템플릿 기반 음향 신호 분리 기술 연구

권기수
서울대학교 전기정보공학부

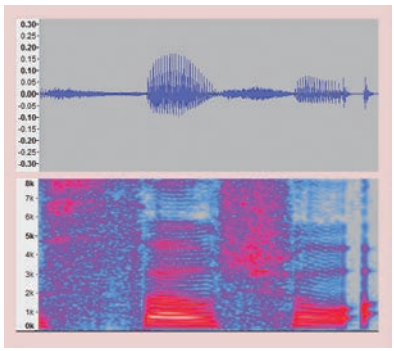
김남수
서울대학교 전기정보공학부

Ⅰ. 서 론

음향 신호는 사람의 청각 기관으로부터 인식이 되는 소리를 나타내며 보통 전기 신호로 표현하여 데이터화 한다. 우리 주변에 발생하는 자연적인 소리들은 모두 주파수라는 값을 가지고 있다. 주파수란 어떠한 진동이 얼마간의 시간 동안 주기적으로 발생하는 지를 말한다. 즉 주파수 값이 클수록 초 당 진동이 많음을 나타낸다.(사람의 귀는 보통 20Hz부터 18,000Hz의 주파수 대역을 인지 할 수 있다[1].) 이러한 음향 신호를 하나의 데이터처럼 보고 다양한 목적에 맞는 작업을 수행하게 된다. 이를 크게 음향 신호 처리라고 한다. 음향 신호 처리 분야 안에는 음향 위치 추정, 음질 향상, 실감 음향, 음악 정보 처리, 음원 분리, 음성 인식, 음성 합성, 음향 코딩, 화자 인식 등이 있다.(〈그림 1〉)

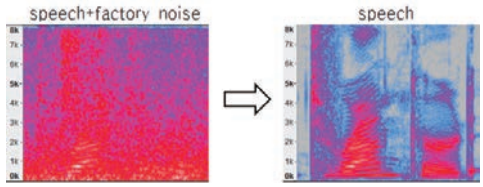
음향 신호는 크게 두 가지 형태로 표현이 가능하다. 첫 번째는 시간 축 표현으로 시간의 흐름에 따른 전기적 신호의 크기로 나타낸 것이다. 예를 들어 샘플링 레이트가 16kHz/s 라면 초당 16,000 개의 전기 신호값으로 이루어지게 된다. 〈그림 2〉의 위에 있는 그래프는 시간 축 음향 신호를 나타낸다. 크기가 음수와 양수 모두 존재하는 것을 볼 수 있다. 두 번째는 주파수-시간 축 표현으로 일정 시간 단위마다 각 주파수에서의 신호값이 얼마인지를 나타내게 된다.(빨간색일수록 큰 값임을, 파란색과 흰색은 작은 값을 의미) 즉 주파수-시간 축 데이터에서는 한 시간 프레임에 여러 차수의 값을 가지게 된다. 그림에서는 1Hz 부터 8,000Hz 까지 표현하고 있다. 시간-주파수 축 값은 첫 번째의 시간 축 데이터를 이용해서 쉽게 구할 수 있다. 이 때 short time fourier transform (STFT)을 이용하여 주파수-시간 값으로 변환을 하고 보통 이를 spectrogram이라고 한다. 음향 신호 처리에서는 두 형태의 데이터를 모두 이용할 수 있지만, 주파수-시간 축 신호가 가지는 여러 장점

〈그림 1〉 음향 신호 처리 분야의 세부 기술



〈그림 2〉 동일한 소리를 시간 축으로 표현 한 음향 신호 그래프(위)와 주파수–시간 축으로 표현 한 음향 신호 그래프(아래)

〈그림 3〉 피아노, 드럼 그리고 콘트라베이스 소리가 섞인 상황에서 목표로 하는 피아노 소리만 분리



〈그림 4〉 공장 소리가 존재하는 곳에서 녹음된 사람의 음성 신호(왼쪽)를 음질 향상 과정을 통해 공장 소리가 제거된 음성 신호(오른쪽)로 복원

들 때문에 이를 더 많이 사용되고 있다.

다양한 음향 신호 처리 분야 중 음원 분리(음향 신호 분리)는 다양한 소리가 존재하는 환경에서 특정 관심 있는 소리만을 복원 또는 추출해내는 기술을 나타낸다. 또는 반대로 원치 않는 소리를 제거하는 것을 의미한다. 이 때 관심 있는 소리는 한 종류가 아닌 여러 종류일 수도 있다. 예를 들어 〈그림 3〉처럼 피아노, 콘트라베이스, 드럼으로 이루어진 재즈 음악에서 드럼 소리와 콘트라베이스 소리를 제외한 피아노 소리만을 듣고 싶은 경우 해당 음악(관찰된 음향)에서 드럼과 콘트라베이스 소리만을 제거 하는 것이다. 음원 분리는 오랜 기간 음향 신호 처리 분야에서 중요하게 다뤄진 음질 향상(잡음 제거)에도 손쉽게 적용 가능하다는 이점이 있다[2–3]. 이때는 음성 신호가 목적 신호가 되고 그 외 소리 신호(ex. 공장 소리, 차량 소리, 키보드 소리 등)를 방해 신호 또는 잡음 신호로 보고 음원 분리를 수행하게 된다. 〈그림 4〉는 사람의 음성과 공장 잡음이 섞인 주파수–시간 축 데이터를 이용하여 오른쪽과 같이 잡음이 제거된 깨끗한 음성 신호만 있는 형태를 복원하게 된다. 음원 분리는 단순히 그 자체만으로도 음향 신호 처리에서 중요한 한 축을 이루고 있으며, 또는 다른 음향 신호 처리 분야의 전처리로 중요한 역할을 하고 있다. 예를 들어 음성 인식의 경우 음성 이외의 잡음이 존재하면 정확도가 떨어지기 때문에 일종의 전처리 역할로 음원 분리를 수행하여 음성 인식의 정확도를 높일 수 있다.

