được viết tắt thành ICs _ independent components) là các biến ẩn (latent variables), có nghĩa là không thể khảo sát trực tiếp chúng và các hệ số a_{ij} cũng không biết. Tất cả thông tin có được chỉ là các biến ngẫu nhiên x_i, và ta phải ước lượng tìm cả các hệ số lai trộn a_{ij} và ICs s_i mà chỉ sử dụng thành phần lai x_i.

Ở đây thông tin chỉ số thời gian t được bỏ qua bởi vì trong mô hình ICA cơ sở, giả thiết rằng mỗi thành phần lai x_i cũng như mỗi thành phần độc lập s_j là một biến ngẫu nhiên, thay vì là một tín hiệu thời gian hay chuỗi thời gian. Các giá trị khảo sát $x_i(t)$, chẳng hạn tín hiệu micro trong bài toán cocktail-party là các mẫu của biến ngẫu nhiên này. Đồng thời bỏ qua các thời gian trì hoãn có thể xuất hiện trong quá trình lai trộn, vì vậy đây được gọi là mô hình lai trộn tức thời (instantaneous mixing model).

ICA là phương pháp thống kê giải quyết bài toán BSS hoặc phân tách tín hiệu mù (blind signal separation). Nguồn "source" ở đây có nghĩa là tín hiệu nguyên thủy, như là âm phát ra từ mỗi người nói trong bài toán cocktail-party. Mù "blind" có nghĩa là biết rất ít về ma trận lai trộn, và các giả thiết về tín hiệu nguồn hầu như không đáng kể.

Để thuận tiện, ta sử dụng các ký hiệu vectơ – ma trận thay cho tổng ở các phương trình trên. Theo đó mô hình lai trộn được viết lại như sau:

$$\mathbf{x} = \mathbf{A}\,\mathbf{s} \quad (2)$$

Điều kiện giới hạn trong ICA

- Các thành phần độc lập được xem là độc lập thống kê.

- Các thành phần độc lập phải có phân bố phi Gauss.

- Ma trận lai trộn là vuông.

Tính nhập nhằng (không xác định) của ICA

- Không thể xác định chính xác phương sai (năng lượng) của các thành phần độc lập.

- Không thể xác định thứ tự của các thành phần độc lập.

Thuật toán tách âm dưới xác định

Tách âm mù dưới xác định được thực hiện qua hai bước: ước lượng ma trận lai trộn và phân tách nguồn [6]. Thuật toán ước lượng các hệ số ma trận lai trộn dựa trên cấu trúc không gian, tìm các vectơ hướng trên phân bố tín hiệu trong đồ thị phân tán hỗn hợp lấy từ các cảm biến khảo sát và yêu cầu các nguồn phải có biểu diễn đủ thưa [4][5]. Mỗi hướng đặc trưng bởi vectơ cột của ma trận lai trộn, và sau đó kết hợp các vectơ hướng tìm được để xác định ma trận lai trộn ước lượng . Sau khi tìm được ma trận lai trộn ước lượng, việc phân tách nguồn (dưới xác định) trở thành bài toán giải hệ phương trình tuyến tính. Bài toán được phát biểu dưới dạng bài toán tối ưu tuyến tính có lời giải tối thiểu chuẩn L_1 biểu diễn thưa [2][7]. Để đảm bảo bài toán đạt tỉ lệ thành công cao với đa dạng tập dữ liệu, đặc tính thưa của tín hiệu được cải thiện qua phép biến đổi STFT, nghĩa là toàn bộ quá trình phân tách nguồn được thực hiện trong miền biến đổi STFT (thời gian-tần số) [8]. Kết quả sau đó chuyển về miền thời gian ban đầu.

Thuật toán ước lượng ma trận lai trộn

1. Biến đổi dữ liệu khảo sát trong miền thời gian $\mathbf{x}_i$ là hàng thứ i của $\mathbf{X}$, i=1,…,M sang miền

thời gian-tần số sử dụng biến đổi STFT, các hệ số được đo bởi kurtosis.

