실시간 오디오 업믹싱 시스템을 위한 비음수 행렬 분해 기반의 단일채널 배경 잡음 추출 기법

이석진
경기대학교
Sjlee6@kgu.ac.kr

Monaural Ambient Sound Extraction for On-line Audio Upmixing System based on Nonnegative Matrix Factorization

Seokjin Lee
Kyonggi University

요 약

본 논문에서는 비음수 행렬 분해 (NMF) 기법을 이용하여 단일 채널에서 배경음 성분을 추출하는 알고리즘에 대해 서술한다. 이러한 배경음 성분 추출은 오디오 업믹싱 시스템을 고려하여 개발되었으며, 기존의 연구를 통하여 분리된 배경음 신호가 서라운드 채널 혹은 상방향 채널에 적용될 경우 청취자의 공간감을 향상시킬 수 있다는 사실이 이미 확인된 바 있다. 다만 기존의 기법은 음향 신호를 모두 축적하여 일괄적으로 처리해야 한다는 단점이 있어, 스트리밍 시스템이나 디지털 신호 프로세서 등을 이용한 시스템에서 사용될 수 없는 단점이 있다. 본 논문에서는 이를 해소하기 위하여 실시간 비음수 행렬 분해 기법을 이용한 배경음 추출 시스템을 고안하여 실험하였다. 실험 결과 실시간 배경음 추출 기법이 신호의 후반부에서는 원하는 대로 동작하나, 초중반에 기저가 과도하게 설정되는 문제점이 있음을 확인할 수 있었으며, 이에 대한 해결이 향후 연구 과제가 될 것이다.

1. 서론

음향 녹음 및 재생 기기의 등장으로 청취자가 가정에서 오디오를 향유하게 된 이래, 음질뿐만 아니라 공간감에 대해서도 보다 충실하고 현장감이 있는 소리를 들을 수 있기를 바라는 요구가 항상 존재해 왔다. 이에 따라 과거 모노채널 오디오 시스템에서부터 시작하여, 스테레오 시스템, 그리고 5.1 채널 서라운드 시스템까지 다채널 오디오 시스템이 발전하면서, 청취자에게 보다 실감나는 음향 재생 성능을 제공할 수 있게 되었다.

보다 향상된 다채널 스피커 시스템의 성능을 즐기기 위해서는, 재생해야 할 음향 신호 또한 스피커 시스템에 맞도록 구성되어야 한다. 특히 최근에는 5.1 채널 서라운드 이상의 스피커 시스템에 대한 연구 개발이 지속되면서, 다양한 다채널 오디오 시스템이 소개되고 있다. 이러한 상황 속에서 보다 적은 채널 개수의 음향 신호를 다채널 시스템에 맞도록 재구성하거나, 혹은 최근에 개발되는 다양한 오디오 시스템에 적합한 컨텐츠를 확보하기 위하여 오디오 업믹싱 기법이 개발되고 있다.

특히, 5.1 채널 서라운드의 좌/우 서라운드 채널이나, 상방향 스피커가 있는 시스템에서 상방향 레이어 (top layer) 채널의 신호들은 주로 뚜렷한 음상을 맺도록 하기 보다는 넓은 음장감을 제공하는 데에 초점을 두고 있다. 이에 따라 다채널 업믹싱 기법에 있어서 주요한 역할을 하는 부분 중 하나가 배경음 분리 기법이다.

최근 독일 Fraunhofer 연구소의 한 연구결과에 따르면, 비음수 행렬 분해 (Nonnegative Matrix Factorization: NMF) 기법을 이용하여 단일채널에서 배경음에 가까운 소리들을 분리하였으며, 해당 기법으로 분리한 배경음을 이용하여 업믹스 시스템에 적용하였을 경우 기존 기법 대비 향상된 청취평가 결과를 얻을 수 있었다 [1].

위의 연구 결과가 좋은 성능을 보여주기는 했으나, 기존 NMF 기법의 경우 전체 음원을 한꺼번에 처리해야 한다는 단점을 가지고 있다. 즉, latency 에 민감한 스트리밍 데이터 시스템이나, 홈시어터와 같이 DSP 를 사용한 시스템에서는 사용할 수 없는 알고리즘이다.