음원 분리(음향 신호 분리)는 다양한 소리가 존재하는 환경에서 특정 관심 있는 소리만을 복원 또는 추출해내는 기술을 나타낸다.

본 학회지에서는 다양한 음향 신호 처리 분야 중 음원 분리에 대해 자세히 다루기로 한다. 두 번째 섹션에서는 음원 분리에 대해 전반적으로 소개 하고, 세 번째 섹션에서는 최근 음원 분리 분야에서 주로 사용되는 알고리즘 하나를 중심으로 설명 한다.

Ⅱ. 음원 분리(음향 신호 분리)

음원 분리는 아래와 같은 근본적인 문제점이 있기에 수행하는데 있어 큰 어려움이 있다. 인간의 귀로 어떠한 소리를 들을 때 경험적으로 '어떤 종류의 소리들이 섞여 있다'를 쉽게 인지는 할 수 있지만, 정확하게 하나의 소리를 추정하기에는 어려움이 있다.(종류가 아닌 개별 주파수마다의 크기를 정확히 인지 나아가 분리하기란 거의 불가능하다.) 즉 인간은 '어떤 종류의 소리가 있구나' 정도는 쉽

게 인지하지만 정확히 해당 시간에서 어떤 소리의 크기가 얼마인지 까지는 추측하기가 어렵다. 이를 마이크가 하나인 단채널 상황(입력 신호가 한 가지)에서 수식으로 보면 $Y=S_1+S_2+\cdots+S_i$ 처럼 볼 수 있다. Y는 우리가 귀로 듣게 되는 소리(관찰된 소리)를 주파수-시간 축 데이터로 나타낸 것이고, S_i는 i번째 종류의 소리를 주파수-시간 축 데이터로 나타낸다. 여기서 우리는 Y밖에 알 수 없기에 나머지 각 소리의 크기를 아는 것은 불가능하다.(식의 개수보다 미지수의 숫자가 더 많다.) 인간이 경험적인 지식으로 소리의 종류를 추정하듯이, 기계적으로도 특정 소리의 데이터베이스를 사용하여 해당 소리의 사전 정보를 최대한 얻고 난 후 음원 분리를 수행해야 높은 성능을 기대할 수 있다.

음원 분리에서 단채널 상황이란 쉽게 말해 마이크가 하나만 있는 환경을 의미한다. 즉 어떠한 목적 소리를 추정하는 데 있어서 쓸 수 있는 입력값이 한 종류인 것이다. 그렇기 때문에 음향들의 공간적인 정보는 활용할 수 없다. 반대로 다채널 상황이란 마이크가 두 개 이상 있는 환경을 의미한다. 복수 개의 마이크로 인해 입력값은 그만큼 많아지고 마이크 위치 등에 따라 음향이 음원으로부터 마이크까지 오는 채널 등 다양한 공간 정보를 활용할 수 있게 된다. 본 학회지에서는 단채널 상황으로 한정하여 음원 분리를 설명하도록 한다.

음원 분리는 일반적으로 Blind source separation (BSS) 이라고도 일컬어지며, 이 분야는 최근 20년 이상 동안 다양한 분야에 중요한 역할을 수행해 왔다[4]. BSS는 음원에 대한 어떠한 정보도 없이 또는 최소한의 정보로 음원 분리를 수행하는 것을 목표로 한다. Independent component analysis (ICA)는 BSS 분야에서 매우 중요한 알고리즘으로서, 일반적으로 두 개 이상의 마이크 신호로부터 음원을 분리하게 된다. 그 이후 nonnegative matrix factorization (NMF), nonnegative tensor factorization (NTF), sparse component analysis, dictionary learning, empirical mode decomposition 과 같은 다양한 알고리즘이 음원 분리에 활발하게 쓰이게 된다[5-12]. ICA와 달리 NMF 등과 같은 알고리즘에서는 목적이 되는 음원에 대해 어느 정도 이상의 정보를 사전에 알아야 가능하다. 이러한 알고리즘은 소리 신호를 어떤 작은 최소한의 소리 단위들이 중첩 또는 더해져서 표현할 수 있다고 가정하고 있다.