2. Khởi tạo ngẫu nhiên N vector hướng $\mathbf{v}_i$ và γ một giá trị đủ lớn.

3. Gán từng phần mỗi điểm dữ liệu $\mathbf{x}(t)$ đến mỗi vector hướng $\mathbf{v}_i$, sử dụng phép gán mềm:

$$q_{it} = \left\|\mathbf{x}(t) - (\mathbf{v}_i.\mathbf{x}(t))\mathbf{v}_i\right\|^2,$$

$$\tilde{q}_{it} = \frac{e^{-\gamma q_{it}}}{\sum_{i'} e^{\gamma q_{i't}}}$$

Tham số γ kiểm soát độ trơn tại biên giữa các vùng đặc trưng cho mỗi đường, $\tilde{q}_{it}$ là các trọng số của dữ liệu khảo sát tại thời điểm t đối với mỗi đường i.

4. Tính ma trận hiệp phương sai của dữ liệu khảo sát có trọng số đã gán đến mỗi đường. Biểu thức ma trận hiệp phương sai và các trọng số được biểu diễn:

$$\sum_i = \frac{\sum_t \tilde{q}_{it}(\mathbf{x}(t)-\mu)(\mathbf{x}(t)-\mu)^T}{\sum_t \tilde{q}_{it}}$$

ở đây μ là vector trị trung bình các hàng của $\mathbf{X}$; Đối với tín hiệu âm, μ có trung bình không; $\mathbf{\Sigma}_i$ là ma trận hiệp phương sai của dữ liệu khảo sát có trọng số kết hợp với đường thứ i.

5. Cập nhật hướng mới theo vectơ riêng chính của mỗi ma trận hiệp phương sai bằng cách khai triển vectơ riêng của $\mathbf{\Sigma}_i$:

$$\sum_i = \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^{-1}$$

ma trận riêng $\mathbf{U}_i$ chứa các vectơ riêng của $\mathbf{\Sigma}_i$ và ma trận chéo $\mathbf{\Lambda}_i$ chứa các trị riêng tương ứng $\lambda_i,\ldots, \lambda_M$. Ước lượng hướng vectơ mới là vectơ riêng chính của $\mathbf{\Sigma}_i$: $\mathbf{v}_i \leftarrow \mathbf{u}_{max}$

vectơ riêng chính $\mathbf{u}_{max}$ là vector riêng có trị riêng lớn nhất λ_{max}.

6. Cập nhật γ sử dụng phương sai theo hướng trực giao hóa mỗi đường: chọn trị riêng lớn nhất thứ hai từ mỗi $\mathbf{\Lambda}_i$ và lấy nghịch đảo:

$$\gamma \leftarrow \frac{1}{\max(\lambda_i,\ldots,\lambda_M)}$$

λ_i là trị riêng lớn nhất thứ hai của $\mathbf{\Sigma}_i$. Trở lại bước 3 và lặp lại cho đến khi $\mathbf{v}_i$ hội tụ.

Sau khi hội tụ, kết hợp các vectơ hướng $\mathbf{v}_i$ ước lượng được lập thành ma trận lai trộn ước lượng: $\hat{\mathbf{A}}=[\mathbf{v}_1|\ldots|\mathbf{v}_N]$.

Thuật toán phân tách nguồn

Thực hiện ước lượng $\hat{\mathbf{A}}$

Trường hợp dưới xác định: Quá trình phân tách được thực hiện qua bài toán tối ưu, lời giải tối thiểu chuẩn L_1 cho mỗi dữ liệu khảo sát trong miền thưa STFT:

$$\arg\min_{\hat{\mathbf{s}}(\omega)\in\mathbb{C}^N} \left\|\hat{\mathbf{s}}(\omega)\right\|_1 \quad \text{theo} \quad \hat{\mathbf{A}}\hat{\mathbf{s}}(\omega) = \mathbf{x}(\omega)$$

Sau đó thực hiện biến đổi ngược ISTFT chuyển về miền tín hiệu ban đầu:

$$\hat{\mathbf{s}}(\omega) \rightarrow \hat{\mathbf{s}}(t)$$

Kết quả cuối cùng là ma trận $\hat{\mathbf{S}}$ kích thước NxT với các hàng là các nguồn ước lượng: $\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_N$

KẾT QUẢ

Chúng tôi triển khai thực nghiệm trên PC sử dụng ngôn ngữ Matlab thực hiện tách âm trong trường dưới xác định với mô hình lai trộn tức thời.

