

EventSep: 사전 학습 모델 결합 기반의 선택적 음원 분리

박경빈

Department of Electrical Engineering
Korea University

overjoy1008@korea.ac.kr

Abstract

텍스트 조건부 음원 분리는 자연어 프롬프트만으로 특정 음향 객체를 선택할 수 있다는 장점이 있으나, 텍스트 표현의 모호성, 시간 정보 부재, 그리고 프롬프트와 오디오 불일치로 인해 안정적인 분리 성능을 유지하기 어렵다. 본 연구는 이러한 구조적 한계를 보완하기 위해, 사전 학습된 분리·검출·복원 모델을 결합하여 의미적 정렬과 시간 선택성을 동시에 강화하는 EventSep 프레임워크를 제안한다. EventSep은 SED 기반 프레임 단위 시간 선택성을 통해 텍스트 기반 분리의 시간적 비정렬(temporal misalignment)를 보정하고, 텍스트 기반 및 확산 기반 음원 분리 모델의 병렬 구조를 활용해 의미적 정확도(semantic fidelity)와 청각적 자연스러움(perceptual quality) 간 균형점을 형성하며, STFT 기반 Ensemble을 통해 잔여 잡음(residual noise)를 완화한다. VGGSound, ESC-50, MUSIC에서 수행된 실험 결과, EventSep은 기존 텍스트 기반 분리 대비 의미적 정렬(CLAPScore) 측면에서 일관된 향상을 보였으며, 특히 시간 선택성이 가장 큰 성능 기여 요인임을 확인하였다. 본 연구는 사전 학습 모델의 구조적 설계를 통해 텍스트 기반 선택적 음원 분리의 안정성과 신뢰도를 높일 수 있음을 보여준다.

1 서론

최근 대규모 오디오-텍스트 및 비전-언어 모델의 발전과 더불어, 텍스트 조건부 음원 분리(text-guided source separation) 기술이 빠르게 확산되고 있다. 텍스트 조건부 음원 분리 모델은 자연어 프롬프트만으로도 여러 오디오가 합성되거나 섞인 복합 오디오로부터 특정 오디오 객체를 선택하여 분리할 수 있어, 음악이나 영상 편집, 언어 학습 등 다양한 분야에서 오디오 이해, 편집, 검색 등 다양한 응용 가능성을 제시하였다. 그러나 텍스트 기반 분리 방식은 본질적으로 텍스트 표현의 모호성, 시간 정보 부재, 프롬프트와 실제 오디오의 불일치 등 근본적인 구조적 한계를 갖는다. 이 문제는 특히 사용자가 입력한 프롬프트가 해당 복합 오디오에 존재하지 않는 경우, 혹은 긴 오디오에서 특정 오디오가 지속적으로 등장했다가 사라지는 패턴을 보이는 환경에서 더욱 두드러지며, 특히 프롬프트가 지시하는 오디오 객체가 실제로 존재하지 않는 구간까지 모델이 억지로 분리하려고 시도

하는 현상으로 나타난다. 기존 텍스트 기반 분리 접근의 한계를 보완하기 위해, 프롬프트에 제시된 특정 음원 대상(예: 'drum')이 실제로 존재하는 시간 구간을 사운드 이벤트 검출(SED, sound event detection) 모델을 통해 먼저 식별하고, 그 식별된 구간에 대해서만 음원 분리를 수행하는 새로운 프레임워크를 제안한다. 구체적으로는 PANNs 기반 SED 모델 [1], AudioSep [2], FlowSep [3]을 결합한 텍스트 기반 및 확산(diffusion) 기반 음원 분리 모델의 병렬 구조 기반의 선택적 음원 분리를 하는 EventSep 방식을 제안한다.