본 학회지에서는 최근 음원 분리에서 활발하게 사용되고 있는 템플릿(template) 기반 음원 분리에 대해 자세히 설명하기로 한다. 템플릿 기반 알고리즘에는 NMF, NTF, dictionary learning, exemplar based approach 등이 있다. 이러한 알고리즘은 분리 과정 전에 훈련 과정을 필수로 수행하게 된다. 이 훈련 과정을 통해서 각 음원 종류마다 적절한 모델 또는 구성 요소(판형처럼) 등을 구하고 이를 분리 과정에 사용하게 된다. NMF는 행렬 인수분해 알고리즘 중 하나로 각 종류의 소리에서 적절한 부분(기저, basis)들을 구하여 음원 분리과정에 사용하게 된다. NTF의 경우 NMF와 매우 유사한 형태를 가지며 행렬이 아닌 3차원 데이터를 대상으로 하기에 시간 축 정보까지 활용하기에 적절하다. dictionary learning의 경우 다양한 접근 방법이 존재하지만 결국 여러 개의 비음수 인 작은 부분의 조합으로 설명하기에 NMF와 같은 프레임에서 볼 수 있다. exemplar 방법의 경우는 넓은 의미의 dictionary learning 방법에 속한다. 즉 템플릿 기반 방법들은 모두 사전에 구할 수 있는 목적 음원의 정보를 하나의 사전 속 단어처럼 활용하는 것을 의미한다. 이 부분은 때때로 compositional model 이라고 불리 우기도 한다[5]. 다음 섹션에서는 NMF를 통해 기본적인 템플릿 기반 알고리즘에 대해 설명을 하고, 이를 이용한 음원 분리 시스템에 대해 설명을 하도록 한다.

> NMF는 행렬 인수분해 알고리즘 중 하나로 각 종류의 소리에서 적절한 부분(기저, basis)들을 구하여 음원 분리과정에 사용하게 된다.

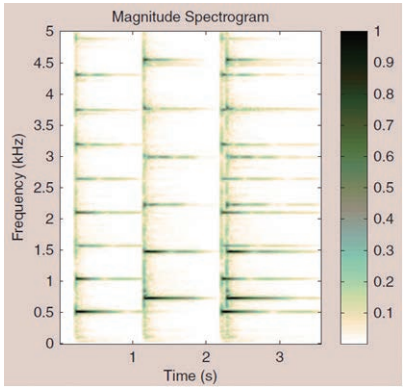
Ⅲ. NMF 기반 음원 분리

1. 조합 모델(compositional model)

앞서 언급한 템플릿 기반 방식은 사전에 활용 가능한 데이터를 이용하여 특정 모델을 만들고 이 모델을 목적에 맞게 사용하는 것을 의미한다. 특정 모델은 일종의 트

〈그림 5〉 여러 부품들의 조합으로 이루어지는 자전거

레이닝 과정을 통해서 구해질 수 있다. 여기서 말하는 특정 모델은 예를 들어 확률 분포를 이용한 통계 모델이 될 수 있다. 특정 종류의 데이터들을 가우시안 확률 분포로 모델링을 하여 이 데이터들이 보통 평균값이 얼마고 분산 정도는 얼마인지 사전에 알고, 이를 이용하여 원하는 작업에 사용하게 되는 것이다. 이러한 특정 모델로 음원 분리에서 최근 몇 년간 각광 받고 널리 쓰이는 것은 조합 모델(compositional model)이다[5]. 조합 모델 이란 말 그대로 어떠한 구성 요소들의 조합으로 이루어지는 모델을 나타낸다. 많은 자연적인 신호들은 보통 특정 부분들의 조합으로 나타낼 수 있다고 본다. 조합이 되는 부분들이 더하기 형태로 조합이 되어 어떤 실제 데이터를 표현하기 위해서는 부분들이 양수의 값을 가져야 한다. 즉 부분을 나타내는 값들 중 음수 부분이 존재한다면 이는 어떤 실제 데이터를 구성 할 때 삭제 또는 빼기의 특징을 같기 때문에 실제 자연 상태의 데이터의 특징과 부합하지 않기 때문이다. 이러한 조합을 설명하기 위해 자전거를 하나의 예로 들을 수 있다. 자전거를 하나의 실제 데이터로 본다면 이를 구성하는 적절한 부분들을 정의해야 한다. 자전거를 구성하고 있는 부분들을 보면, 첫 번째로 바퀴가 있다. 이 때 바퀴가 두 개 일수도 있고 외발 자전거처럼 한 개 일수도 있다. 두 번째로 프레임을 생각할 수 있다. 그



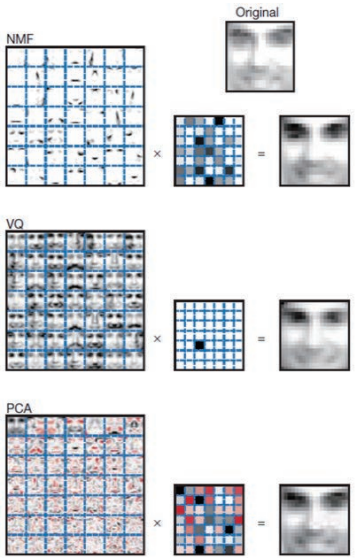
〈그림 6〉 음향 신호의 부분 및 조합 특성을 볼 수 있는 피아노 소리의 주파수–시간 축 그래프[5]

