한국음향학회지 제35권 제4호 pp. 303~309 (2016)
The Journal of the Acoustical Society of Korea Vol.35, No.4 (2016)
http://dx.doi.org/10.7776/ASK.2016.35.4.303

pISSN : 1225-4428
eISSN : 2287-3775

복층 자기부호화기를 이용한 음향 신호 군집화 및 분리

Audio signal clustering and separation using a stacked autoencoder

장길진†

(Gil-Jin Jang†)

경북대학교 전자공학부

(Received June 9, 2016; revised June 22, 2016; accepted July 13, 2016)

초 록: 본 논문은 자기부호화기를 이용한 음향신호 분리방법을 제안한다. 사용된 복층구조 신경망 자기부호화기는 입력 신호의 효율적인 표현방법을 자동으로 학습하며, 유사한 특징을 가지고 있는 요소신호들을 군집함으로써 다른 특징의 신호들을 분리할 수 있다. 시간영역과 주파수영역의 변이특성을 추출하기 위하여 단구간푸리에변환(Short-Time Fourier Transform, STFT)을 수행하였으며, 정해진 크기의 사각형 창을 모든 가능한 위치에 적용하여 얻은 단구간 주파수 스펙트럼을 자기부호화기의 입력으로 사용하였다. 자기부호화기의 부호노드들의 값을 이용하여 유사한 스펙트럼 창들을 군집하고, 이를 이용하여 원래의 음원들로 분리해 낼 수 있었다. 분리된 원음들은 원래의 입력신호의 특징을 확실히 나타내었으며, 기존의 비음수 행렬분해(Non-negative Matrix Factorization, NMF) 결과와 주파수 스펙트럼 비교를 통해 그 유효성을 보일 수 있었다.

핵심용어: 음향신호 분리, 자기부호화기, 심화신경회로망, 음향신호 군집

ABSTRACT: This paper proposes a novel approach to the problem of audio signal clustering using a stacked autoencoder. The proposed stacked autoencoder learns an efficient representation for the input signal, enables clustering constituent signals with similar characteristics, and therefore the original sources can be separated based on the clustering results. STFT (Short-Time Fourier Transform) is performed to extract time-frequency spectrum, and rectangular windows at all the possible locations are used as input values to the autoencoder. The outputs at the middle, encoding layer, are used to cluster the rectangular windows and the original sources are separated by the Wiener filters derived from the clustering results. Source separation experiments were carried out in comparison to the conventional NMF (Non-negative Matrix Factorization), and the estimated sources by the proposed method well represent the characteristics of the orignal sources as shown in the time-frequency representation.

Keywords: Audio signal separation, Autoencoder, Deep neural networks, Audio clustering

PACS numbers: 43.60.-c, 43.71.-k

I. 서 론

음성인식과 실감음향과 같은 응용분야에서는 입력 오디오 신호에 필요하지 않은 음원이 포함되어 있을 경우 그 성능이 크게 저하된다. 특히 하나의 마이크로폰만이 주어져 있을 경우, 원하지 않는 음원을 제거하는 것은 매우 어려운 문제이다. 기존의 해결 방법의 하나로 혼합신호의 기본주파수를 추정하고, 이의 정수배로 구성한 마스크를 단구간 주파수 스펙트럼에 곱하여 분리하는 방법이 제안되었다.[1] 이 방법은 혼합된 신호에서 잡음을 배재하고 음성신호만을 효과적으로 분리해낼 수 있지만, 입력신호가 음성이 아닌 일반적인 음향신호일 경우 성능이 저하되는 문제점이 있다. 다른 방법으로는 단구간 주파수 스펙트럼에서 반복되는 구조를 추정하고 이를 음원분리에 사용하는 비음수 행렬 분해(Non-negative Matrix

†**Corresponding author:** Gil-Jin Jang (gjang@knu.ac.kr)
School of Electronics Engineering, Kyungpook National University, 80 Daehakro, Bukgu, Daegu 41566, Republic of Korea
(Tel: 82-53-950-5517, Fax: 82-53-950-5055)

Factorization, NMF) 기반 음원분리 방법이 있으며,[2] 다양한 종류의 악기음 분리가 가능하다는 장점이 있다. 하지만 선행조건인 비음수 선형 결합으로 인하여 악기음이 아닌 일반적인 음성신호의 분리에서는 그 성능이 저하되는 단점이 있다.