텍스트와 오디오 매칭 기반 분리는 CLAP [4]을 활용함으로써 다양한 자연어 프롬프트를 처리할 수 있는 장점을 갖는다. CLAP은 텍스트와 오디오를 동일한 잠재 공간으로 매핑하는 대표적 대조 학습 모델로, 두 modality 간의 의미적 정렬(semantic alignment)을 안정적으로 측정할 수 있어 최근 텍스트 기반 분리·검색·편집 연구의 기반으로 활용된다. 그러나, 'bass', 'vocal', 'drum roll', 'running water' 등의 표현은 문맥에 따라 서로 다른 음향 패턴을 가리킬 수 있다. 동일한 텍스트 프롬프트라도 장르, 연주 스타일, 녹음 환경에 따라 분리되어야 할 타깃 신호가 달라질 수 있다. 텍스트는 프롬프트가 지시하는 대상이 무엇인지 설명하는 데에는 유용하나, 해당 음향 이벤트가 언제 발생하는지에 대한 시간 정보를 제공하지는 못한다. 그 결과, 모델은 전체 시간축에 대해 분리 마스크를 예측하려 하며, 이는 타깃 이벤트가 존재하지 않는 구간까지 불필요하게 분리하려는 over-separation 현상을 야기한다. 또한, 프롬프트와 실제 오디오 신호의 존재 여부가 불일치하는 경우가 빈번하게 발생할 수 있다. 예를 들어 입력 혼합 신호에 바이올린 성분이 포함되어 있지 않음에도 'violin solo'와 같은 텍스트 프롬프트를 제공하면, 모델은 바이올린이 아닌 음원을 hallucination 형태로 분리해 출력하는 문제가 존재한다. 이러한 이유로 현재 존재하는 모델들을 사용한 텍스트 기반 분리만으로는 안정적인 선택적 음원 분리를 구현하는 데에는 근본적인 한계가 존재한다.

2 관련 연구

2.1 텍스트 기반 음원 분리

텍스트 기반 음원 분리 모델(AudioSep[2] 등)은 기존의 보컬·베이스·드럼처럼 미리 정의된 소수의 카테고리만 분리할 수 있던 방식과 달리, "running water", "child"

laughing”, “violin harmonics”처럼 사용자가 자연어로 지정한 임의의 음향 대상을 직접 타깃으로 삼을 수 있다는 점에서 높은 유연성을 가진다. 이는 CLIPSep[5], LASS[6], SepDrafter[7]와 같은 텍스트-오디오 매칭 기반 분리 모델에서도 등장하는 공통적 추세이다.

그러나 텍스트만으로는 해당 소리가 언제 등장하는지를 알 수 없기 때문에, 전체 구간에서 타깃을 찾으려는 과분리, 존재하지 않는 소리를 생성하는 환각, 그리고 시간축 불일치로 인한 SDR(signal-to-distortion ratio)[8] 변동 등이 발생할 수 있다. 본 연구에서는 이러한 시간 정보 부재 문제를 보완하기 위해 SED 기반의 프레임 단위 시간 선택성을 AudioSep과 결합하였다.

2.2 확산 기반 음원 분리

최근 FlowSep과 같은 확산(diffusion) 모델 기반 오디오 생성기가 등장하면서[9], 기존 UNet 기반 분리 모델의 약점을 보완하는 방향이 제시되었다[3]. AudioLDM[10], AudioLDM2[11], 그리고 Stable Audio[12] 등 다양한 확산 모델은 음색 복원, 잡음 제거, 고주파 세부 복원 측면에서 특히 우수한 결과를 보인다.

FlowSep과 같은 latent diffusion 기반 분리 모델은 기타의 배음 구조나 사람 목소리의 고주파 성분처럼 STFT 마스크 기반 모델에서 쉽게 손상되는 미세한 질감을 자연스럽게 복원하는 데 강점을 갖는다. 특히 고역대에너지 단절, 잔향 왜곡, 과도한 평탄화(smoothing) 같은 중·고역 잡음을 효과적으로 보정할 수 있어, 청각적 자연스러움 측면에서 뛰어난 결과를 보인다. 다만, 텍스트 의미를 정밀하게 해석하여 특정 타깃을 직접 분리하는 능력은 AudioSep보다 제한적이므로, 본 연구에서는 FlowSep을 정확한 분리보다는 고주파 복원 및 질감 향상을 담당하는 보조 prior로 활용하였다.

2.3 Sound Event Detection (SED)

SED는 오디오 신호 속에서 어떤 소리가 포함되어 있는지, 그리고 그 소리가 언제 등장하고 사라지는지를 추정하는 기술이다. 즉, 입력 오디오에 존재하는 다양한 사운드 이벤트를 식별하고, 각 이벤트의 발생 구간을 자동으로 찾아내는 역할을 수행한다. 대표적인 SED로는 PANNs[1], HTS-AT[13], 그리고 PaSST[14] 등이 있다.

