한국음향학회지 제28권 제8호 pp. 697~705 (2009)

# 블라인드 방식의 리듬 음원 분리

# Blind Rhythmic Source Separation

김 민 제*, 유 지 호**, 강 경 옥*, 최 승 진**
(Minje Kim*, Jiho Yoo**, Kyeongok Kang*, Seungjin Choi**)

*한국전자통신연구원 방통융합미디어연구부, **포항공과대학교 컴퓨터공학과
(접수일자: 2009년 10월 30일; 채택일자: 2009년 11월 16일)

본 논문에서는 단일 채널 다성 음악에서 리듬 악기 신호를 블라인드 (blind) 방식으로 추출하는 방법을 제안한다. 상업적으로 판매되는 음악 신호는 대부분 2개 이하만의 혼합된 채널 형태로 사용자에게 제공되는 반면, 그 혼합 채널 신호에는 각각 가창 음원 (vocal)을 비롯한 많은 종류의 악기가 포함되어 있는 형태이다. 따라서, 혼합 신호의 개수가 음원 개수와 같거나 더 많은 상황을 가정하는 기존의 음원 분리 방법처럼, 혼합 환경이나 신호의 통계적 특성을 모델링하는 것 보다는, 특정 음원의 고유 특성을 활용하는 것이 이처럼 적은 개수의 혼합 신호만을 가지고 있는 환경 (underdetermined)에 더욱 적합하다. 본 논문에서는 다른 화성 악기와 혼합되어 있는 상황에서 리듬 악기 음원만을 추출하는 것을 목표로 한다. 비음수 행렬 인수분해 (NMF: Nonnegative Matrix Factorization)의 변형된 알고리즘인 비음수 행렬의 부분적 공동 분해 (NMPCF: Nonnegative Matrix Partial Co-Factorization)가 입력 행렬의 시간적인 속성과 주파수적인 속성에서 다양한 관계성을 분석하기 위해 활용된다. 또한 특정 시간 단위로 입력 신호를 파편화 (segmentation)하고, 파편들에서 반복적으로 발생하는 성분을 리듬 악기가 공통적으로 포함하고 있는 특성이라고 가정한다. 본 논문에서 제안하는 방법은 일반적으로 받아들여질 수 있을 정도의 성능을 보여주지만, 기본적으로는 사전 정보를 활용하는 타악기 음원 분리 방식보다 우수하지는 않다. 그러나 블라인드 방식의 특성상, 사전 정보를 획득하기에 용이하지 않은 경우, 또는 사전 정보와 현격히 다른 리듬 악기가 연주되는 경우 등에 보다 유연하게 대응할 수 있다.

**핵심용어:** 비음수 행렬 분해, 비음수 행렬의 부분적 공동분해, 리듬 음원 분리, 음악 정보 검색
**투고분야:** 음향 신호처리 분야 (1.2)

An unsupervised (blind) method is proposed aiming at extracting rhythmic sources from commercial polyphonic music whose number of channels is limited to one. Commercial music signals are not usually provided with more than two channels while they often contain multiple instruments including singing voice. Therefore, instead of using conventional modeling of mixing environments or statistical characteristics, we should introduce other source-specific characteristics for separating or extracting sources in the underdetermined environments. In this paper, we concentrate on extracting rhythmic sources from the mixture with the other harmonic sources. An extension of nonnegative matrix factorization (NMF), which is called nonnegative matrix partial co-factorization (NMPCF), is used to analyze multiple relationships between spectral and temporal properties in the given input matrices. Moreover, temporal repeatability of the rhythmic sound sources is implicated as a common rhythmic property among segments of an input mixture signal. The proposed method shows acceptable, but not superior separation quality to referred prior knowledge-based drum source separation systems, but it has better applicability due to its blind manner in separation, for example, when there is no prior information or the target rhythmic source is irregular.

***Keywords:*** Blind Source Separation, Nonnegative Matrix Factorization, Nonnegative Matrix Partial Co-Factorization, Rhythmic Source Separation, Musical Information REtrieval
***ASK subject classification:*** Acoustic Signal Processing (1.2)

책임저자: 김 민 제 (mkim@etri.re.kr)
305-700 대전광역시 유성구 가정로 138번 한국전자통신연구원
방송통신융합연구부문 방통융합미디어연구부
(전화: 042-860-1361; 팩스: 042-860-5479)

# I. 서 론

모노 또는 스테레오 신호로부터 음원을 분리하는 것은 음악 정보 검색 (MIR: musical information retrieval) 분야에서 깊이 있게 연구되고 있는 주제 중 하나이다. 그 이유는, MIR의 많은 연구 주제들이 일반적인 혼합 신호에 더해서 단독 음원이 확보되었는지 여부에 따라 성능이 좌우되기 때문이다. 원본 단독 음원을 확보하기 힘든 대부분의 경우, 성능이 우수한 음악 음원 분리 결과물은 원본 단독 음원에 대한 좋은 대안이 될 수 있다. 예를 들어, 자동 악보 생성 (automatic music transcription), 음악 유사도 분석 (musical similarity analysis), 허밍에 의한 질의 (query by humming), 음악 분위기/장르 분류 (music mood/genre classification) 등 다양한 MIR 분야들이 음원 획득에 의해 성능이 향상될 수 있다. 또한, 고품질 노래방 서비스 또는 Music 2.0과 같은 객체 기반 오디오 서비스의 경우, 자동화된 음악 음원 분리 도구를 활용할 수 있다면, 그 시장 영역을 보다 확대할 수 있다.

음악 음원 분리 (MSS: Musical Source Separation)는 수득한 혼합 신호의 개수가 음원의 개수보다 적은 (underdetermined) 환경에서의 대표적인 음원 분리 (BSS: Blind Source Separation) 주제이다. 보통 MSS는 전통적인 BSS와는 다른 형태의 음원에 대한 가정이 필요한데, 이는 혼합 환경이나 음원의 통계적 특성을 모델링하는 BSS 방법이 MSS에서는 그대로 적용되기 힘들기 때문이다. 그 이유는 전술된 바와 같은 혼합 신호 개수의 제한과 같은 문제 뿐 아니라, 마이크와 음원의 고정된 배치 또는 녹음실의 반향 환경 등이 정해져 있다는 전통적인 BSS의 가정이 다양한 사운드 효과를 사용하는 음악 콘텐츠의 제작 과정에는 적용될 수 없는 경우가 많기 때문이기도 하다.