본 논문에서는 단일 마이크로폰 입력에 대하여 심화신경망(Deep Neural Network, DNN)의 일종인 복층 자기부호화기(stacked autoencoder)를 적용하고, 이를 이용하여 음원분리를 가능하게 하는 새로운 방법을 제안한다. 기존의 자기부호화기는 목적신호가 없이 입력신호만으로 학습하는 비교사 학습 방식의 신경회로망의 한 종류로써, 이전의 연구에서는 영상신호의 잡음제거 및 압축에 사용되었다.[3,4] 제안된 방법에서는 혼합된 입력신호를 복층 자기부호화기로 학습하여 음원들의 특징이 구분되도록 자동으로 군집화하고, 적절하게 분류된 요소신호들을 구분하는 방법을 제안하였다. 제안된 방법의 유효성을 검증하기 위하여 5종류의 음악과 2종류의 음성신호를 하나씩 선별하고 더하여 혼합신호를 생성하였으며, 자기부호화기의 중간 단계의 노드에서의 출력값을 적절한 군집화 방법을 적용하여 각 요소신호들을 분류하고 이를 합쳐 원음을 복원하였다. 분리된 음원들의 스펙트럼을 보았을 때 제안된 방법이 음원들의 요소신호의 분류에 매우 효과적이었음을 알 수 있었다.

II. 자기부호화기를 이용한 음원 분석

효율적인 입력신호의 특징을 추출하기 위하여 시간축 입력신호에 단구간 푸리에 변환(Short-Time Fourier Transform, STFT)을 수행하였다. 정수 k가 주파수 계수, 정수 m은 단구간 푸리에 변환의 시간축 프레임 번호일 때, 주파수성분의 크기 스펙트럼값 $X_{m,k}$은 다음의 식과 같이 계산된다.

$$X_{m,k} = \left| \sum_{n=0}^{N_f-1} x(n_2) \cdot e^{\frac{-2\pi i k n}{N_f}} \right|, \qquad (1)$$

$$n_2 = n + (m-1)N_s .$$

Eq.(1)에서 $x(n)$은 시간영역 신호이고, 정수 N_f는

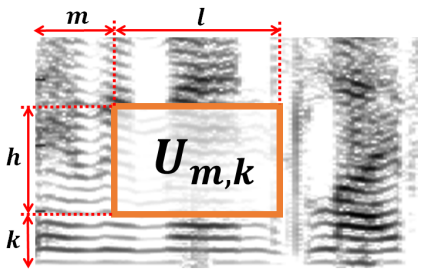

Fig. 1. Position of the convolutional window given the time-frequency indices to generate the input to the stacked autoencoder.

단구간 푸리에 변환을 수행하는 프레임의 샘플수, 그리고 N_s는 프레임을 취득하는 시작위치를 이동시키는 시간영역에서의 샘플의 수이다. 자기부호화기의 입력은 단구간 주파수 스펙트럼에서 일정한 크기의 직사각형의 창으로 구성하였으며, 이는 다음과 같이 정의된다.

$$\mathbf{U}_{m,k} = \begin{pmatrix} X_{m,k} & \cdots & X_{m+l-1,k} \\ \vdots & \ddots & \vdots \\ X_{m,k+h-1} & \cdots & X_{m+l-1,k+h-1} \end{pmatrix}. \qquad (2)$$

프레임 계수 m과 주파수 계수 k에서 시작되는 사각창은 Eq.(2)와 같이 행렬 $\mathbf{U}_{m,k}$로 나타낼 수 있으며, Fig. 1과 같이 시간축 계수 m부터 $m+l-1$까지의 l개의 프레임, 주파수축 계수 k부터 $k+h-1$까지의 h개의 주파수 성분을 취하는 $l \times h$ 크기의 직사각형으로 표현된다. 입력신호에서 충분한 수의 학습자료를 추출하기 위하여 시간축에서 하나의 프레임 단위, 주파수축에서도 하나씩 이동하면서 추출하였다. 자기부호화기 신경회로망의 학습의 입력으로 사용하기 위하여 $\mathbf{U}_{m,k}$를 다음과 같이 1차원 열벡터로 재구성하였다.