SED는 시간 정보를 추정하는 방식에 따라 크게 두 종류로 나뉜다. Clip-wise SED는 전체 오디오 단위에서 특정 소리의 포함 여부에 대한 정보만 제공한다. 반면, Frame-wise SED는 프레임 단위로 소리의 존재 확률을 예측하므로, 시간축을 따라 speech가 어느 구간에서 어떤 확률로 나타남과 같은 보다 세밀한 시간 구조를 확인할 수 있다. 본 연구에서는 분리해야 하는 소리가 어느 시점에 나타나는지를 확인하는 것이 중요하므로 보다 정밀한 시간 정보를 제공하는 frame-wise SED를 사용한다.

3 본론

EventSep은 Figure 1와 같이, 입력 오디오에서 최종 분리 결과까지를 네 단계로 구성한 4-stage 파이프라인을 따른다. 첫 단계에서는 입력 혼합 신호에 존재하는 이벤트의 시간적 발생을 SED 모델을 이용해 탐지한다. 두 번째 단계에서는 텍스트 프롬프트를 의미 기반으로 분석하여 가장 가까운 타깃 클래스를 선택한다. 세 번째 단계에서는 선택된 클래스의 SED 확률로부터 시간 선택적 마스크를 생성해 입력 오디오를 정제한다. 마지막 단계에서는 마스킹된 오디오를 두 분리 모델(Text-Conditioned UNet Separation, Diffusion-Based Separation)에 병렬로 통과시킨 뒤 두 신호를 결합하여 최종 오디오를 출력한다.

3.1 입력 오디오의 시간 기반 정보 추출

혼합 오디오에는 음악, 말소리, 배경 소음, 악기 연주 등 서로 다른 소리가 시간적으로 겹쳐 나타난다. 텍스트 기반 분리 모델은 어떤 소리를 분리해야 하는지는 잘 구분 하지만, 그 소리가 언제 나타나는지에 대한 시간 정보를 직접 제공하지 않는다. 따라서, EventSep은 먼저 SED를 통해 프레임 단위 시간 정보를 확보한다.

Figure 2는 이를 보완하는 SED의 예시로, 상단에는 mixture의 스펙트로그램, 하단에는 여러 사운드 카테고리에 대한 프레임 단위 존재 확률이 heatmap으로 제시되어 있다. 예시에서처럼 음악이 지속되고, 특정 지점에서 speech나 violin이 짧게 등장하는 패턴을 통해 각 소리가 실제로 언제 나타났는지를 직관적으로 확인할 수 있다. 이러한 과정은 프롬프트가 지시하는 소리가 mixture 안에서 실제로 등장한 구간만을 효과적으로 분리할 수 있도록 돋는다.

본 연구에서 SED 모듈은 PANNs-Cnn14를 사용하며, AudioSet 전체 527개 클래스[15]에 대해 약 100 fps 수준의 frame-wise 존재 확률을 예측한다. AudioSet은 약 2백만 개의 10초 길이 오디오 클립으로 구성된 대규모 인간 라벨링 데이터셋으로, 다양한 환경음과 음악 신호를 포함한다. AudioSet의 라벨 예시는 다음과 같다. speech, dog, water, violin, engine, footsteps, applause, musical instrument 등 일상 환경음과 음악 신호를 폭넓게 포함한다. 이러한 라벨 집합은 이후 텍스트 프롬프트를 타깃 클래스로 투사하는 기준이 된다.

오디오 파형이 입력되면 PANNs는 프레임 인덱스 k 에 대해 다음과 같은 존재 확률($p[k]$)을 출력한다.

$$p[k] = \Pr(c_i \in \text{frame } k)$$

여기서 $\Pr(\cdot)$ 는 특정 클래스(c_i)가 프레임 k 내에 존재 할 확률(probability)을 의미하며, c_i 는 AudioSet 클래스 중 하나이다. EventSep은 이 중 텍스트 프롬프트와 가장 의미적으로 가까운 클래스만을 선택하여 시간 선택적 마스킹의 근거로 사용한다.