입력 비음수 행렬을 2개의 또다른 비음수 행렬의 곱으로 인수분해하는 비음수 행렬 분해 (NMF: Nonnegative Matrix Factorization) [6, 7] 는 많은 연구 분야에서 좋은 성능을 보여주었다. 그것은 NMF가 기본적으로 입력 데이터를 부분 기반 표현 (parts-based representation) 방식으로 표현해 주기 때문이다. 부분 기반 표현 방식은 사람의 두뇌가 정보를 처리할 때도 나타나는 주요한 현상이다. NMF의 이러한 부분 기반 또는 희소 (sparse) 표현 방식은 MIR 분야에서도 역시 의미있는 진보를 이루어냈다. NMF는 타악기 연주 신호 또는 단일 악기의 다성 음악 연주 신호를 자동으로 악보화 하는 것에 사용되었으며 [9], 음향 신호 분류기의 특징 추출 (feature extraction) 방식으로도 활용되었다 [1]. 그러나, 본 논문에서 더 주목하는 NMF의 흥미있는 사용예는 MSS 분야에서 찾을 수 있다. NMF는 두 세 가지의 악기로 이루어진 간단한 MIDI 신호를 대상으로 의미있는 음원 분리 성능을 보였으며 [2, 3, 5, 10], 지지 벡터 기계 (SVM: Support Vector Machine)와 연동하여 상업 음악으로부터 타악기 신호를 분리하는 노력 또한 이루어졌다 [4]. 이와는 별도로, NMF의 확장된 형태인 비음수 행렬의 부분적 공동 분해 (NMPCF: Nonnegative Matrix Partial Co-Factorization) 방식을 이용하여 분류기 학습 과정 없이 타악기 사전 정보를 도입하는 타악기 음원 분리 방식 및 관련된 블라인드 형태의 알고리즘 역시 소개되었다 [12, 13].

비록 단일 채널 음악 음원 분리에서 NMF가 효과적인 도구로 사용될 수 있다고 하더라도, 부분 기반 표현 방식은 다양한 종류의 악기 음원에 대해 언제나 만족스러운 성능을 제공하는 것은 아니다. 특히 악기 조합이 2개 이상이 되거나, 가창에 의한 사람 목소리가 혼합되는 경우 등에 대해서는 NMF의 음악 신호에 대한 여러 가지 가정들이 잘 들어맞지 않게 된다. 예를 들어서, 악기의 개수가 몇 개인지 미리 알아야 한다는 점은 논외로 하더라도, 복수의 음을 연주하는 것이 가능한 악기의 연주를 분석하기 위해서는, NMF는 최소한 몇 개의 음정이 연주되고 있는지를 미리 알 필요가 있다. 또한, NMF는 선형 인수분해를 수행하기 때문에, 현악기 연주 시 사용되는 벤딩 (bending) 또는 슬라이딩 (sliding)과 같은 연속적인 악기 음정의 움직임, 또는 발음에 의해 달라지는 가창 신호의 주파수 특성 등을 분석하기에는 적합한 방식이 아니다.

그러나, 전술된 화성 악기와는 달리, 대부분의 리듬 악기의 경우, 주파수 특성이 변하지 않는다는 가정을 할 수 있다. 예를 들어 피아노와 같은 화성 악기가 노래 전체에서 지속적으로 연주된다고 하더라도, 노래에 사용되는 수 십 개의 모든 피아노 음정이 노래의 전 구간에 반복적으로 나타나지는 않는다. 반대로 리듬 악기는 그 연주 특성상, 몇 가지 서로 다른 음색 (timbre)을 가지는 타악기류의 악기가 반복적으로 연주되는데, 이 때 악기들은 음높이를 거의 가지지 않는 충격성 (impulsive) 전 주파수 대역 잡음 신호이거나, 화음 성분을 가지더라도 음 높이가 서로 다른 타격 시점에 따라 변하지 않는다는 특징을 가진다. 주파수 영역의 특성이 변하지 않는 악기군이 반복적으로 연주된다는 리듬 악기의 특성은 본 논문에서 제시하는 주요한 가정이다. 또한, 이러한 특성은 선형 분해 방식에 적절하다.

본 논문에서는 NMF 계열의 분석을 수행한 뒤 일부 성분 (component)들은 리듬 악기 음원에 배당하고, 나머지 성분들은 화성 악기 음원에 배당하는 일반적인 NMF의 사용 방식을 그대로 따른다. 그러나, 드럼 연주로만 이루어진 사전 정보가 활용 가능하지 않다는 좀 더 어려운 상황을 목표로 한다. 이러한 상황을 고려하여, 입력 단일 채널 혼합 신호를 보다 짧은 몇 개의 동일 길이 파편 (segment)으로 나누고, 일반적인 NMF 대신 [8]에서 제시된 고정 효과 분석 NMF (FFX-NMF: Fixed-effects Analysis NMF) 방식을 도입하여, 여러 개의 파편들에 걸쳐서 공통적으로 나타나는 성분과, 각각의 파편에 개별적으로 나타나는 성분을 따로 학습한다. 이 때 전술된 리듬 악기의 반복성에 의해, 리듬 악기 성분이 공통 성분으로 표현되며, 반대로 화성 악기는 개별 성분으로 표현될 것이다.

본 논문은 다음과 같은 순서로 구성된다. II 장에서는 기본적인 NMF 기술에 대한 소개와, 짧은 구간 퓨리에 변환 (STFT: Short Time Fourier Transform)에 의한 신호의 시간-주파수 영역 표현인 스펙트로그램(spectrogram)에의 NMF 적용에 대해서 설명한다. FFX-NMF에 대한 상세한 분석과 NMPCF와 FFX-NMF 간의 상관관계 및 그들의 새로운 정의에 대해서 III 장에서 논하며, IV 장에서는 블라인드 방식의 리듬 음원 분리 문제에 대한 새로운 정의 및 모델을 제안한다. V 장에서는 10 개의 실제 상업 음악 신호에 대한 실험 결과가 기존의 방법과의 비교와 함께 제시되며, VI 장은 본 논문에 대한 결론을 제시한다.