외에 페달, 안장, 조향기, 브레이크, 안전등, 바퀴 커버, 전방등, 경적기 등을 생각할 수 있다. 이러한 다양한 부분들이 조합 되어 자전거라는 형태를 이루게 된다. 당연히 어떤 자전거에는 전방등이 있을 수도 있고 없을 수도 있고 이러한 부분들로 다양한 자전거 형태가 표현 가능할 것이다. 즉 적절히 자전거 부분들을 정의한다면 수많은 자전거를 그 보다 적은 수의 부품의 유무로 표현 할 수 있다. 이러한 부분들을 표현에 따라 템플릿이라고 볼 수 있고 이러한 형태의 모델을 크게 템플릿 기반 방식 또는 조합 모델 방식이라고 한다. 예를 든 자전거 외에도 실제 음향 신호는 이러한 부분 기반의 조합 모델로 표현이 가능하다고 일반적으로 가정한다[5]. 예를 들어 〈그림 6〉을 보면, 피아노 소리에서 도 음과 미 음을 동시에 친 음향을 시간–주파수 축 영역에서 분석을 하면 단독으로 쓰인 도 음과 단독으로 쓰인 미 음의 시간–주파수 축 영역이 더해진 모양과 같음을 확인 할 수 있다. 이러한 템플릿 기반 방식 안에는 다양한 알고리즘이 존재 한다. 그 중에서도 이 학회지에서는 1999년도에 제안된 NMF를 중심으로 설명하기로 한다.

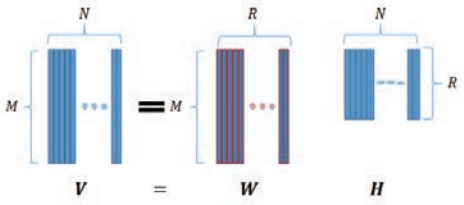
2. NMF (비음수 행렬 인수분해)

NMF는 말 그대로 음수가 없는 하나의 행렬을 두 개의 행렬의 곱 형태로 근사시키는 것을 의미한다.[12] 여기서 두 개의 행렬 또한 음수가 없는 형태이어야 한다. 여기서 음수가 없다는 것이 매우 중요한 의미를 가지게 된다. 음수가 없기 때문에 목적으로 하는 데이터를 구성하는 부분

들이 모두 양수의 형태로 더하기 특징을 갖게 해주고, 또한 이 부분들의 조합비를 나타내는 행렬 부분도 양수이기 때문에 부분들이 해당 데이터에 '존재 한다' 또는 '존재하지 않는다'의 의미를 갖게 되기 때문이다. 이러한 이유로 NMF에 사용되는 데이터는 비음수 형태이어야 한다. NMF가 가지는 비음수의 의미를 이해하기 위해 〈그림 7〉을 참고 할 수 있다. 〈그림 7〉은 사람 얼굴 이미지를 데이터로 쓰며 목표가 되는 사람의 얼굴 이미지를 복원하는 것을 목표로 한다. vector quantization (VQ)의 경우 사전에 가지고 있는 데이터들 중 가장 목표 얼굴에 근접한 한 데이터로 복원을 하게 된다. 즉 가능한 많은 사람들의 실제 얼굴 이미지를 여러 개 저장한 상태에서 가장 유사한 이미지 하나만을 사용하게 된다. 인코딩 행렬 부분을 보면 하나를 제외한 모든 데이터는 0이고 단 하나의 데이터만 1의 크기로 쓰이게 된다.(검은색은 쓰인 정도가 1임을, 빨간색은 쓰인 정도가 −1임을, 마지막으로 흰색은 전혀 쓰이지 않았음을 의미.) principal component analysis (PCA) 알고리즘의 경우 각 기저 벡터들이 직교인 특징을 가지고 있고 비음수라는 제한이 없기 때문에 각 기저 모양이 실제 존재하는 사람 얼굴과는 다른 모습을 가지게 된다. 또한 인코딩 또한 음수 부분이 존재하여 어떤 기저는 빼주게 되는 등의 복잡한 형태를 갖게 된다.



〈그림 7〉 특정 얼굴 이미지 데이터를 복원 시 NMF, PCA 그리고 VQ의 사용 예[12]



〈그림 8〉 NMF에서 데이터, 기저행렬, 인코딩 행렬의 관계

하지만 NMF의 경우 기저의 형태에서 의미를 찾을 수 있다. 어떤 기저는 사람의 코 부분의 윤곽을, 어떤 기저는 눈매의 윤곽을, 어떤 기저는 입술 부분을 나타낸다고 볼 수 있다. 인코딩 또한 모두 양수로 각각의 기저가 얼마나 쓰이는 지에 대해 비교적 간단하게 조합의 모습을 추측할 수 있게 된다. 즉 NMF에서는 테트리스 게임처럼 어떤 목표 데이터를 표현하는데 있어서 필요한 기저를 쌓아서 더하는 형태를 보이게 된다. 이는 앞서 설명했듯이 실제 자연에 존재하는 데이터 들이 쌓아서 더하는 형태를 가진다고 가정 할 수 있기 때문에 더욱 적절하다고 볼 수 있다.