$$\mathbf{u}_{m,k} = \begin{pmatrix} X_{m,k} \\ X_{m,k+1} \\ \vdots \\ X_{m,k+h-1} \\ X_{m+1,k} \\ \vdots \\ X_{m+1,k+h-1} \\ \vdots \\ X_{m+l-1,k} \\ \vdots \\ X_{m+l-1,k+h-1} \end{pmatrix}. \qquad (3)$$

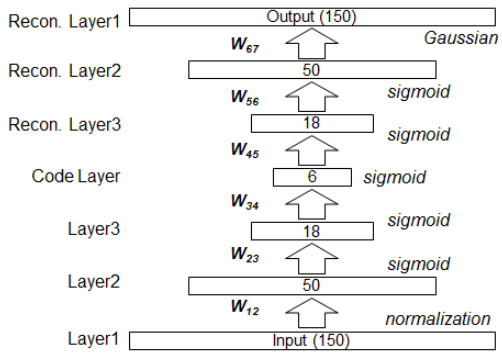

Fig. 2. The structure of autoencoder for audio source separation.

Eq.(3)에서 열벡터 $\mathbf{u}_{m,k}$는 행렬 $\mathbf{U}_{m,k}$의 열들을 하나씩 수직으로 연결하여 재구성한 벡터의 형태이다. 제안된 방법에서는 $l=5, h=30$으로 설정하여 150차원의 입력벡터가 하나의 학습샘플로 구성되었다. 일반적인 자기부호화기의 정의에 따라 신경회로망의 출력 역시 같은 차원의 벡터이다.

제안된 방법에서 사용한 복층구조 자기부호화기는 입력층과 출력층, 그리고 총 5개의 은닉층으로 구성하였으며, 각 은닉층의 노드의 수는 각각 50, 18, 6, 18, 50개이며, Fig. 2에 전체 구성을 나타내었다. 1-3층은 코드층(code layer)으로 이어지는 전향경로이며, 복원층(recon. layer)들은 입력신호를 복원하는데 사용된다. 입력층에서는 각 주파수별로 스펙트럼의 값의 표준편차가 1이 되도록 정규화하였다.

$$X_{m,k}^{(N)} = \frac{X_{m,k}}{\sigma_X(k)},$$
$$\sigma_X(k) = \sqrt{\frac{1}{M}\sum_{m=1}^{M} X_{m,k}^2}. \quad (4)$$

Eq.(4)에서 M은 하나의 입력 신호의 프레임의 개수이고 $\sigma_X(k)$는 이 프레임들을 이용하여 구한 주파수 k에서의 표준편차, $X_{m,k}^{(N)}$은 정규화된 스펙트럼 값이다. Fig. 2에서 두 개의 층 (i,j)를 연결하는 가중치 행렬은 $\mathbf{W}_{ij}$로 표현되었으며, 코드 층을 포함한 모든 중간 단계의 층에서는 sigmoid 활성화 함수를 사용하여 0과 1사이의 값이 출력되도록 하였으며, 이는 다음과 같은 계산식으로 정의된다.

$$\mathbf{o}^{(j)} = f\left(\mathbf{W}_{ij}\mathbf{o}^{(i)} + \mathbf{b}_{ij}\right),$$
$$j = i+1,\ \mathbf{o}^{(1)} = \mathbf{u}_{m,k}. \quad (5)$$

Eq.(5)에서 $\mathbf{o}^{(i)}$는 i층의 출력값이고, $f(\cdot)$은 sigmoid 활성화 함수, $\mathbf{b}_{ij}$는 편차벡터이다. 본 논문에서 사용한 자기부호화기는 Fig. 2와 같이 바로 위의 층에만 연결이 존재하기 때문에 항상 $j=i+1$이 성립한다. 최하위 층의 출력값 $\mathbf{o}^{(1)}$은 Eq.(3)에서 정의한 입력벡터 $\mathbf{u}_{m,k}$이다. 연결 가중치 $\mathbf{W}_{ij}$는 deep Boltzmann machine(DBM) 알고리즘으로 신경망의 출력과 목적값, 즉 자기부호과 신경망에서는 입력값과의 교차엔트로피(cross entropy)가 최소화되도록 학습하였다.[4,5]