## II. 비음수 행렬 분해 (NMF)를 이용한 스펙트로그램 분석

주어진 비음수 행렬 $X$에 대해, NMF는 2 개의 비음수 인수 행렬 $A$와 $S$를 찾는다. 즉, 입력 비음수 행렬 $X$는 2 개의 인수 행렬의 내적 (inner product) $AS$로 표현 가능하다. NMF는 유클리드 방식의 오차 (Euclidean error) 또는 I-divergence 방식의 오차를 이용하여 원본 행렬 $X$와 복원 행렬 $AS$의 차이를 표현하며, 이것을 최소화하는 방식으로 최적화를 수행한다. 무작위 비음수로 초기화된 인수 행렬 $A$와 $S$는 상기 최적화를 통해 점점 그 내적이 원본 행렬 $X$와 가까워지도록 갱신된다. 예를 들어, 유클리드 방식의 오차 함수는 식 (1)과 같이 주어질 수 있다.

$$\begin{aligned} &\arg\ \min \mathfrak{I} = \frac{1}{2} \| X - AS \|_F^2 \\ &A, S \geq 0 \end{aligned} \tag{1}$$

식 (1)에서 $\| \bullet \|_F^2$는 Frobenius norm (행렬 요소별 제곱의 총합에 대한 2의 제곱근)을 나타낸다. [7]에서는 곱셈에 의한 인수 행렬 $A$, $S$의 갱신 규칙이 유도되었는데, 이는 목적 함수 (1)에 대한 일반적인 점진 하강 방식 (gradient descent method)의 최적화에서 스텝 크기를 매번 정교하게 교정함으로써 얻을 수 있다. 인수 행렬이 비음수에 의해 초기화 되어 있고 곱해지는 갱신 요소 역시 비음성을 유지한다면, 인수 행렬의 부호는 갱신에 의해 변화되지 않는다. 즉, NMF는 비음성을 유지하면서 식 (1)의 목적함수를 최소화하기 위해서, 곱셈에 의한 갱신 규칙을 통해 인수 행렬을 갱신하는 것이다. 식 (1)을 최소화하기 위한 곱셈 갱신 규칙은 식 (2)와 같이 정의된다.

$$\begin{aligned} A &\leftarrow A \odot \frac{XS^T}{ASS^T}, \\ S &\leftarrow S \odot \frac{A^T X}{A^T AS} \end{aligned} \tag{2}$$

식 (2)에서 $\odot$는 행렬의 요소 단위 곱셈 (Hadamard product)이며, 나눗셈 역시 행렬의 요소 단위로 수행된다.

곱셈 갱신 규칙은 점진 하강 방식의 스텝 크기를 교정하는 방식 외에, 보다 간단한 방식으로도 정의될 수 있다. 이는, 목적 함수를 각 인수 행렬 $A$와 $S$로 편미분한 도함수에서, 양수 항을 분자로 삼고, 음수 항을 분모로 삼는 방식으로 정의된다. 예를 들어, 식 (1)의 목적 함수 $\mathfrak{I}$의 기저 요소 행렬 $A$에 대한 편미분은 다음과 같은 양수 항과 음수 항을 가진다.

$$\begin{aligned} \frac{\partial \mathfrak{I}}{\partial A} &= \left[ \frac{\partial \mathfrak{I}}{\partial A} \right]^+ - \left[ \frac{\partial \mathfrak{I}}{\partial A} \right]^- \\ &= ASS^T - XS^T \end{aligned}$$

이러한 개념을 임의의 인수 행렬 $\Theta$에 대한 일반적인 곱셈 갱신 규칙을 도출하는 것에 확장하면, 식 (3)과 같은 규칙을 얻을 수 있다.

$$\Theta \leftarrow \Theta \odot \left( \frac{[\partial \mathfrak{I} / \partial \Theta]^-}{[\partial \mathfrak{I} / \partial \Theta]^+} \right)^{\cdot \eta} \tag{3}$$

여기서 $.\eta$는 행렬의 요소 단위 승수를 나타내며, $\eta$를

$0 < \eta \leq 1$과 같이 정의함으로써, 곱셈의 정도를 조정하기 위해 쓰인다. 이하 본 논문에서 제안하는 갱신 규칙들은 상기 식 (3)과 같은 방식을 통해 도출하였다.

다음으로, NMF의 스펙트로그램에 대한 활용에 대해 살펴본다. NMF에 의해 얻어지는 인수 행렬 $A^{N\times R}$와 $S^{R\times M}$은 내적에 의해 입력 행렬 $X^{N\times M}$의 근사치를 생성한다. 이 때, 인수 행렬 $A$의 열벡터 개수 $R$ (또는 인수 행렬 $S$의 행벡터 개수 $R$)은 어떠한 값이든 가능하지만, 일반적으로는 $R < N$, $M$과 같이 정의된다. NMF는 주어진 $R$에 대해 $X$를 분해하며, 이 때 스펙트로그램의 절대값을 입력 행렬 $X$로 삼으면, 스펙트로그램은 NMF에 의해 분해되어 인수 행렬 $A$와 $S$의 곱으로 표현된다.

스펙트로그램은 널리 알려진 STFT에 의해 구해진다. 이 때 복소수 스펙트로그램 행렬의 각 행은 저대역부터 고대역까지의 주파수 성분 별 시간에 따른 신호 변화를 표현하며, 각 열은 특정한 짧은 시간 구간에서의 주파수 성분을 의미한다. 일반적으로 실수값을 가지는 신호에 대해, STFT는 그것을 복소수 행렬로 표현하기 때문에 NMF의 비음수 가정에 맞지 않으나, 대부분의 응용에서는 STFT의 절대값을 취하여 NMF의 입력 행렬로 사용한다.

NMF에 의해 분석된 스펙트로그램은 주파수 영역의 기저 열벡터로 이루어진 인수 행렬 $A$와, 각각의 기저 열벡터에 대한 시간 영역의 가중치 $S$의 곱에 의해 표현된다. 다시 말해, $A$의 $r$번째 열벡터와 $S$의 $r$번째 행벡터의 곱은, NMF가 산출하는 전체 $R$개의 성분 중 $r$번째 성분을 나타낸다. 스펙트로그램에 대한 NMF 분석을 통해 얻어지는 이와 같은 $R$개의 성분은, 각각 신호에 포함된 다양한 특성들을 가능한한 배타적으로 표현하게 되는데, 이는 NMR의 희소 표현 또는 부분 기반 표현 방식에 의해 가능하다.