본 논문에서는 위와 같은 비음수 행렬 분해에 기반한 실시간 배경음 추출 기법에 대해 논의하며, 특히 실시간 비음수 행렬 분해에 기반한 기법을 도출한다. 이를 토대로 기존의 일괄적 비음수 행렬 분해 기법에 의해 추출된 결과와 비교 분석한 후, 향후 연구 방향에 대해 논의한다.

2. 비음수 행렬 분해 기법에 기반한 배경음 추출 기법

2.1. 비음수 행렬 분해 기법

비음수 행렬 분해 기법이란 다음과 같이 $K\times N$ 크기의 비음수 행렬 $\mathbf{V}$ 를 $K\times R$ 크기의 비음수 행렬 $\mathbf{W}$ 와 $R\times N$ 크기의 행렬 $\mathbf{H}$ 의 곱으로 나타낸 후, 각 행렬을 추정하여 분해하는 기법이다[2, 3].

$$\mathbf{V}=\mathbf{WH}+\mathbf{E} \tag{1}$$

일반적으로 오디오 신호처리 시스템에서 행렬 $\mathbf{V}$ 는 음향 신호의 magnitude spectrogram 을 사용하며, 이 경우 행렬 $\mathbf{W}$ 는 주파수 영역 기저를, 행렬 $\mathbf{H}$ 는 시간 영역 기저를 나타낸다.

위와 같은 비음수 행렬 분해 모델에서 각 기저 행렬, 즉 $\mathbf{W}$ 와 $\mathbf{H}$ 를 추정하는 기법에는 여러 방법이 있으나, 주로 많이 사용되는 기법은 다음과 같이 multiplicative update 방법에 의해 순차적으로 추정하는 기법이다[4].

$$\mathbf{W}\leftarrow\mathbf{W}\otimes\left[\left(\mathbf{V}\mathbf{H}^{T}\right)\varnothing\left(\mathbf{W}\mathbf{H}\mathbf{H}^{T}\right)\right] \tag{2}$$

$$\mathbf{H}\leftarrow\mathbf{H}\otimes\left[\left(\mathbf{W}^{T}\mathbf{V}\right)\varnothing\left(\mathbf{W}^{T}\mathbf{W}\mathbf{H}\right)\right] \tag{3}$$

여기서 $\otimes$는 Hadamard product 로 행렬 원소 간의 곱셈을 의미하고, $\varnothing$ 는 원소 간의 나눗셈을 의미한다.

2.2. 비음수 행렬 분해에 기반한 배경음 분리 기법

비음수 행렬 분해는 오디오 신호의 희박 표현 (Sparse Representation)에 의거하여, 해당 음악 신호를 의미를 가진 여러 개의 음표 성분으로 분리하는 방법으로 주로 사용된다. 이 과정에서 희박 성분으로 잘 표현되지 않는, 전 주파수 영역에 걸쳐 퍼져 있는 소리의 성분들은 기저 행렬로 잘 표현되지 않는 결과를 보인다. 이를 이용하여, 비음수 행렬 분해로 분석하고 남은 나머지 신호들을 따로 추출하면 배경음 성분을 추출해 낼 수 있다[1]. 단, 여기서 추출되는 배경음 성분은 기존의 공간적인 배경음과 조금 다른 성질을 가진다. 추출된 배경음 성분은 단일 채널의 주파수 구조를 통해서 얻어낸 배경음 성분이며, 따라서 공간적인 분포와는 관계없이 주파수 영역에 걸쳐 고르게 분포되는 성분들을 배경음으로 간주하여 추출해내기 때문이다.

배경음 성분을 추출해내기 위하여, 먼저 위의 비음수 행렬 분해 결과를 이용하여 배경음 성분의 크기 성분을 먼저 다음과 같이 계산한다[1].

$$\left|\mathbf{A}_{kn}\right|=\beta_{kn}\cdot\left(\mathbf{V}-\mathbf{WH}\right)_{kn} \tag{4}$$

여기서 $\left|\mathbf{A}_{kn}\right|$ 는 추출할 배경음 성분의 k 번째 주파수 및 n 번째 시간 윈도우의 스펙트럼 크기 성분을 나타낸다. 여기서 크기 성분은 항상 양의 값을 가져야 하지만, $\mathbf{V}-\mathbf{WH}$ 의 성분은 항상 양의 값을 가지지 않는다. 따라서, 양의 값을 가지도록 의 가중치를 다음과 같이 설정한다[1].

$$\beta_{kn}=\begin{cases}\gamma\ :\ \left(\mathbf{V}-\mathbf{WH}\right)_{kn}<0\\ 1\ :\ otherwise\end{cases} \tag{5}$$

여기서 γ 는 -1 과 0 사이의 상수 값으로 설정한다.