이러한 NMF를 수식적으로 간단히 나타내면 $V \approx WH$처럼 볼 수 있고, 여기서 V는 데이터 행렬을, W와 H는 각각 기저행렬과 인코딩 행렬을 나타낸다. 〈그림 8〉은 위 행렬의 관계를 쉽게 그래프로 보여주고 있다. 데이터는 M차수를 가지고 있고, 이 데이터를 R개의 기저(각 열 벡터)들로 표현하게 된다. N개의 데이터가 있다면 인코딩 행렬 또한 N개의 벡터를 가지고 있게 된다. W의 각 열 벡터들이 하나의 기저를 의미하게 된다. 이를 구하기 위해 아래와 같은 목적함수를 정하게 된다.

$$f(W,H) = D(V|WH) \qquad (1)$$

여기서 $D(\bullet \mid \bullet)$는 일종의 거리 또는 다이버전스(divergence)로 두 데이터가 같을수록 0의 값을, 차이가 클수록 큰 값을 갖게 된다. 이를 위해 보통 유클리디언 또는 Kullback−liebler divergence(KL−D)가 쓰이게 된다. V를 표현하는 WH를 구하기 위해 이 목적함수의 값이 최소화 되는 방향으로 최적화를 수행하게 된다. 하지만 위 목적함수가 두 값에 대해 convex하지 못 하기 때문에 단순한 방법으로 최적값을 구할 수 없다. 이 때문에 alternative 업데이트 방식을 택하게 된다. 즉 W를 업데이트 할 때는 H를 상수처럼 고정하고 수행하는 것이다.

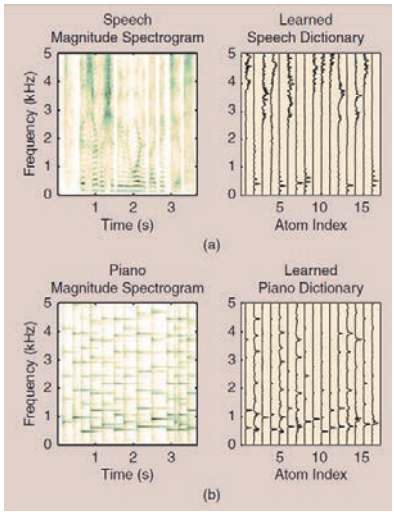
이를 최적화하기 위해 다양한 방법이 제안 되었고, 최초의 논문 [12]에서 제안된 multiplicative update rule로 최적화 업데이트 수식을 구하게 되면 아래와 같다.

$$H \leftarrow H \otimes \frac{W^T \frac{V}{WH}}{W^T 1} \quad (2)$$

$$W \leftarrow W \otimes \frac{\frac{V}{WH} H^T}{H^T 1} \quad (3)$$

앞서 설명했듯이 두 값을 구하기 위해 수식 (2)와 (3)을 번갈아 여러 번 수행하게 된다. 위 수식에서 $\cdot^T$는 transpose를 의미하고, 1은 적절한 크기의 모든 원소가 1인 행렬 또는 벡터를 나타낸다. 위 수식들을 여러 번 교대로 수행하다가 수렴했다고 판단 될 때 업데이트를 마치고 그 결과를 최적값으로 사용하게 된다.

이러한 알고리즘 또한 여러 가지 단점을 가지고 있다. 음향 신호는 다른 신호와 달리 시간의 흐름에 따른 정보가 매우 중요하다. 하지만 기본적인 NMF에서는 이 정보가 전혀 반영되지 않기 때문에 한계가 있다. 이를 해결하고자 연속된 시간 데이터 여러 개를 하나의 데이터로 구성하여 NMF과정을 수행하기도 하며 또는 convolutive NMF와 같은 알고리즘이 제안되기도 했다[13]. 또 한 가지 큰 단점은 모든 음향 종류의 기저는 직교하지 않는 다는 것이다. 이는 달리 말하면, 다른 종류의 음향의 기저들이 조금씩은 비슷하다는 것이다. NMF 입장에서는 한 종류의 음향을 해당 기저들의 선형 합으로 표현하게 되는데, 다른 종류의 음향 데이터 중 일부가 표현 가능하다는 것이다. 이는 바로 복원 에러로 이어지며 큰 성능 저하의 원인이 된다. 이를 해결하고자 기존에 discriminative NMF 라는 이름으로 다양한 알고리즘이 제안되었다[14]. 마지막으로 NMF 목적함수가 convex 하지 못하다는 점이다. 이로 인해 NMF로부터 얻어지는 기저들은 초기값에 매우 민감한 모습을 보인다. 일반적으로 무작위 비음수 값으로 초기값을 준 경우 안정적인 성능이 나오기 때문에 주로 무작위 초기값 방법이 쓰이며, 최근에는 singular value decomposition을 활용한 초기화 방법이 제안되기도 했다[15].