따라서, 음악 신호의 스펙트로그램이 $R$개의 배타적인 성분으로 분해된다면, 각각의 성분은 군집화되어 특정한 악기 음원을 표현하고 있을 가능성이 높다. 만일 $R = R_D + R_H$와 같다고 할 때, 우리는 $R_D$개의 성분들이 리듬 악기 음원을 재구성할 수 있고, $R_H$ 개의 성분들이 화성 악기를 재구성할 수 있을 것으로 추정이 가능하다. 그러나, 리듬 악기 음원을 복원하기 위해서는, $R$ 개의 성분들 중 어떤 성분들이 리듬 악기에 속하는지를 분간할 수 있어야 한다. 그 이유는, NMF 분석 과정에는, 배타적인 성분들 중 비슷한 성분들 일부를 정렬하거나 따로 모으는 과정이 포함되어 있지 않기 때문이다. 즉, 교환 모호성 (permutation ambiguity)가 존재하는 상황에서, 임의로 앞부분 $R_D$만큼의 성분을 모아서 내적한다고 해서 목표하는 음원이 복원되는 것은 아니다. NMF의 희소 표현이 음악 신호로부터 리듬 악기 음원을 실제로 분리할 수 있는지 여부를 떠나서, 상기와 같은 결과 성분의 군집화 또는 정렬 작업은 또 다른 부담이 될 수 있으며, 이와 같은 문제를 해소하기 위해 목표 음원에 대한 사전 정보를 활용하는 방안들이 제시되어 왔다.

## III. 비음수 행렬의 부분적 공동분해와 FFX-NMF의 재정의

FFX-NMF는 EEG (Electroencephalogram)의 분류를 위해 제안되었으며 [8], 복수의 대상에 공통적으로 공유될 수 있는 NMF의 기저 벡터와, 각각의 대상이 개별적으로 가지고 있는 특성을 표현하는 NMF의 기저 벡터를 효과적으로 분해하는 것을 목표로 한다. $l$ 번째 대상 입력 행렬 $X^{(l)}$에 대해 FFX-NMF는,

$$X^{(l)} = A_C S_C^{(l)} + A_I^{(l)} S_I^{(l)} \tag{4}$$

와 같은 형태로 행렬 인수분해를 수행하며, 여기서 $A_C$는 모든 입력 행렬들이 공유하는 공통의 기저 열벡터들로 이루어져 있다. 반면, $A_I^{(l)}$는 $l$ 번째 입력 행렬이 개별적으로 가지고 있는 성분을 표현하는 기저 열벡터로 이루어져 있다. $S_C^{(l)}$와 $S_I^{(l)}$는 각각 해당되는 가중치 행렬이다.

이와는 별도로, 비음수 행렬의 부분적 공동분해 (NMPCF)는 비음수 행렬 공동분해 (NMCF: Nonnegative Matrix Co-Factorization)를 개선하기 위해 제안되었다 [11]. NMPCF는 분해 결과의 일부 기저 벡터만을 다른 입력 행렬의 전체 기저 벡터와 공유함으로써 타악기 음원 분리에 활용되었다 [12]. NMPCF에서 기저 벡터의 부분적인 공유 개념은 일부 나머지 기저 벡터를 다른 입력 행렬과 공유되지 않도록 허용하는 역할을 하는데, 이는 FFX-NMF가 개별 기저 벡터 행렬 $A_I^{(l)}$를 허용하는 것과 비슷한 개념이다. 이것은 또한, NMCF에 대한 NMPCF의 주요한 개선점이기도 하다. 그러므로, FFX-NMF는 NMPCF의 특수한 경우로 볼 수 있으며, 단, 모든 입력 행렬이 공유 기저 벡터와 개별 기저 벡터를 모두 가지도록 분해되는 경우이다. 반면, NMPCF는 입력 행렬들 중 일부가 공유 기저 벡터로만 분해되는 상황 또한 가정한

다. 이처럼, NMPCF는 식 (4)에서 제시된 분해 방식을 포함하며, 동시에 $A_I^{(l)}$가 빈 행렬인 경우까지도 정의할 수 있다. 그림 1 (a)는 입력 행렬들 중 일부가 공유 기저 벡터로만 분해되는 상황을 표현하고 있다. 그림 1에 관한 보다 상세한 설명은 IV 장에서 제시한다.

상기 소개된 바와 같이 통합된 NMPCF 모델을 통해 주어진 $L$개의 입력 행렬에 대한 NMPCF 수행 방식은 아래와 같은 목적 함수를 최적화하는 것으로 수식화할 수 있다.

$$\Im_{NMPCF}=\sum_{l=1}^{L}\lambda_l \| X^{(l)}-A_C S_C^{(l)}-A_I^{(l)}S_I^{(l)} \|_F^2+\gamma\left\{\sum_{l=1}^{L}\| A^{(l)} \|_F^2\right\} \quad (5)$$

여기서 조정 (regularization) 항은 다음과 같이 정의된다.

$$\sum_{l=1}^{L}\| A^{(l)} \|_F^2=L\| A_C \|_F^2+\sum_{l=1}^{L}\| A_I^{(l)} \|_F^2$$

식 (5)를 최적화하는 $S^{(l)}$, $A_C^{(l)}$, $A_I^{(l)}$에 대한 곱셈 갱신 규칙은,

$$S^{(l)} \leftarrow S^{(l)} \odot \left( \frac{A^{(l)^T}X^{(l)}}{A^{(l)^T}A^{(l)}S^{(l)}} \right)^{\cdot\eta},$$

$$A_C \leftarrow A_C \odot \left( \frac{\sum_{l=1}^{L}\lambda_l X^{(l)}S_C^{(l)^T}}{\sum_{l=1}^{L}\lambda_l A^{(l)}S^{(l)}S_C^{(l)^T}+\gamma L A_C} \right)^{\cdot\eta}, \quad (6)$$

$$A_I^{(l)} \leftarrow A_I^{(l)} \odot \left( \frac{\lambda_l X^{(l)}S_I^{(l)^T}}{\lambda_l A^{(l)}S^{(l)}S_I^{(l)^T}+\gamma A_I^{(l)}} \right)^{\cdot\eta},$$

과 같이 표현될 수 있으며, 이는 식 (3)에서와 같이 식 (5)를 각각의 인수 행렬에 대해 편미분함으로써 얻을 수 있다. 각각의 입력 행렬에 대한 가중치 $\lambda_l$은 입력 행렬에 대한 중요도를 반영할 수 있다.