DBM은 모든 입력이 양의 값을 가지는 것을 가정하기 때문에 입력값은 Eq.(4)와 같이 표준편차만으로 정규화하였고, 평균은 차감하지 않았다. 또한 자기부호화기의 마지막의 출력층은 입력값의 복원을 위하여 Gaussian 활성화 함수를 사용하였으며, Eq.(4)의 주파수별 표준편차 $\sigma_X(k)$를 곱하여 원래의 범위로 복원하였다. 마지막 층의 최적화함수는 자기부호화기의 정의에 따라 입력 스펙트럼과 신경망의 최종 출력 사이의 평균제곱오류이며, 목적값을 필요로 하지 않는 비교사 학습으로 분류된다.[5]

III. 음원의 분리 및 복원

실제 환경에서는 무수히 많은 종류의 음원이 존재하며, 같은 음원이라도 녹음환경 및 장비, 음원과 마이크의 거리 등의 여러 가지 조건에 의하여 선형적이거나 비선형적인 차이가 발생한다. 기존의 음원분리 알고리즘들은 별도의 학습 자료를 요구하지 않고 혼합신호만을 사용하여 음원 특징 모델을 학습하고 적절한 방법으로 음원을 분리하는 방식을 사용하였으며,[1,2] 본 논문에서도 자기부호화기를 혼합된 입력신호마다 개별적으로 학습함으로써 학습 자료가 필요 없는 음원분리 방법을 제안한다.

Fig. 3은 자기부호화기 학습에 기반하여 제안된 음원 분리방식의 블록도이다. 시간영역에서의 입력신호 $x(n)$은 두 개의 서로 다른 특성을 가진 음원들 $s_1(n)$과 $s_2(n)$의 합으로 가정되며, 이를 Eqs.(1)과 (4)

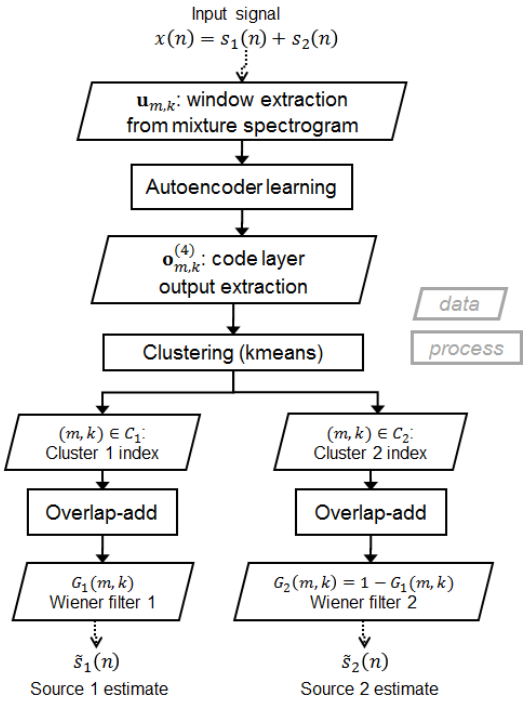

Fig. 3. Block-diagram of the proposed audio source separation method using stacked autoencoder (2 source case).

를 이용하여 단구간 주파수 스펙트럼으로 변환한다. 그리고 Eqs.(2)와 (3)을 적용하여 자기부호화기의 학습을 위한 벡터로 변형한다. Fig. 3은 두개의 음원이 혼합된 경우를 가정하여 도식화되었으나, 다수의 음원이 혼합되었을 경우에도 군집의 개수를 늘려 쉽게 확장이 가능하다. 다수의 음원들이 혼합된 입력신호가 자기부호화기에 의하여 학습되었을 경우, 자기부호화기 학습 결과에 따라 Fig. 2의 코드층에서 음원들의 차이를 효과적으로 모델링할 수 있으며, 이는 기존 연구에서 필기숫자 영상에 대하여 성능향상으로 입증되었다.[4,5]

본 연구에서는 각각의 단위 윈도우 $\mathbf{U}_{m,k}$가 서로 다른 음원들에 의해 생성되었다고 가정하고, 윈도우들을 서로 다른 성분으로 분류하기 위하여 k-means 군집화를 사용하였다. 입력벡터 $\mathbf{u}_{m,k}$이 q번째 클러스터에 속할 확률은 다음의 Bernoulli 확률식으로 정의된다.