〈그림 9〉 (a) 음성 신호로부터 얻어진 음성 기저들(dictionary)
(b) 피아노 신호로부터 얻어진 피아노 기저들(dictionary)[5]

3. NMF 기반 음원 분리

음향 신호에 NMF를 적용하기 위해서는 일반적으로 supervised 상황이 가정 된다. 즉 테스트 음향 신호에 존재하는 음향 종류를 사전에 알고 있는 것이다. 예를 들어 어떠한 재즈 음악을 대상을 할 때는 음원 분리 작업 전에 그 음악 파일은 피아노, 콘트라베이스 그리고 드럼 소리가 있다는 정보를 미리 안다고 가정하는 것이다. 그렇다면 훈련 과정에서 각 악기 소리의 기저행렬을 훈련 DB와 NMF 알고리즘을 이용하여 구할 수 있다. 이렇게 구해진 각 악기의 기저행렬에서 각 컬럼 벡터는 해당 악기의 작은 단위의 소리라고 볼 수 있다. 〈그림 9〉에서는 음성 데이터베이스(database, DB)로부터 얻어진 음성 기저들과 피아노 DB로부터 구해진 각 기저들의 모양을 보여주고 있다. 음성 기저 중 몇 개는 하모닉 성분이 있음을 확인할 수 있다. 이러한 작업을 위해서는 주파수-시간 축의 데이터의 절대값을 V로 만들어 사용하여야 한다. NMF 는 비음수 데이터만 쓰일 수 있기 때문이다. 이렇게 사전에 구한 각 악기의 기저행렬을 하나의 행렬로 아래와 같이 만들게 된다.

$$W = [W_1 \; W_2 \; \cdots \; W_i] \quad (4)$$

위에서 W_i는 i번째 음원으로부터 얻은 기저행렬을 의미한다. 예를 들어 각 음원에서 10개의 기저를 구하고 5개

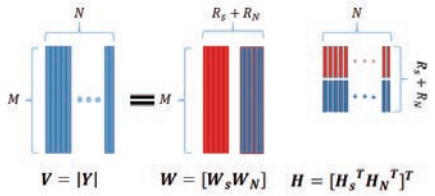
의 음원을 고려한다면, 총 50개의 기저 벡터들이 하나의 큰 기저행렬을 이루게 되는 것이다. 음원 분리 과정 전에 이런 기저행렬을 만들어 놓으면 모든 사전 훈련 작업이 끝나게 된다. 아래의 분리 과정에서는 쉬운 이해를 돕기 위해 음원이 두 가지 있다고 가정하겠다. 즉 사용되는 기저는 $W=[W_S W_N]$으로 W_S는 목적 신호를 표현하게 될 기저행렬을, W_N은 제거하고자 하는 방해 신호를 표현하게 될 기저행렬을 나타낸다.

실제 분리 과정에서는 관찰 신호 y를 시간-주파수 영역의 데이터 Y로 변환한 후 $V \approx WH$가 되도록 H를 구하게 된다. 여기서 Y는 STFT를 적용한 관찰된 신호이고, $y=s+n$이라고 가정한다. s는 목표 신호를, n은 제거하고자 하는 신호를 의미한다. 여기서 H는 〈그림 10〉과 같은 형태로 볼 수 있다. 기저행렬 부분에서 왼쪽 반은 목표 음원의 기저들을 나타내고, 나머지 오른쪽 부분은 잡음 기저들을 나타낸다. 당연히 목표 음원과 잡음의 기저 개수(R_S,R_N)는 자유롭게 정할 수 있기에 둘의 숫자가 달라도 문제가 없다. 그림에서 인코딩 행렬 부분은 위와 아래 두 부분으로 나눌 수있다. 위 부분은 목표 음원 기저가 얼마나 쓰였는지를 의미하고, 아래 부분은 잡음 기저가 얼마나 쓰였는지 나타낸다. 즉 $W_S H_S$는 목적 신호의 복원값으로 볼 수 있다. 이를 구현하기 위해 훈련 과정과 마찬가지로 같은 목적 함수를 사용하게 된다.

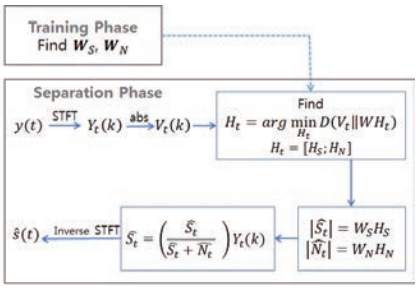
$$f(H) = D(V|WH) \tag{5}$$

훈련과정과 달라진 점은 변수가 H하나로 줄었다는 것이다. 일반적으로 이 상황에서는 convex하기 때문에 최적값을 찾을 수 있다.(특정 상황에서는 그렇지 않다) H를 구하기 위해 수식(2)를 사용하고 이 또한 한 번이 아닌 여러 번의 과정을 통해 최종 H를 구하게 된다.

이렇게 구한 값을 이용하여 각 음원의 추정치로 사용할 수 있지만 이는 완벽하지 않다. NMF 알고리즘 특성 상 어느 크기 이상의 오차가 존재하게 되고 이를 최소화하기 위해 아래와 같은 이득 함수를 사용하게 된다.