## IV. 블라인드 방식의 리듬 음원 분리와 NMPCF

본 논문에서 가정하는 리듬 악기 음원의 가장 큰 특징은, 리듬 악기가 곡 전체에 대해서 반복적으로 연주되는 동시에, 주파수 특성이 변하지 않는다는 점이다. 이는 NMPCF의 모델처럼, 파편을 구성하는 성분들 중 일부가 여러 파편들에서 공유될 수 있다는 것을 뜻한다. 먼저, 입력 혼합 신호 절대값 스펙트로그램 $X$를 $L$ 개의 파편, $X^{(1)}$, $X^{(2)}$, $\cdots$ ,$X^{(L)}$로 분할하며, 이를 통해 혼합 음악 신호에 대한 복수 개의 조각을 획득한다. 다음으로, 이 파편들을 NMPCF의 입력 행렬들로 취하며, 인수 행렬 $A_C$와 $A_I$, $S_C$와 $S_I$를 비음수 무작위 숫자로 초기화한다. 식 (6)에서 제시된 갱신 규칙에 의해 얻어지는 인수 행렬들 중, 리듬 악기 음원의 반복성에 대한 가정에 따라, $A_C$와 $S_C^{(l)}$의 곱이 리듬 악기 음원의 성분을 복원하는 것으로 본다. 한 편, 개별 파편에만 포함되어 있는 화성 악기 음원의 연주 부분은 $A_I^{(l)}$와 $S_I^{(l)}$의 곱으로 표현할 수 있다. 복원된 각 음원 별 절대값 스펙트로그램과 혼합 신호에서 따로 추출된 위상 정보 $\Phi$를 활용하여 출력 시간 영역 신호를 생성할 수 있다.

그림 1 (b)는 입력 혼합 스펙트로그램이 2 개의 파편으로 분할되는 경우에 대한 예를 표현하고 있다. $X^{(2)}$가

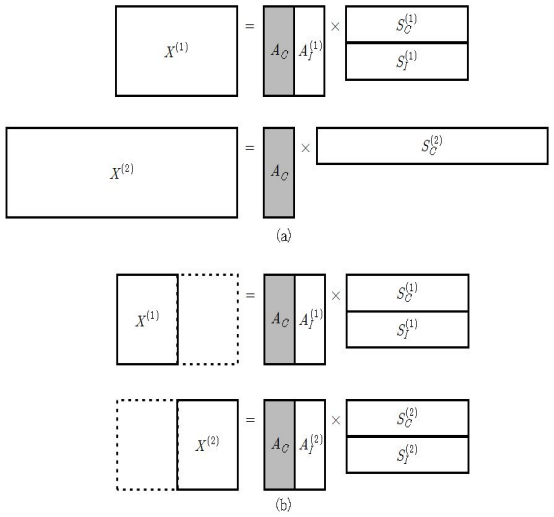


그림 1. NMPCF를 이용한 음악 음원 분리 모델.
(a) NMPCF를 이용한 사전정보 기반 타악기 음원 분리 모델. 타악기 솔로 연주로 이루어진 스펙트로그램 절대값 행렬 $X^{(2)}$는 여러 악기의 혼합 신호로 이루어진 $X^{(1)}$로부터 타악기 음원 성분을 추출하기 위하여 사전 정보로서 활용된다.
(b) NMPCF를 이용한 블라인드 방식의 리듬 음원 분리 모델. 사전 정보로 활용될 수 있는 데이터가 없는 상황이며, 여기서 $X^{(1)}$과 $X^{(2)}$는 각각 혼합 신호의 중첩되지 않은 파편이다. 본 논문에서 제안하는 방법은 모든 $X^{(l)}$에 걸쳐서 반복적으로 연주되는 음원, 즉 리듬 악기 음원을 $A_C$의 형태로 표현하는 것이다.

Fig 1. Figure 1. MSS models using NMPCF.
(a) Prior knowledge-based drum source separation model using NMPCF
(b) Blind rhythmic source separation model using NMPCF and segmentation

사전 정보로 활용되는 드럼 독주 신호의 스펙트로그램인 그림 1 (a)에서와는 다르게, 그림 1 (b)에서 $X^{(2)}$는 혼합 신호의 한 파편일 뿐, 따로 리듬 신호에 대한 사전 정보가 없는 상황을 가정한다. 이처럼 본 논문에서 제안하는 방식은, 대상 음원에 대한 사전 정보를 이용하지 않고 이미 알려진 리듬 악기 음원의 반복적 연주 특성을 활용하여 분리를 수행하므로, 블라인드 방식의 일종이라고 할 수 있다. 그림 2는 본 논문에서 제안하는 반복적 리듬 악기의 추출 방식을 블록 다이어그램으로 도식화한 것이다.

한 편, 음악에 사용되는 악기는 그 분류 기준에 따라 다양한 군집으로 나눌 수 있다. 예를 들어 MIDI (Music Instrument Digital Interface) 사용 여부에 따라 합성된 가상 악기와 어쿠스틱 (acoustic) 악기로 분류할 수 있으며, 소리의 생성 방식에 따라 타악기, 현악기, 관악기 등으로 분류할 수 있다. 더 나아가서 현악기의 경우, 기타나 피아노처럼 줄을 두드리거나 튕겨서 연주하는 현악기와, 바이올린이나 첼로처럼 활로 현을 켜서 소리를 내는 악기로도 분류할 수 있다. 타악기의 경우, 주기성 (periodicity), 주파수 평탄도 (spectral flatness), 주파수 중점 (spectral centroid), 에너지 분포 (roll-off point), 잡음성 (noiselikeness) 등 다양한 기준을 종합하여 타악기인지 타악기가 아닌지 여부를 분간할 수 있다.

본 논문에서 분리의 목표로 하는 대상 음원은 방식은 "리듬" 악기로 정의하고자 하며, 이는 기존의 "드럼" 또는 "타악기" 음원 분리와는 다른 개념을 포괄한다. 리듬 악기는 반복성 (주기성 보다는 넓은 개념임)을 지닌 모든 악기를 통칭하는 것이며, 타악기 뿐만 아니라 화성 악기라고 하더라도 반복성을 지닌다면 리듬 악기로 분류될 수 있다. 참고로 타악기는, 일반적으로 타격에 의해 시간 영역에서 충격성이 있는 신호를 만들어내며, 또한 시간이 지남에 따라 그 음량이 빠르게 감소한다. 또한, 주파수 성분 분석에 의하면, 많은 주파수 대역에 넓게 펼쳐진 잡음성을 가지고 있다. 대부분의 타악기는 일반적으로 반복적으로 연주되기 때문에, 리듬 악기의 조건을 충족한다.