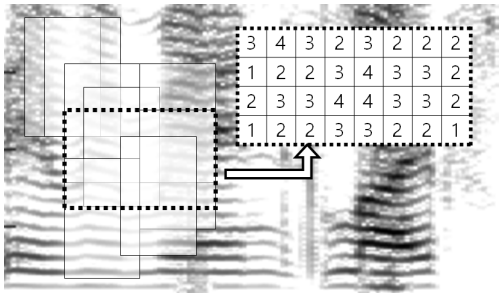

Fig. 4. Reconstruction of the original sources by constructing a mask from the clustering results. The left boxes are sampled windows that belong to a specific cluster. The numbers in the right table represents the number of overlapped windows to a specific position.

$$p(\mathbf{u}_{m,k} \in C_q) = \begin{cases} 1, & \text{if } \underset{p}{\operatorname{argmin}}\, D(\mu_p, \mathbf{o}_{m,k}^{(4)}) = q. \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Eq.(6)에서 C_q는 q번째 클러스터에 속한 윈도우들의 집합이며, μ_p는 p번째 클러스터의 대표벡터이다. 함수 $D(\cdot)$은 k-means의 정의에 따른 거리척도이고 일반적으로 벡터간의 L_2-norm, 즉 Euclidean 거리가 사용된다. $\mathbf{o}_{m,k}^{(4)}$는 입력벡터 $\mathbf{u}_{m,k}$에 대한 코드층의 6차원 출력벡터이며, 군집화를 위한 확률계산에 사용되었다.

각 클러스터에 해당되는 스펙트럼은 해당 클러스터에 속한 입력창을 중첩하여 구하며, Fig. 3의 6번째의 단계에 해당된다. 이 과정은 Fig. 4 에 도식화되어 있다. 왼쪽의 다수의 회색의 직사각형들은 하나의 클러스터에 속한 윈도우들을 중첩한 것이며, 점선으로 표시된 중첩된 임의의 사각형 영역을 확대한 오른쪽의 직사각형 격자에는 스펙트럼의 각각의 위치 (m,k)에서 몇 개의 입력창이 중첩되었는지 나타낸다. 이를 최대 가능한 중첩수인 $l \times h$로 나누어 프레임 m에서 주파수 성분 k가 클러스터에 속할 확률을 계산한다.

$$p(X_{m,k} \in C_q) = \frac{1}{lh} \sum_{\tau=1}^{l} \sum_{\omega=1}^{h} p(\mathbf{u}_{m-l+\tau,k-h+\omega} \in C_q) \,. \quad (7)$$

Eq.(7)은 입력신호에 대한 클러스터 q의 마스크이며,

이를 Wiener 필터 계수로 적용한다.

$$G_q(m,k) = p(X_{m,k} \in C_q). \quad (8)$$

원래의 신호는 단구간 주파수 스펙트럼에 Eq.(8)의 Wiener 필터 계수를 곱하고, 역 푸리에 변환을 사용하여 복원한다.

IV. 실험결과

제안된 방법의 성능을 평가하기 위하여 5개의 음악에서 하나를 선택하고, 2개의 음성신호 중 하나를 선택하여 혼합하여 입력신호를 생성하였다. 음성신호는 음성인식 표준으로 사용되는 TIMIT(Texas Instruments and Massachusetts Institute of Technology) 자료로부터 선택하였으며, 음악은 재즈, 드럼, 기타, 전자기타와 피아노 등 음색이 다른 5가지 일반 음악으로 선택하였다. 전체 가능한 조합은 $2 \times 5 = 10$가지이다. 음향신호들은 8 kHz의 샘플링 주파수로 변환되었고, 각 혼합신호별로 8 s를 추출하였다. 단구간 주파수 영역 스펙트럼 행렬은 단구간 푸리에 변환을 적용하여 구하였으며, 프레임 길이 40 ms, 시프트 길이 10 ms를 적용하였다. Eq.(1) 1에서 $h = 30$과 $l = 5$를 사용하였다. Eqs.(6)과 (7)에서 전체 클러스터의 수는 혼합된 음원의 개수인 2를 사용하여 2개의 원음을 복원하도록 하였다.