추출할 배경음 성분의 위상 정보는 입력 신호의 위상 정보를 그대로 이용한다.

3. 실시간 배경음 추출 기법

3.1. 실시간 비음수 행렬 분해 기법

앞서 언급된 바와 같이, 기존의 비음수 행렬 분해 기법은 모든 데이터를 축적한 뒤 일괄적으로 계산해야 한다는 단점이 있으며, 이를 극복하기 위하여 실시간으로 기저 행렬을 추정하는 알고리즘들이 개발되고 있다. 그 중 하나가 재귀적 최소 자승법 (Recursive Least Squares: RLS)의 원리를 이용하여 개발된 실시간 비음수 행렬 기법 알고리즘이다[5, 6].

실시간 비음수 행렬 분해 기법은 매 n 번째 시간 윈도우마다 다음의 연산을 통해 주파수 기저 행렬 및 시간영역 기저 벡터를 추정한다. 먼저 n 번째 시간 윈도우의 시간영역 기저벡터는 다음과 같이 계산된다[5, 6].

$$\mathbf{h}(n)=\left[\mathbf{W}^{\dagger}(n-1)\mathbf{v}(n)\right]_{+} \tag{6}$$

여기서 $\mathbf{h}(n)$ 은 $R\times1$ 크기의 시간영역 기저 벡터이며, $\mathbf{v}(n)$ 은 $K\times1$ 크기의 n 번째 시간 윈도우의 magnitude spectrum 데이터를 나타낸다. 또한 † 는 Moore-Penrose pseudo inverse matrix 를 나타내며, $[\ ]_{+}$ 는 0 보다 작은 값을 0 이 되도록 연산하는 반파 정류 연산을 의미한다.

위와 같이 추정된 시간영역 기저 벡터의 값을 이용하여, 주파수 기저 벡터는 다음과 같이 추정된다[5, 6].

$$\mathbf{k}(n)=\left(\mathbf{P}(n-1)\mathbf{h}(n)\right)/\left(\lambda+\mathbf{h}^{T}(n)\mathbf{P}(n-1)\mathbf{h}(n)\right) \tag{7}$$

$$\mathbf{P}(n)=\lambda^{-1}\mathbf{P}(n-1)-\lambda^{-1}\mathbf{k}(n)\mathbf{h}^{T}(n)\mathbf{P}(n-1) \tag{8}$$

여기서 (7), (8)의 연산 결과를 이용하여 매 k 번째 주파수 빈 마다 다음의 연산을 통해 주파수 기저를 추정한다.

$$\xi_{k}(n)=v_{k}(n)-\mathbf{w}_{k}(n-1)\mathbf{h}(n) \tag{9}$$

$$\mathbf{w}_{k}^{T}(n)=\left[\mathbf{w}_{k}^{T}(n-1)+\mathbf{k}(n)\xi_{k}(n)\right]_{+} \tag{10}$$

여기서 $v_{k}(n)$ 는 $\mathbf{v}(n)$ 의 k 번째 주파수 성분 값을 나타내고, $\mathbf{w}_{k}(n)$ 는 $\mathbf{W}(n)$ 행렬의 k 번째 행 벡터를 나타낸다. 그리고 망각 인자 λ 는 일반적으로 1 에 가까운 값을 가지며, 과거 값에 대한 가중치를 나타낸다.

3.2. 실시간 배경음 분리 기법

위와 같이 추정된 기저 행렬 및 벡터 값을 이용하여, 배경음 성분의 크기 $a_{k}(n)$ 은 다음과 같이 계산된다.

$$a_{k}(n)=\beta_{k}(n)\cdot\left(v_{k}(n)-\mathbf{w}_{k}(n)\mathbf{h}(n)\right) \tag{11}$$