이에 반해, 대부분의 화성 악기는 배음 구조 (harmonics)와 느린 음량 감소 등 타악기와 많은 차이점을 가지고 있지만, 이외에도 다양한 음정을 분산 또는 동시에 연주하는 특성 때문에, 설령 노래의 모든 구간에 걸쳐서 연주된다고 하더라도, 각각의 음정 단위로 보아서는 반복성이 높다고 보기 어렵다. 특이한 예로, 클래식 음악에서 사용되는 타악기인 팀파니 (timpani)의 경우, 음정을 가지고 있지만, 타격에 의해 소리를 내기 때문에 타악기의

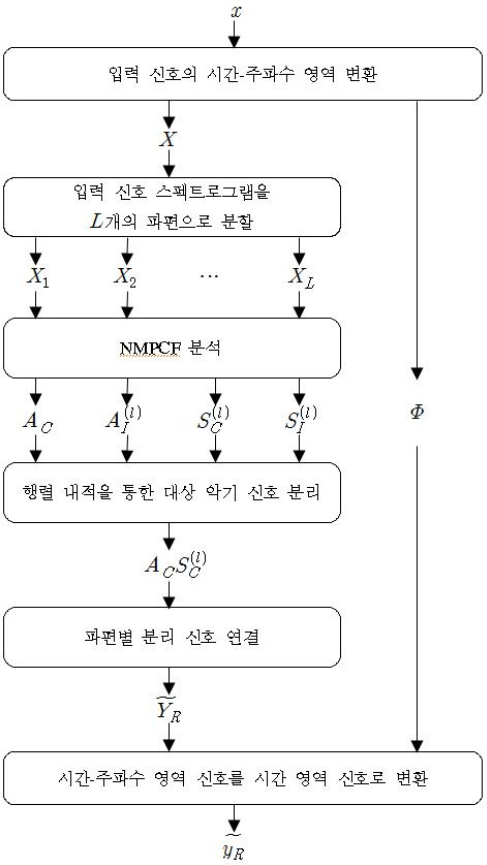


그림 2. 블라인드 방식의 반복적인 리듬 악기 음원 분리 방법 흐름도

Fig 2. Flowchart of blind repeating rhythmic source separation.

특성 또한 가지고 있다. 특히, 상업 음악에서 많이 쓰이는 베이스 기타 (bass guitar)의 경우, 충격성 또는 잡음성을 가지고 있지는 않으나, 동시에 여러 음정이 연주되는 경우가 많지 않으며, 무엇보다도, 베이스 드럼 (bass drum)의 타격에 맞추어 해당 코드 (chord)의 근음과 관련 꾸밈음을 연주하는 경우가 많다. 따라서, 베이스 기타가 노래의 전 영역에 걸쳐 너무 많은 음정을 연주하지 않는다면, 비록 음정이 변하기는 하지만 리듬 악기의 반복성을 가지고 있다고 볼 수 있다. 베이스 기타를 리듬 악기에 포함시키는 것은, 실제 음악에서 베이스 기타를 리듬 악기와 함께 분류하기도 하기 때문에, 어떤 면에서는 어느 정도 일리가 있는 분류라고 할 수 있다.

## V. 실험 결과

실험에 사용된 데이터는 10 개의 상업 음악이며, CD 음질과 같은 44.1 kHz의 표본율 (sampling rate)과 16

비트로 인코딩된 WAV 형태의 파일이다. 음악의 길이는 모두 100초이다. 또한, 분리 성능을 보다 정량적으로 측정하기 위해, 각 음악에 포함되어 있는 드럼 신호 원본과 가창을 포함한 나머지 화성 악기의 원본을 확보하였다. 혼합 신호는 미리 정해진 길이만큼의 연속적인 파편으로 분할하였다. 본 알고리즘을 수행함에 있어서 고려해야 할 매개 변수는 아래와 같다.

STFT를 위한 매개변수: 2048 표본만큼의 길이를 가지는 윈도우 함수가 7/8의 비율로 다음 윈도우와 중첩되도록 함.

$\eta$, $\gamma$, $\lambda_l$: 이 값들은 분리 결과에 큰 영향을 미치지 않았으며, 모두 1로 고정했음.

기저 벡터의 개수: $A_C$에 대해서는 30 개의 무작위 비음수 값을 가지는 열벡터를 기저 벡터로 할당하였으며, $A_I^{(l)}$는 15 개를 할당하였음. 이 개수에 대한 약간의 변동은 결과에 크게 영향을 미치지 않으나, $A_C$를 통해 복원하는 리듬 악기가 곡의 전체를 포괄할 수 있도록 넉넉한 기저 벡터 개수를 할당하였음.

파편의 길이: 파편의 길이는 SNR 기준으로 보아 1에서 2dB 정도의 영향을 미침. 그러나 블라인드 방식을 가정하는 상황에서 적절한 길이를 미리 아는 것은 불가능함. 본 실험에서는 모든 파편 길이를 4 초로 고정하였으나, 리듬 음원의 여러 성분을 충분히 포함할 수 있을 정도로 길다면 길이 변화에 따른 분리 성능의 변동은 심각하지 않음.

최적의 갱신 회수: 식 (6)에서 제안된 갱신 규칙에 따라 인수 행렬을 갱신하면, 식 (5)에서 제안된 목표 함수는 그 값이 감소하며, 수백 회 이내의 갱신으로 수렴이 가능함. 그러나 더 작은 목표 함수 값이 언제나 더 좋은 분리 성능을 보장하는 것은 아님. 실제로 원본 드럼 신호와의 SNR 비교를 해보면, 5에서 20회 이내의 갱신에서 가장 높은 SNR 값을 가지며, 그 이후 추가적인 갱신은 SNR 값을 감소시킴. 목적 함수 값을 감소시키는 것과는 별개로, SNR 값 관점에서 최적의 갱신 회수는 입력 음악 신호의 종류에 따라 변화가 있음. 파편의 길이 결정과 마찬가지라, 블라인드 상황에서 최적의 갱신 회수를 미리 아는 것은 불가능하므로, 본 실험에서는 모든 노래에 대해 15회 갱신하는 것으로 고정하였음. 이는 모든 노래에 대해서 가장 좋은 결과를 보장하는 값이 아님.