제안된 방법을 이용하여 분리한 음원분리결과를 NMF와 비교하였다. NMF 기반 음원분리 방법은 단일채널 음원분리에 널리 사용되는 방법으로, 단구간 주파수 스펙트럼에서 비음수 기저벡터를 추정하고, 그 기저벡터들의 선형결합으로 음원들이 혼합되어 있다고 가정하여 음원들을 분리한다.[2] NMF의 기저벡터의 수는 제안된 방법의 클러스터의 개수와 동일한 2개를 사용하였으며, 각각의 기저벡터에 대해 음원을 하나씩 추출할 수 있으므로 NMF의 결과로 하나의 입력신호가 2개의 음원으로 분리된다.

Fig. 5는 음성과 재즈음악의 혼합신호에 대한 음원분리 비교 결과이다. 전체 8 s의 입력에서 처음의 3 s의 스펙트럼만을 도식화하였다. 첫 번째 행의 두 개의 스펙트럼은 혼합되기 전의 신호들이고, 두 번째 행은 혼합된 신호의 스펙트럼이다. 세 번째 행은 NMF를 이용하여 얻은 Wiener 필터 $G_{m,k}^{NMF}$이며, 선택된 단구간 주파수 스펙트럼 요소는 어두운 색으로 표현되었다. 네 번째 행은 이를 이용하여 재합성한 스펙트럼이다. 마지막 행의 AE1과 AE2는 제안된 방법인 자기부호화기로 얻은 Wiener 필터 $G_{m,k}^{AE}$와 재합성 스펙트럼이다. NMF를 이용하여 구한 Wiener 필터는 일반적으로 시간축에서 변이가 크지 않았으며, 전체

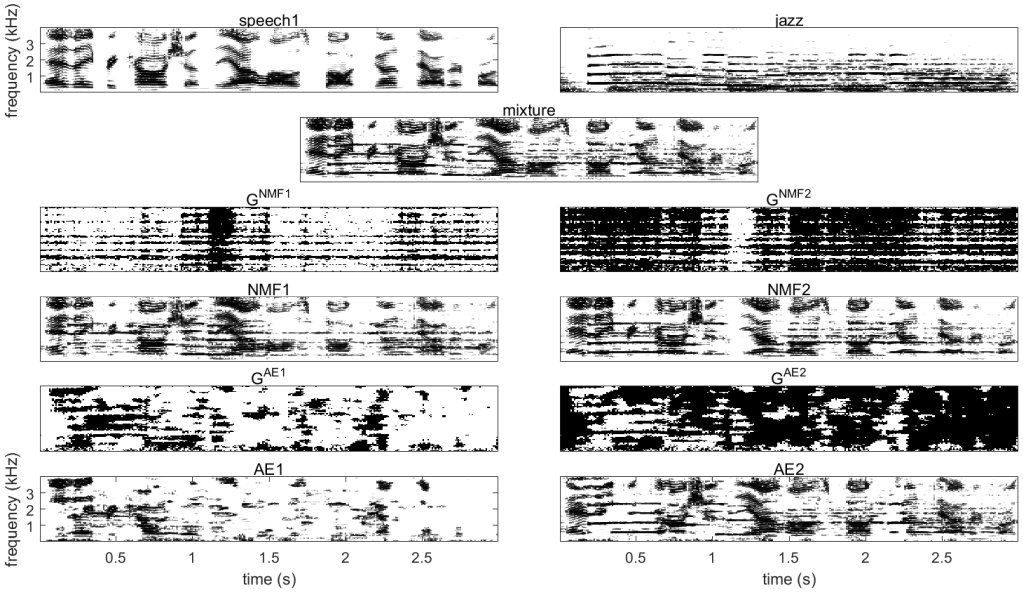

Fig. 5. Separation results of the proposed method and NMF for speech and jazz mixture input.