〈그림 10〉 분리 과정에서의 NMF 관계식을 그래프로 표현한 것



〈그림 11〉 NMF 기반 음성 향상 과정을 표현한 블록 다이어그램

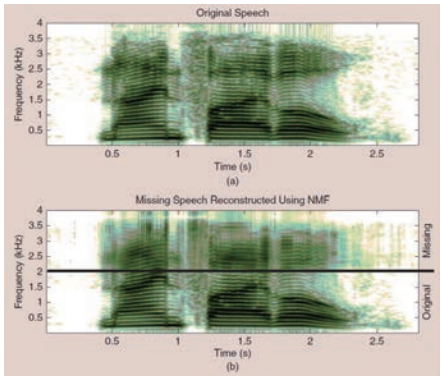
$$G = \frac{W_S H_S}{W_S H_S + W_N H_N} \tag{6}$$

NMF 알고리즘으로부터 1차 추정한 $W_S H_S$ 값은 불완전하지만 상대적으로 이 값들을 이용한 각 음원의 비는 그 보다 더 정확하다고 볼 수 있다. 그렇기 때문에 위 이득함수 G수식처럼 각 음원의 1차 추정치를 이용하여 해당 음원의 비를 사용하게 된다. 최종적으로 아래와 같이 각 음원의 추정 주파수-시간 값을 구할 수 있다.

$$\hat{S} = G \otimes Y \tag{7}$$

위의 NMF를 이용한 음원 분리를 〈그림 11〉에 블록 다이어그램으로 정리하였다. 즉 훈련 과정에서 각 음원의 기저행렬을 저장하고 음원 분리과정에서 기저행렬을 이용하여 관찰 신호를 최대한 적절하게 복원하는 과정을 거치는 것으로 볼 수 있다. 〈그림 11〉의 블록 다이어그램에서는 전체 관찰된 신호를 이용하여 배치 형태로 하는 것이 아니고 특정 시간 단위로 들어온 값을 이용하여 STFT 기준으로 한 시간 프레임 단위로 분리 과정을 수행하고 있다. 그렇기 때문에 각 패러미터에 t가 붙어 시간을 의미

훈련 과정에서 각 음원의 기저행렬을 저장하고 음원 분리과정에서 기저행렬을 이용하여 관찰 신호를 최대한 적절하게 복원하는 과정을 거치는 것으로 볼 수 있다.



〈그림 12〉 원본 신호 (a)에서 1~2000Hz까지의 정보만을 이용하여 그림 (b)처럼 2000~3800Hz까지 복원을 한 결과

하고 있다. 마지막의 $\hat{s}(t)$는 시간 축 데이터로 표현된 목표 신호이다.

4. NMF의 음원 분리 외에 적용 분야

NMF와 같은 방식으로는 exemplar-based 방식과 NTF 등이 있다. 또한 이 분야는 dictionary learning 또는 compositional model 등으로 표현하기도 한다. NMF와 같은 알고리즘들은 위에서 설명한 음원 분리 외에도 다양한 곳에 활용 될 수 있고 활발하게 쓰이고 있다.

음향 신호 처리 분야 중 band width extension 문제는 일종의 missing data를 다루는 것으로 볼 수 있다[5]. 기존의 유선 상의 통신으로는 전송의 효율성 등의 이유로 8kHz/s 정도의 샘플링 레이트 환경에서 동작을 하였다. 이 환경에서는 사람의 음성 정보가 3.4kHz 까지 존재하게 되어 음성 명료도가 약간 저하되고 고음 부분이 없는 먹먹한 느낌의 음성이 들리게 된다.[1] 이러한 문제를 해결하기 위해 샘플링 레이트가 16kHz/s 인 정도의 정보를 복원하는 여러 방법들이 제안되어 왔다. 이 또한 NMF 등으로 처리할 수 있다. 목표가 되는 음성을 16kHz/s 인 파일에서 주파수-시간 축의 정보로 변환을 한다. 그렇게 하면 주파수 부분에서 1/2 부분은 저주파 대역(~4kHz)을 표현하고, 나머지 1/2부분은 4kHz에서 8kHz 부분을 표현하게 된다.(실제로 통신 상 6.8kHz 정도의 주파수 대역까지 표현하지만 설명의 편의 상 8kHz까지 표현하는 것으로 가정한다.) 이 데이터를 이용하여 여러 기저들을 만들고 만들어진 기저 또한 앞의 반은 저주파 대역을 표현하는 기저들이라고 볼 수 있고, 나머지 반은 고주파 대역을 표현하는 부분이라고 볼 수 있다. 이렇게 구한 저주파 대역 기저들을 이용하여 실제 입력 된 8kHz/s 신호를 복원하고 이때 얻은 인코딩 값을 그대로 고주파 대역 기저들에 곱해주면 복원하고자 하는 고주파 대역의 음성이 구해지게 된다. 이렇게 band width extension 외에 missing data를 처리하는 다양한 분야에 적용 가능하다. 〈그림 12〉 는 논문 [5]에서 예로 든 것으로, (a)는 실제 3.8kHz 대역까지 표현된 목표 음성 신호를 주파수-시간 축 데이터로 나타낸 것이다. 이 원본 신호에서 2kHz~3.8kHz 대역의 신호를 강제로 없앤 후 1~2000Hz 까지의 신호만을 이용하여 복원한 결과이다. 100% 복원까진 아니지만 음성의 하모닉한 성분 등을 적절히 잘 복원할 수 있음을 확인할 수 있다.