$L$ 개의 시간–주파수 영역 절대값 파편 행렬은 NMPCF 분석 과정의 입력이 된다. 주어진 $l$ 번째 입력 $X^{(l)}$에 혼합되어 있는 리듬 악기 음원의 절대값 스펙트로그램 $Y_R^{(l)}$는,

$$X^{(l)} = Y_R^{(l)} + Y_H^{(l)}$$

NMPCF 분석 이후의 공통 기저 벡터로 이루어진 행렬 $A_C$와 그에 대한 파편 고유의 시간 영역 가중치 $S_C^{(l)}$의 행렬 내적으로 복원된다. 복원된 전체 리듬 악기 신호의 절대값 스펙트로그램 $\widetilde{Y_R}$는, 파편 별 복원 값 $\widetilde{Y_R^{(l)}}$의 연결을 통해 얻어진다. 화성 악기 음원의 파편 $Y_H^{(l)}$의 복원 값 $\widetilde{Y_H^{(l)}}$ 역시 비슷한 방법으로 얻을 수 있으나, 공통 기저 벡터로 이루어진 행렬 대신, 파편 개별적으로 산출하는 기저 벡터의 행렬 $A_I^{(l)}$과 연관된 가중치 행렬 $S_I^{(l)}$의 내적을 통해 얻어진다. 상기 과정을 수식을 통해 포현하면 아래와 같다. 먼저, 리듬 악기와 화성 악기의 복원 스펙트로그램은 파편별 복원값의 연결을 통해 얻어진다.

$$Y_R \approx \widetilde{Y_R} = \left[ \widetilde{Y_R^{(1)}}, \widetilde{Y_R^{(2)}}, \cdots, \widetilde{Y_R^{(L)}} \right]$$
$$Y_H \approx \widetilde{Y_H} = \left[ \widetilde{Y_H^{(1)}}, \widetilde{Y_H^{(2)}}, \cdots, \widetilde{Y_H^{(L)}} \right]$$

여기서 [,] 연산은 스펙트로그램의 가로 연결, 즉 시간 영역으로의 연결을 의미한다. 그리고 $l$ 번째 파편의 복원 $\widetilde{Y_R^{(l)}}$, $\widetilde{Y_H^{(l)}}$을 수식으로 표현하면 아래와 같다.

$$\widetilde{Y_R^{(l)}} = A_C S_C^{(l)}$$
$$\widetilde{Y_H^{(l)}} = A_I^{(l)} S_I^{(l)}$$

상기와 같은 방법을 통해 얻어진 복원 행렬 $\widetilde{Y_R}$, $\widetilde{Y_H}$는, 혼합 신호로부터 획득한 위상값 $\Phi$를 이용하여 시간 영역으로 역변환되어, 시간 영역에서의 복원 신호 $\widetilde{y_R}$, $\widetilde{y_H}$를 얻는다. 경우에 따라, 목표 음원 외의 나머지 신호의 음질 역시 중요한 경우, 혼합 신호에서 목표 음원의 복원 신호를 빼는 것으로 대신할 수 있다.

$$\widetilde{y_H} = x - \widetilde{y_R}$$

본 실험에서는 복원된 신호의 음질을 정량적으로 측정하기 위해 신호 대 잡음비 (SNR: Signal to Noise Ratio)를 기준으로 삼았으며, 아래와 같은 수식을 통해 얻는다.

$$SNR = 10\log_{10} \frac{\Sigma\, s(t)^2}{\Sigma \left(s(t) - \tilde{s}(t)\right)^2}$$

여기서 $s(t)$와 $\tilde{s}(t)$는 각각 원본 음원과 복원 음원을 나타낸다.

표 1은 세 가지 리듬 또는 타악기 음원 분리 시스템의 성능 비교표이다. NMF+SVM은 [4]에서 제안된 시스템으로, NMF 분석된 혼합 신호의 기저 벡터를 미리 학습된 SVM으로 분류하여 복원한다. S-NMPCF는 [12]에서 제안된 감독형 (supervised) 방식으로, 드럼 독주 데이터를 학습용으로 활용한다. U-NMPCF는 [13]에서 제안된 본 논문의 비감독형 (unsupervised) 방식이다. 기존 시스템을 활용한 실험은, 참고 문헌에서 제안된 것과 최대한 같은 조건 하에서 실험하였으며, 다만 시간-주파수 영역 변환에 따른 매개 변수는 모든 실험에 대해 동일하게 수행하였다. S-NMPCF 방식의 경우는 130초의 드럼 독주 데이터를 학습용으로 활용하며, NMF+SVM의 경우는 130초의 드럼 독주 데이터 이외에도 가창과 화성악기들로 이루어진 130초의 추가적인 학습 데이터를 활용하였다. 표 1에 따르면, 대부분의 경우 S-NMPCF 방식이 높은 성능을 보였으며, U-NMPCF는 기존의 시스템보다 낮은 성능을 보여준다. 그러나, 제안된 알고리즘이 블라인드 방식에 기반하여 사전 정보를 활용하지 않는다는 측면을 고려하면, U-NMPCF는 상대적으로 낮은 성능에도 불구하고 많은 경우에 활용 가치가 높다.

또한, 특히 낮은 성능을 보여주는 입력 신호, 예를 들어 6번 곡의 경우, 실제로 최적의 갱신 회수를 가정하면 보다 높은 성능을 얻을 수 있다. 그러나 원본 리듬 신호가 없다는 블라인드 방식의 가정에 따라, 15 회의 최적이 아닌 갱신 회수를 사용하였다. U-NMPCF의 낮은 성능에 대한 또다른 설명은, 복원된 신호가 드럼 신호 뿐만 아니라 베이스 기타의 신호까지 포함하고 있다는 점이다. 이는 SNR 측정에서 원본으로 활용한 드럼 신호와의 차이를 야기하며, SNR 수치의 저하에 영향을 미칠 것으로 생각된다. 응용에 따라, 갱신을 거듭하면서 산출되는 복원 신호를 청취하면서, 성능이 일정 정도에 도달하면 갱신을 멈추는 식의 방법 또한 생각해 볼 수 있다.