적으로 음성과 재즈음악이 두 번째 분리음원에 몰리는 결과가 나왔다. 제안한 방법은 0.2-0.8 s 구간에서는 재즈 음악과 음성을 잘 잘 구분해 내었으나, 그 외의 구간에서는 전체적으로 음성과 재즈음악이 역시 두 번째 분리음원에 같이 포함되었다.

Fig. 6은 음성과 드럼이 포함된 전자기타음을 분리한 결과이다. NMF는 시간영역의 주파수 성분 전체를 선택하였으나, 제안된 방법은 특히 저주파 구간에서 음성신호의 성분을 잘 선택하였다. 하지만 음악신호의 단구간 주파수 스펙트럼에서 반복되는 구조를 선택하는 데에는 두 방법 모두 적절하지 않았다. Fig. 7은 어쿠스틱 기타와 음성의 혼합신호이다. Wiener 필터와 복원된 음원에서 0.5 ~ 1.5 s의 저주파 구간에서는 두 가지 방법 모두 기타음을 잘 선택하였으나, 그 밖의 구간에서는 완전하게 구분하지 못했다.

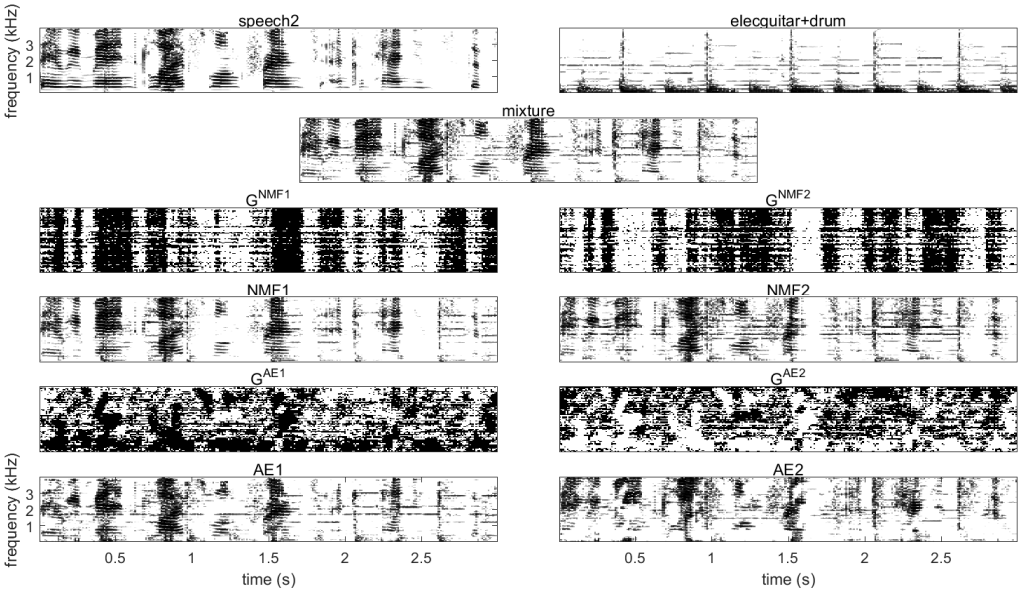

Fig. 6. Separation results of the proposed method and NMF for speech and electric guitar mixture.