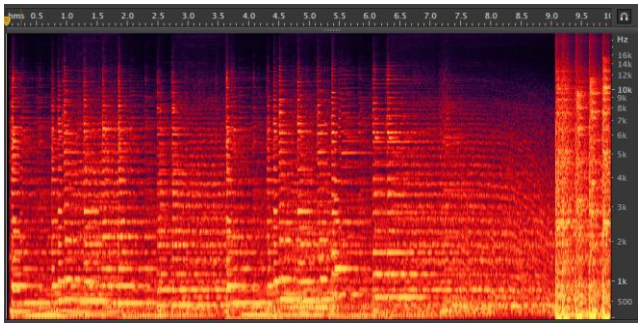

그림 1. 입력 신호의 스펙트로그램

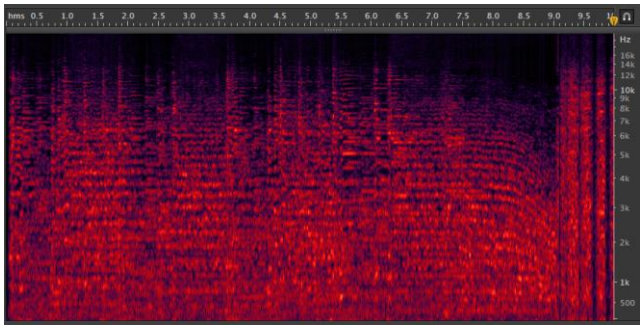

그림 2. 일괄 비음수 행렬 분해 기법을 통해 추출된 배경음 (R=80)

매 시간 윈도우 결과 사이의 변화가 너무 큰 경우 시간적 불연속에 의한 잡음이 들리는 경우가 있으므로, 이를 방지하기 위하여 다음과 같이 시간 영역 평활화(smoothing) 작업을 추가로 진행한다.

$$a_k(n) \leftarrow (1-\eta)\cdot a_k(n-1) + \eta \cdot a_k(n) \tag{12}$$

기존의 배경음 추출 기법과 마찬가지로, 추출되는 배경음 신호의 위상 값은 입력 신호의 위상 값과 동일한 값을 사용한다.

4. 실험 결과 및 토의

본 논문에서 소개 및 도출된 배경음 추출 기법의 동작을 살펴보기 위하여, 음악 신호에 대하여 배경음 추출 실험을 진행하였다. 실험 대상으로는 RWC database [7]의 popular music database 중 1 번 음원을 선정하였으며, 스펙트럼 분석은 2048 길이의 Hamming 윈도우를 이용하여 50% 겹침(overlap)을 통해 수행되었다. 일괄 비음수 행렬 분해 기법의 기저 개수는 80 개를 사용하였으며, 실시간 비음수 행렬 분해 기법의 기저 개수는 80 개와 TBD 개를 사용하였다. 실시간 비음수 행렬 분해 기법의 망각 인자 λ 는 1 로 설정하였고, 배경음 분리에 이용되는 시간 평활화 인자 η 는 0.5 로 설정하였다.

그림 1 은 해당 실험 음원의 처음 10 초간 크기 스펙트럼을 보여주고 있으며, 그림 2 는 분리된 배경음의 크기 스펙트럼을 나타내고 있다. 그림 2 의 결과를 보면 비음수 행렬 분해 기법을 이용해 추출된 배경음이 처음 의도된 대로 주파수 영역에 걸쳐 퍼져 있는 성분이 잘 추출된 것을 확인할 수 있다.

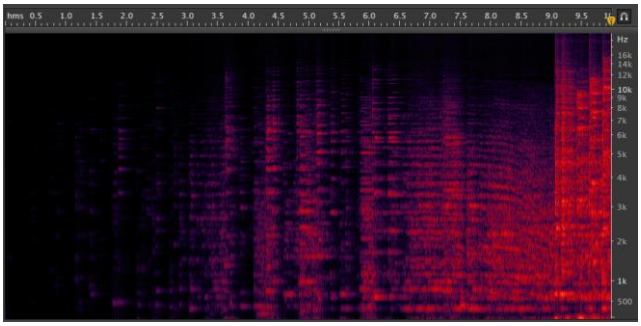

그림 3. 실시간 배경음 추출 알고리즘을 통해 추출된 배경음 (R=80)

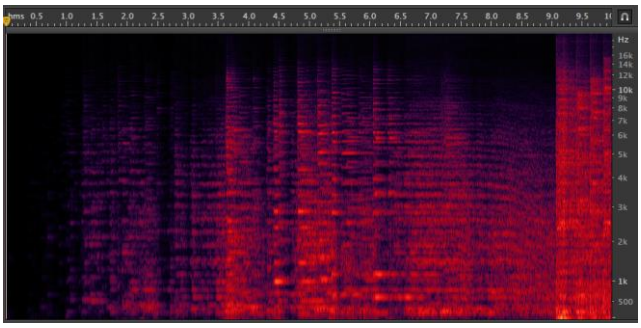

그림 4. 실시간 배경음 추출 알고리즘을 통해 추출된 배경음 (R=40)