Thu các nguồn âm thực nghiệm ở tốc độ lấy mẫu 16kHz và 22.05kHz, mã hóa PCM 16 bit,

chiều dài mỗi đoạn dữ liệu âm là 10s. Thực hiện lai trộn âm từ các nguồn âm có sẵn. Mô hình lai trộn là ngẫu nhiên hoặc tự định nghĩa để phù hợp với điều kiện môi trường thực tế. Số cảm biến khảo sát được giữ ở mức tối thiểu là hai cảm biến cho tất cả thực nghiệm.

Dữ liệu lai trộn các nguồn âm trong miền thời gian được biến đổi qua miền thưa (thời gian – tần số) sử dụng phép biến đổi STFT với cửa sổ 1024 điểm. Sau cùng, dữ liệu trong miền biến đổi đưa qua bộ tách âm dưới xác định với thuật toán đã trình bày trong phần 4. Kết quả là các âm phân tách độc lập được đánh giá khách quan với các tỉ số SDR, SIR, SAR [6], và đánh giá chủ quan qua việc quan sát dạng sóng, nghe âm phát lại ở loa so với nguồn âm gốc.

Thực nghiệm 1 (nguồn âm gốc: ba giọng nữ)

Thực hiện phân tách hai hỗn hợp lai trộn tức thời từ ba nguồn âm độc lập đặc trưng giọng nữ. Kết quả nhận được rất tốt với các tỉ số đánh giá khách quan cho từng nguồn âm ước lượng:

Tỉ số/âm ước lượng	$\mathbf{se}_1$ ($\mathbf{s}_1$)	$\mathbf{se}_2$ ($\mathbf{s}_2$)	$\mathbf{se}_3$ ($\mathbf{s}_3$)
SDR (dB)	12,1	9,2	10,8
SIR (dB)	13,5	12,9	12,5
SAR (dB)	12,4	9,4	11,3

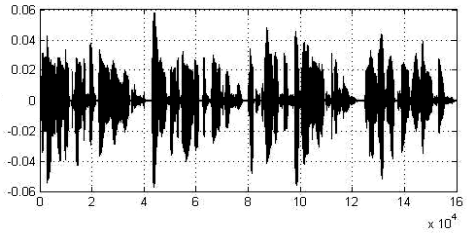
Về đánh giá chủ quan (hình 2), dạng sóng tín hiệu âm được khôi phục đúng với nguồn âm gốc và âm nghe được phân biệt từng nguồn rõ ràng. Thứ tự tương ứng các nguồn âm $\mathbf{se}_1$ là $\mathbf{s}_1$; $\mathbf{se}_2$ là $\mathbf{s}_2$, và $\mathbf{se}_3$ là $\mathbf{s}_3$. Lưu ý rằng đặc trưng về biên độ, pha và thứ tự các nguồn âm ước lượng có thể không đúng với nguồn âm gốc, đây cũng chính là giới hạn của ICA [1].