NMF는 반향 제거에도 쓰일 수 있다. 음향 신호는 작은 공간 또는 반향이 심한 곳에서는 원 신호 외에 반향이라는 것이 발생하게 된다. 이러한 반향이 일정 크기 이하일 때는 듣기 좋을 수도 있지만 너무 큰 경우 잡음처럼 생각되어 음성에서는 명료도를 떨어트리는 원인이 된다. 이 또한 NMF를 이용하여 원음을 표현하는 기저들과 반향을 나타내는 기저들로 따로 훈련을 하여 어느 정도 제거를 할 수 있다. 이외에도 화자 인식 분야에도 쓰이고 있으며, 음향 신호가 아닌 이미지 신호 분야에서 또한 얼굴 인식 등의 목적으로 활발하게 쓰이고 있다.

Ⅵ. 향후 연구 및 결론

음향 신호 처리 기술은 청각 기관으로 인지 가능한 신호를 다루며 소리 신호를 활용하여 다양한 목적에 맞게 가공 및 처리한다. 대표적인 음향 신호 처리 기술에는 음악 정보 처리, 음성 인식, 음성 향상, 음성 합성, 음원 분리 등이 있다. 이러한 목적을 위해 기존에 통계모델 기반 방식이 활발히 제안되고 사용되었고, 최근 20년 전부터 다양한 템플릿 기반 방식이 제안되고 주목 받아 왔다. 템플릿 기반 방식은 사전 정보로 특정 종류의 음원을 사용하여 모델을 얻고 이를 실제 과정에서 사용하는 것을 말한다. 템플릿 기반 방식 중 본 학회지에서는 NMF를 중심

으로 설명을 하였다. NMF 는 특정 종류 신호에서 그 신호를 구성하는 작은 단위의 부분들을 찾아내어 활용하는 것이다. 자연적인 신호는 대부분 작은 단위의 조합으로 나타낼 수 있기에 NMF 는 다양한 신호에서 탁월한 성능을 보였다.

이를 음원 분리 분야에 적용하게 되면 관찰된 소리에 몇 종류의 음원이 섞여 있더라도 간단히 분리 과정을 수행할 수 있다. 훈련 과정에서 사전에 알고 있는 음원 종류의 DB를 사용하여 각 종류마다의 기저행렬을 구하고 이 기저행렬들을 하나의 큰 행렬로 이어줘 사용하면 관찰된 소리에서 특정 원하는 소리만을 다시 복원할 수 있다. 음원 분리 외에도 NMF는 band width extension, 반향 제거, 화자 인식 등에 다양한 분야에 활용 될 수 있고 쓰이고 있다.

더 높은 성능의 음원 분리를 위해서 NMF를 중심으로 시간 흐름의 정보와 음원 간의 차이를 더 극명하게 하는 기저들의 활용 등을 고려하여 연구해야 할 것이다.

참고문헌

[1] R. Larsen and R. M. Aarts. Audio bandwidth extension: application of psychoacoustics, signal processing and loudspeaker design. John Wiley & Sons, 2005.

[2] K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based speech enhancement using bases update," IEEE Signal Process. Lett., vol. 22, no. 4, pp. 450-454, Apr. 2015.

[3] K. Kwon, J. W. Shin, H. Y. Kim, and N. S. Kim, "Discriminative nonnegative matrix factorization using cross-reconstruction error for source separation," Proc. of ISCA Interspeech, 2015.

[4] Naik, Ganesh R., and Wenwu Wang, eds. Blind Source Separation: Advances in Theory, Algorithms and Applications. Springer, 2014.

[5] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional model for audio processing," SPM, Mar. 2015.

[6] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. Hoboken, NJ: Wiley, 2009.

[7] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," IEEE Trans. Audio, Speech, Lang. Processing, vol. 19, no. 7, pp. 2067-2080, 2011.

[8] Y.-C. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classification," Pattern Recognit. Lett., vol. 26, no. 9, pp. 1327-1336, 2005.

[9] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," IEEE Trans. Audio, Speech, Lang. Processing, vol. 18, no. 3, pp. 550-563, 2010.

[10] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies,"Sparse representations in audio & music: From coding to source separation," Proc. IEEE, vol. 98, no. 6, pp. 995-1005, 2009.

[11] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in Proc. IEEE Spoken Language Technology Workshop, 2012, pp. 313-317.

[12] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," Nature, vol. 401, pp. 788-791, 1999.

[13] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," IEEE Trans. on Audio, Speech, and Language process., vol. 15, no. 1, Jan. 2007.

[14] K. Kwon, J. W. Shin, and N. S. Kim, "Target source separation based on discriminative nonnegative matrix factorization incorporating cross-reconstruction error," IEICE Transactions on Information and Systems, Vol. E98.D, No. 11, pp. 2017-2020, 2015

[15] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," Pattern Recognition, vol. 41, no. 4, pp. 1350-1362, 2008.

권기수

• ~2011년 2월 서울대학교 전기정보공학부 졸업
• 2011년 3월~현재 서울대학교 전기정보공학부 석박통합과정

〈관심분야〉
음성 신호 처리, 음향 신호 분리

김남수

• 1988년 2월 서울대학교 전자공학과 졸업
• 1990년 2월 한국과학기술원 전기공학과 석사
• 1994년 8월 한국과학기술원 전기공학과 박사
• 1998년 3월~현재 서울대학교 교수

〈관심분야〉
음성 인식, 음성 향상, 음성 합성, 머신러닝, 인공지능, 실감 음향, 음원 분리