그림 3 은 실시간 배경음 추출 알고리즘을 통해 추출된 배경음 결과의 크기 스펙트럼을 보여준다. 후반부의 추출된 배경음은 비교적 유사한 결과를 보여주지만, 중반부까지 추출된 배경음의 에너지가 과도하게 적은 것을 볼 수 있다. 이는 처음부터 끝까지 동일한 데이터의 크기를 가지는 일괄 비음수 행렬 분해 기법에 반해, 실시간 비음수 행렬 분해 기법의 경우 과거부터 현재까지의 데이터를 가지고 분석하는 효과를 가지기 때문에, 중반까지는 데이터의 특성에 비해 기저의 개수가 과도하게 많이 설정되는 효과를 야기한다. 따라서, 배경음 성분까지 희박 표현에 의한 기저 행렬로 추정되어, 결과적으로 추출된 배경음 성분이 없어지는 결과를 야기하게 된다.

이는 그림 4 의 결과에서 한번 더 확인할 수 있다. 그림 4 는 그림 3 의 결과와 동일한 실험 결과이지만 단순히 기저의 개수만 80 에서 40 으로 줄인 결과이다. 실시간 배경음 추출 알고리즘의 경우 기저의 개수가 적을 때 보다 많은 에너지의 배경음을 얻을 수 있으며, 이는 특히 초반 및 중반부의 신호에서 더욱 확실하게 나타난다. 따라서, 향후 연구를 통해 기저의 개수를 시간에 따라 다르게 설정하여 배경음을 추출하는 기법의 개발이 필요하다.

5. 결론

본 논문에서는 단일 채널의 음향 신호에서 비음수 행렬 분해 기법을 통하여 배경음 성분을 추출하는 알고리즘에 대해 살펴보았다. 또한, 이를 보완하여 일괄 처리가 아닌 실시간으로 배경음 성분을 추출할 수 있도록 새로운 알고리즘을 도출하였다. 도출된 알고리즘은 실시간 비음수 행렬 분해 알고리즘을 통하여 음향 신호를 분석하고, 분석되지 않는 나머지 신호를 배경음 성분으로 추출함으로써, 주파수 영역에

걸쳐 퍼져있는 에너지를 가지는 성분을 추출하도록 한다.

실제 음악 신호를 이용하여 배경음 추출 실험을 진행한 바, 일괄 비음수 행렬 분해 기법이 배경음 추출을 성공적으로 수행할 수 있음을 살펴볼 수 있었으며, 동시에 실시간 배경음 추출 기법 또한 그에 준하는 동작을 수행함을 확인할 수 있었다. 단, 신호의 분석 과정에서 신호의 초반 및 중반에 기저의 개수가 과도하게 설정되는 문제점이 있었으며, 이러한 문제를 해결하는 것이 향후 연구 과제가 될 것이다.

6. 참고문헌

1. C. Uhle, A. Walther, O. Hellmuth, and J. Herre, " Ambience Separation from Mono Recordings using Non-negative Matrix Factorization," *Proc. AES 30th International Conference on Intelligent Audio Environments*, Mar. 2007.
2. D. D. Lee and H. S. Seung, " Learning the Parts of Objects by Non-negative Matrix Factorization," *Nature*, pp. 788-791, Aug. 1999.
3. D. D. Lee and H. S. Seung, " Algorithms for non-negative matrix factorization," *Proc. Advances in Neural Information Processing 13 (NIPS' 2000)*, pp. 556-562, 2001.
4. A. Cichocki, R. Zdunek, A. H. Pahn, S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multiway data analysis and blind source separation*, John Wiley & Sons, Chichester, 2009.
5. S. Lee, S. H. Park, K.-M. Sung, "On-Line Nonnegative Matrix Factorization Method Using Recursive Least Squares for Acoustic Signal Processing Systems," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Science*, vol.E94-A, no. 10, pp. 2022-2026, Oct., 2011.
6. S. Lee, " RLS-based on-line sparse nonnegative matrix factorization method for acoustic signal processing systems," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Science*, vol. E96-A, no. 5, pp. 980-985, May 2013.
7. Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "RWC Music Database: Popular, Classical, and Jazz Music Database", *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pp.287-288, Oct. 2002.