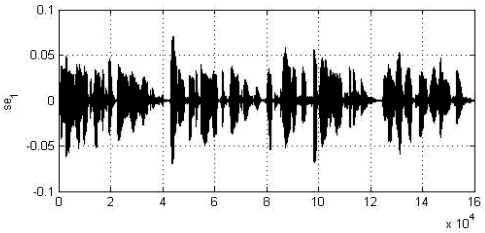
Thực nghiệm 2 (nguồn âm gốc: hai giọng nói và một nhạc)

Thực nghiệm 2 thực hiện với hỗn hợp các đặc trưng khác nhau: hai giọng nói $\mathbf{s}_1$, $\mathbf{s}_2$ và một nhạc $\mathbf{s}_3$. Các tỉ số đánh giá khách quan cho từng nguồn âm ước lượng:

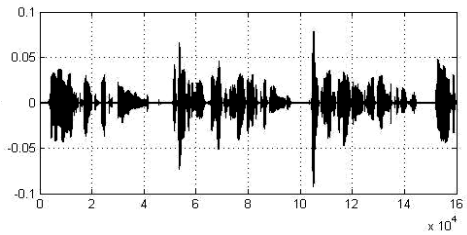
Tỉ số/âm ước lượng	$\mathbf{se}_1$ ($\mathbf{s}_2$)	$\mathbf{se}_2$ ($\mathbf{s}_1$)	$\mathbf{se}_3$ ($\mathbf{s}_3$)
SDR (dB)	11,2	11,9	14,1
SIR (dB)	14,1	14,1	14,8
SAR (dB)	11,3	12,1	14,3

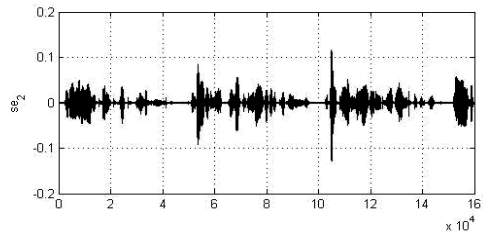
Kết quả đánh giá khách quan và chủ quan (Hình 3) cho thấy việc phân tách tốt mặc dù hỗn hợp là lai trộn các âm có đặc trưng khác nhau, trong đó nguồn âm nhạc ước lượng $\mathbf{se}_3$ có chất lượng rất tốt.



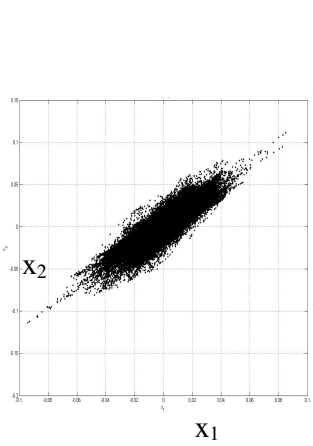


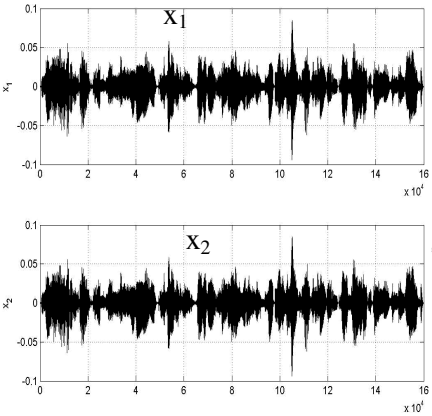
Nguồn âm gốc $\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$

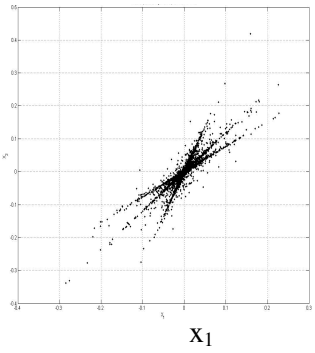




Nguồn âm ước lượng $\mathbf{se}_1$, $\mathbf{se}_2$, $\mathbf{se}_3$



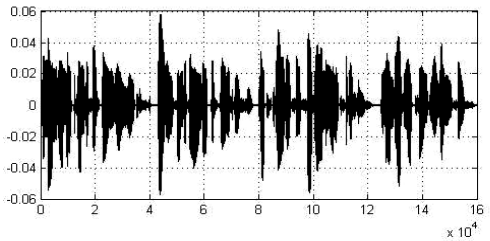




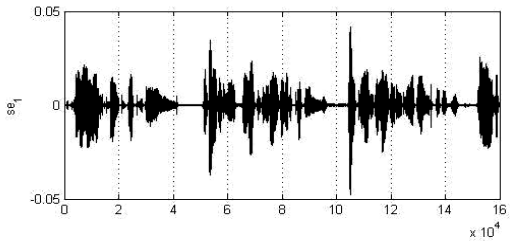
Đồ thị phân tán (thời gian) Hỗn hợp lai trộn $\mathbf{x}_1$, $\mathbf{x}_2$ Đồ thị phân tán (thưa)