## VI. 결론

본 논문에서는 단일 채널에서 리듬 악기 음원을 분리하는 새로운 방법을 제시했다. NMF 알고리즘의 확장된 형태인 NMPCF 알고리즘을 활용하여, 반복적으로 연주되는 리듬 악기의 성분을 학습하였다. 이를 위해 입력 신호를 시간-주파수 영역으로 변환 및 절대값을 취하였고, 미리 정해진 길이 만큼의 짧은 시간 단위로 파편화시켜서 NMPCF 알고리즘의 입력 행렬로 삼았다. 리듬 악기에 대한 본 논문에서의 정의에 따라, 제안된 방법은 의미있는 분리 성능을 보여주었으며, 특히 대부분의 드럼 신호와 반복적으로 연주되는 베이스 기타 신호를 복원하였다. 본 논문에서 제시하는 방법은 몇 가지 최적의 매개 변수를 찾는 것에 있어서 개선의 여지가 있다. 그러나, 블라인드 방식으로 인한 본 논문의 장점은 타악기에 대한 사전 정보가 활용 가능하지 않는 경우 충분히 발휘될 수 있다. 예를 들어 드럼 독주 신호 등의 사전 정보가 아예 없는 경우, 제안된 방법을 사용할 수 있다. 한 가지 더 중요한 사용예는, 사전 정보로 확보한 드럼 독주 신호가 혼합 신

표 1. 10 개의 상용 음악에 대한 3 가지 음원 분리 시스템의 SNR 성능 비교
Table 1. SNR of separated sources for 3 different separation systems.

| 노래 번호 | SNR (드럼 원본에 대해) | | | SNR (화성 악기에 대해) | | |
|---|---|---|---|---|---|---|
| | NMF+SVM | S-NMPCF | U-NMPCF | NMF+SVM | S-NMPCF | U-NMPCF |
| 1 | 8.02 | 8.84 | 6.21 | 8.43 | 7.95 | 4.71 |
| 2 | 4.58 | 5.48 | 4.91 | 3.49 | 4.66 | 3.56 |
| 3 | 4.29 | 5.04 | 4.23 | 4.69 | 5.98 | 4.28 |
| 4 | 3.62 | 3.01 | 1.64 | 5.14 | 4.21 | 2.34 |
| 5 | 5.56 | 5.20 | 1.32 | 6.17 | 6.47 | 4.52 |
| 6 | 4.82 | 6.90 | 0.27 | 1.35 | 5.40 | 4.26 |
| 7 | 3.81 | 3.94 | 3.92 | 7.08 | 6.68 | 4.49 |
| 8 | -0.68 | 2.76 | 1.80 | 3.91 | 6.36 | 3.20 |
| 9 | 4.19 | 4.32 | 4.16 | 7.30 | 7.04 | 4.41 |
| 10 | 7.90 | 7.81 | 5.08 | 8.41 | 8.08 | 5.29 |
| 평균 | 4.62 | 5.33 | 3.35 | 5.60 | 6.28 | 4.11 |

호에 포함된 리듬 악기를 표현하는데에 부적절한 경우이다. 특히 최근과 같이 실제 악기가 아닌 가상의 음원으로 합성된 음악의 경우, 기존의 사전 정보와는 다른 특징을 가지는 리듬 악기 신호가 포함될 수 있으며, 이러한 경우 이 악기가 반복적인 연주 패턴을 보인다면 제안되는 알고리즘으로 분리가 가능할 것이다.

## 감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2009년도 문화콘텐츠산업기술지원사업의 연구결과로 수행되었음.

## 참 고 문 헌

1. Y. C. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classfication," *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1327 - 1336, 2005.
2. D. FitzGerald, M. Cranitch, and E. Coyle, "Shifted nonnegative matrix factorisation for sound source separation," *in Proc. IEEE Workshop on Statistical Signal Processing*, pp.1132-1136, July, 2005.
3. D. FitzGerald, M. Cranitch, and E. Coyle, "Sound source separation using shifted non-negative tensor factorisation," *in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. V653-656, May, 2006.
4. M. Helen and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Proc. European Signal Processing Conference*, Sept., 2005.
5. M. Kim and S. Choi, "On spectral basis selection for single channel polyphonic music separation," in *Proc. International Conference on Artificial Neural Networks (ICANN)*, vol. 2, pp. 157 - 162, Sept., 2005.
6. D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788 - 791, 1999.
7. D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 13. MIT Press, pp. 556-562, 2001.
8. H. Lee and S. Choi, "Group nonnegative matrix factorization for EEG classification," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 320-327, April, 2009.
9. P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177 - 180, Oct., 2003.
10. T. O. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066 - 1074, 2007.
11. J. Yoo and S. Choi, "Weighted nonnegative matrix co-tri factorization for collaborative prediction," in *Proc. of 1st Asian Conference on Machine Learning*, pp. 396-411, Nov., 2009.
12. J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010 (submitted for publication)*.
13. M. Kim, J. Yoo, K. Kang, and S. Choi, "Blind Rhythmic Source Separation: Nonnegativity and Repeatibility," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010 (submitted for publication)*.

## 저자 약력

**• 김 민 제 (Minje Kim)**
2004년: 아주대학교 정보 및 컴퓨터공학부 (학사)
2006년: 포항공과대학교 대학원 컴퓨터공학과 (석사)
2006년~현재: 한국전자통신연구원 (연구원)
※ 주관심 분야: 기계 학습, 음원 분리, 음악 정보 검색, 음성 및 오디오 코덱

**• 유 지 호 (Jiho Yoo)**
2006년: 포항공과대학교 컴퓨터공학과 (학사)
2006년~현재: 포항공과대학교 대학원 컴퓨터공학과 석박사통합과정
※ 주관심 분야: 기계 학습, 행렬 분해, 음원 분리, 음악 정보 검색

**• 강 경 옥 (Kyeongok Kang)**
1985년: 부산대학교 물리학과 (학사)
1988년: 부산대학교 물리학과 (석사)
2004년: 한국항공대학교 전자공학과 (박사)
2006년: 영국 University of Southampton (방문 연구원)
1991년~현재: 한국전자통신연구원 (책임연구원, 미디어응용연구팀장)
※ 주관심 분야: 오디오 신호처리, 객체 기반 오디오, 3D 오디오, 음성 및 오디오 코덱

**• 최 승 진 (Seungjin Choi)**
1987년: 서울대학교 전기공학과 (학사)
1989년: 서울대학교 대학원 전기공학과 (석사)
1996년: 미국 University of Notre Dame, Department of Electrical Engineering (박사)
1996~1997년: 일본 이화학연구소 (RIKEN) (Frontier Researcher)
1997~2001년: 충북대학교 전기전자공학부 (조교수)
2001~현재: 포항공과대학교 컴퓨터공학과 (교수)
※ 주관심 분야: 기계 학습, 확률그래프모델