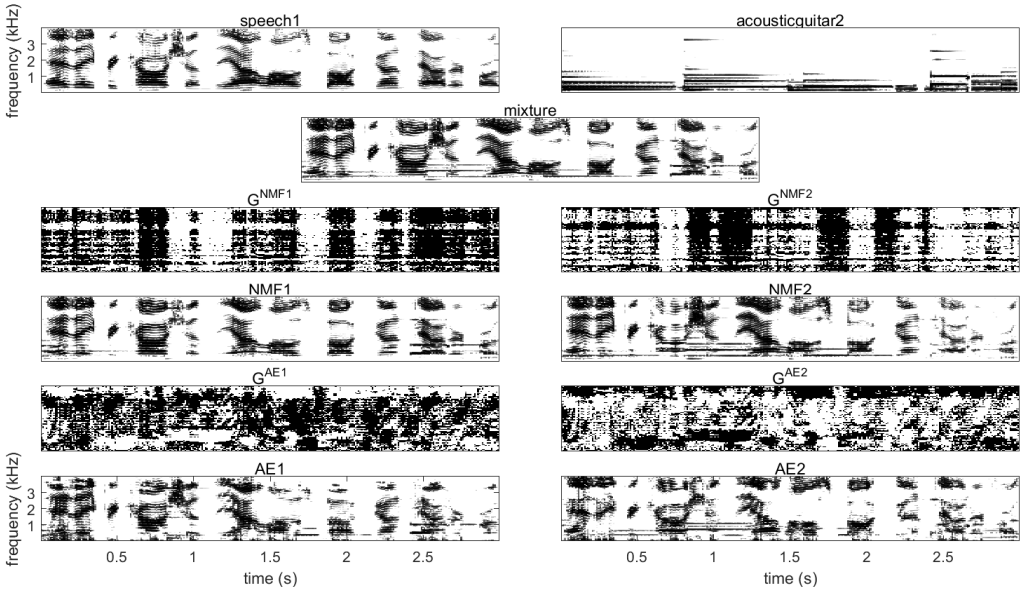

Fig. 7. Separation results of the proposed method in comparison to NMF for speech and acoustic guitar mixture input.

일반적으로, NMF로 얻은 Wiener 필터는 단구간 주파수 스펙트럼에서 안정적인 형태로 음원들을 추출하였으며, 제안된 방법은 윈도우의 위치를 중첩함으로써 Wiener 필터를 얻었기 때문에 매우 복잡한 형태로 나타났다. 복원된 신호를 청취한 결과, NMF 결과는 분리가 되지 않더라도 부드러운 소리가 재생되었으며, 제안된 방법은 분리성능은 조금 더 좋았으나 불연속적인 소리가 재생되었다. 음성신호를 추출하는 데에는 일반적으로 제안된 방법이 더 적합하였다.

V. 결 론

본 논문에서는 자기부호화기를 이용하여 두 가지 이상의 다른 특징을 가진 음원들의 분리에 적용할 수 있는 새로운 방법을 제안하였다. 혼합신호의 단구간 주파수 분석을 통하여 얻은 스펙트럼 행렬에 대하여 복층자기부호화기 학습을 통하여 서로 다른 음원들의 주요 특징이 학습되었다. 자기부호화 신경회로망의 중심의 부호층 출력은 각 음원들을 기술하는 특징들로 사용되었으며, 이를 적절한 군집하여 서로 다른 음원들을 구분하는데 사용하였다. 제안된 방법은 서로 다른 특징의 음원들을 재합성할 수 있었으며, 현재 실제 응용분야에 적용가능하도록 최적화 연구를 진행중에 있다.

감사의 글

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.R0126-15-1034, 채널 객체 융합형 하이브리드 오디오 콘텐츠 제작 및 재생기술 개발).

References

1. G. Hu and D. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” IEEE Trans. on Neural Networks **15**, 1135-1150 (2004).
2. B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, “Non-negative matrix factorization based compensation of music for automatic speech recognition,” Proc. Interspeech, 717-720 (2010).
3. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion,” JMLR **11**, 3371-3408 (2010).
4. G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science **313**, 504-507 (2006).
5. G. E. Hinton, S. Osindero, and Y. Teh, “A fast learning algorithm for deep belief nets,” Neural Computation **18**, 1527-1554 (2006).

저자 약력

▸ 장 길 진 (Gil-Jin Jang)

1997년 2월: KAIST 전산학과 학사
1999년 2월: KAIST 전산학과 석사
2004년 2월: KAIST 전산학과 박사
2004년 2월 ~ 2006년 8월: 삼성종합기술연구소 전문연구원
2006년 9월 ~ 2009년 10월: Postdoctoral Researcher, Univ. California, San Diego, USA
2009년 11월 ~ 2014년 2월: 울산과학기술대학교 전기전자컴퓨터공학부 조교수
2014년 3월 ~ 현재: 경북대학교 전자공학부 조교수