$$\mathbf{A} = \begin{bmatrix} 0.4158 & 0.7815 & 0.9646 \\ 0.9425 & 0.9314 & 0.7402 \end{bmatrix} \qquad \mathbf{A\,e} = \begin{bmatrix} 0.3417 & 0.9396 & 0.7917 \\ 0.9398 & 0.7687 & 0.6109 \end{bmatrix}$$

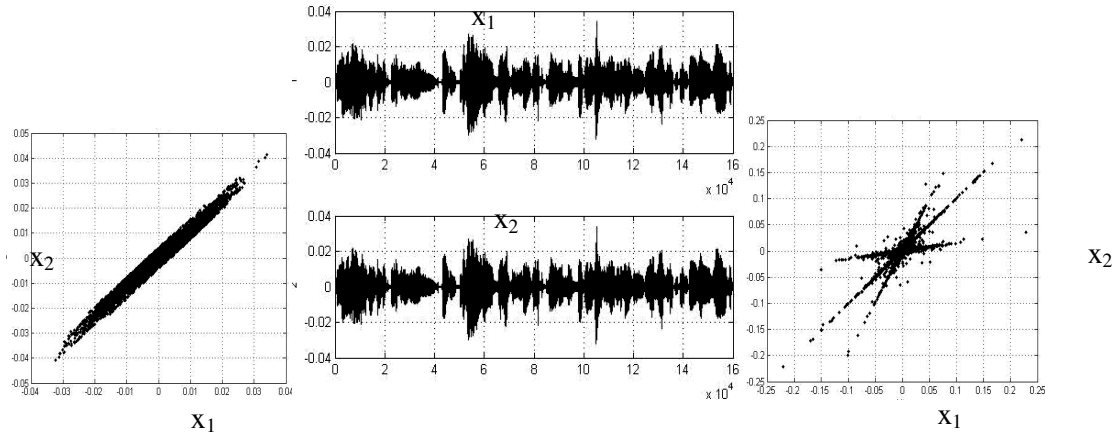
Hình 2. Kết quả dạng sóng, đồ thị phân tán của hỗn hợp trong thực nghiệm 1



Nguồn âm gốc $\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$



Nguồn âm ước lượng $\mathbf{se}_1$, $\mathbf{se}_2$, $\mathbf{se}_3$



Đồ thị phân tán (thời gian) Hỗn hợp lai trộn $\mathbf{x}_1$, $\mathbf{x}_2$ Đồ thị phân tán (thưa)

$$\mathbf{A} = \begin{bmatrix} 0.3102 & 0.3245 & 0.3503 \\ 0.3592 & 0.4041 & 0.2689 \end{bmatrix}$$

$$\mathbf{A_e} = \begin{bmatrix} 0.6261 & 0.6519 & 0.7931 \\ 0.7798 & 0.7583 & 0.6091 \end{bmatrix}$$

Hình 3. Kết quả dạng sóng, đồ thị phân tán của hỗn hợp trong thực nghiệm 2

KẾT LUẬN

Như vậy, chúng tôi đã thực hiện thành công ứng dụng kỹ thuật ICA giải quyết bài toán BSS trong việc tách âm. Các thực nghiệm được tiến hành với các đặc trưng âm khác nhau như nguồn giọng nam, nguồn đặc trưng giọng nữ, có nguồn là nhạc không lời. Điều này cho thấy sức mạnh và tính hiệu quả của kỹ thuật ICA trong ứng dụng tách âm. Các kết quả thực nghiệm được đánh giá khách quan với các tỉ số S/N và chủ quan bằng việc quan sát dạng sóng tín hiệu, nghe âm phát lại đều đạt kết quả tốt.

Đồng thời, điều kiện ràng buộc về số nguồn bằng số cảm biến khảo sát trong mô hình ICA cơ sở đã được tháo gỡ minh chứng qua các thực nghiệm có số nguồn âm lớn hơn số tín hiệu hỗn hợp thu được tại cảm biến (micro) trong trường hợp dưới xác định. Tuy nhiên, vẫn còn rất nhiều điều kiện ràng buộc và giới hạn khác cần được nghiên cứu để giúp hoàn thiện kỹ thuật ICA còn rất mới mẻ này.

Kỹ thuật ICA là phương pháp xử lý tín hiệu dựa trên đặc trưng thống kê, cho phép xử lý nhiều nguồn tín hiệu thay vì chỉ một nguồn tín hiệu đơn thuần dựa trên đặc tính phổ. Thật sự phương pháp ICA đã trở thành một công cụ phân tích thống kê mới bên cạnh các kỹ thuật truyền thống như PCA, FCA … Các kỹ thuật về phân tích nguồn mù này đã có nhiều nghiên cứu phát triển và chắc rằng các giới hạn cũng như điều kiện ràng buộc của mô hình ICA cơ sở sẽ được giải quyết trong một tương lai gần.

BLIND SOURCE SEPARATION (BSS) APPLIED TO SOUND IN VARIOUS CONDITIONS

Truong Tan Quang, Tran Quang Huy, Nguyen Huu Phuong
University of Science, VNU-HCM

ABSTRACT: *Our ears often simultaneously receive various sound sources (speech, music, noise . . .), but we can still listen to the intended sound. A system of speech recognition must be able to achieve the same intelligent level. The problem is that we receive many mixed (combined) signals from many different source signals, and would like to recover them separately. This is the problem of Blind Source Separation (BSS). In the last decade or so a method has been developed to solve the above problem effectively, that is the Independent Component Analysis (ICA). There are many ICA algorithms for different applications. This report describes our application to sound separation when there are more sources than mixtures (underdetermined case). The results were quite good.*

Key words: *blind source separation, independent component analysis, underdetermined.*

TÀI LIỆU THAM KHẢO

[1]. A. Hyvarinen, Karhunen, J., and Oja, E. *Independent Component Analysis*. John Wiley & Sons, Inc, (2001).

[2]. A. Cichocki, S. Amari, *Adaptive Blind Signal and Image Processing*. John Wiley & Sons, (2002).

[3]. T-Won. Lee, H. Sawada, *Blind Speech Separation.* Springer, ISBN 978-1-4020-6478-4, (2007).

[4]. B. A. Pearlumutter, Asari and Zador, *Sparse Representations for the Cocktail Party Problem*, CVS: hrtf source.tex 1.326, http://www.bcl.hamilton.ie/~barak/papers/hrtf-1ear-jns.pdf, (2006).

[5]. P. Bofil, M. Zibulevsky. Underdetermined blind source separation using sparserepresentations. *Signal Processing,* 81, 2353–2362, (2001).

[6]. P. D. O'Grady, *Sparse Separation of Under-Determined Speech Mixture.* Department of Computer Science National University of Ireland, Maynooth, *www.hamilton.ie/publications/ogrady2007_phd.pdf*, (2007).

[7]. Takigawa, M. Kudo, A. Nakamura, J. Toyama. *On the Minimum L_1-Norm Signal recovery in Underdetermined Source Separation.* Springer-Verlag, (2004).

[8]. P. Bofill, M. Zibulevsky. *Blind separation of more sources than mixtures using the sparsity of the short-time fourier transform*. 2nd International Workshop on Independent Component Analysis and Blind Signal Separation, pages 87–92, Helsinki, Finland, June 19–20 (2000).