자연어 프롬프트는 “acoustic guitar”, “running water”, “female speaking softly”와 같이 다양한 표현으로 주어진다. 반면 SED 모델은 AudioSet 라벨 단위로만

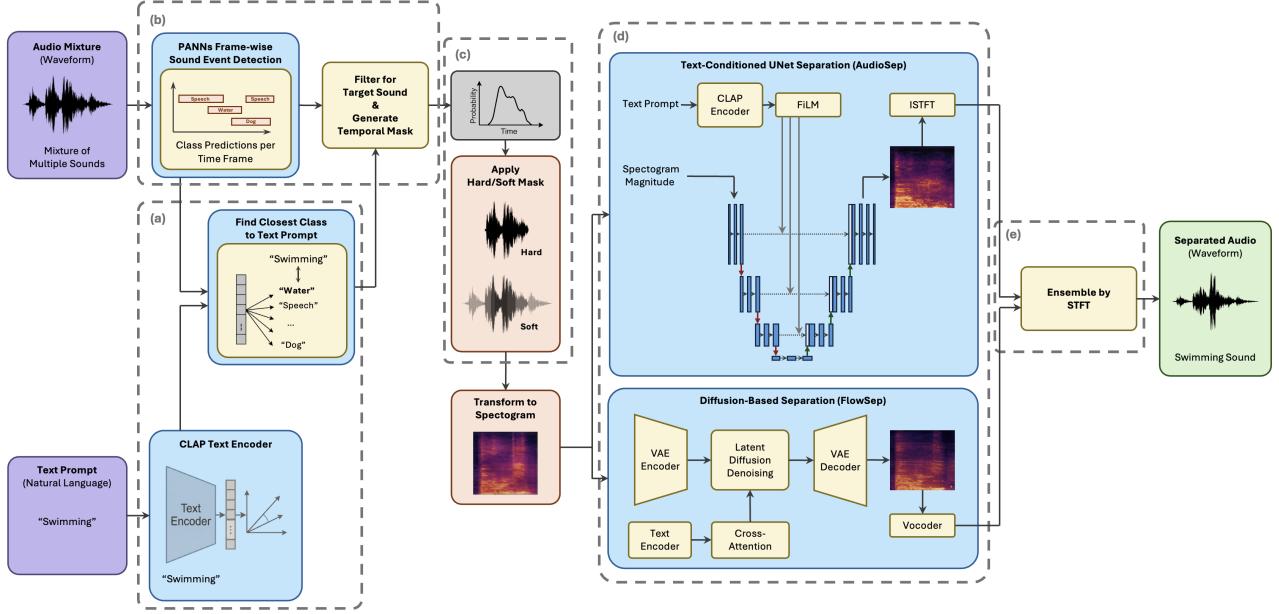


Figure 1: EventSep 전체 아키텍처 개요. (a) 텍스트 임베딩과 AudioSet 라벨 매핑, (b) PANNs 기반 프레임 단위 SED, (c) SED 확률을 이용한 시간 선택적 마스킹, (d) Text-Conditioned UNet Separation · Diffusion- Based Separation의 병렬 분리, (e) STFT 기반 신호 결합 단계를 포함한다.

학습되어 있으므로, 프롬프트를 AudioSet 클래스 중 하나의 타깃 클래스로 대응시켜야 한다.

텍스트 기반 분리에서는 “human talking”, “narration voice”, “someone speaking softly”처럼 다양한 표현이 동일한 음향 범주를 의미하는 경우가 많다. SED가 올바르게 동작하려면 이러한 프롬프트가 공통적으로 speech 클래스로 매핑되어야 하며, 이를 위해 텍스트 임베딩 모델이 활용된다. 임베딩 모델은 프롬프트와 클래스 이름들의 유사도를 비교해 가장 적합한 타깃 클래스를 선택하고, 그 클래스에 대한 frame-wise SED 확률을 시간 선택의 기준으로 삼는다.

예를 들어, 프롬프트가 “water”이고 AudioSet 라벨에 동일한 문자열이 존재하면 이 라벨과 매핑되지만, 정확히 일치하는 라벨이 없으면 부분 문자열 포함 관계를 검사하여, “car engine”과 “engine”처럼 명시적으로 포함 관계가 드러나는 경우를 우선적으로 매칭한다.

이 두 단계를 통해 적절한 클래스가 정해지지 않으면, SentenceTransformer 기반 MiniLM[16] 모델을 사용하여 텍스트 임베딩을 계산한다. AudioSet 각 라벨의 임베딩 z_{c_i} 와 프롬프트 임베딩 z_{text} 를 정규화한 뒤, 코사인 유사도에 해당하는 내적을 계산하여 가장 유사한 하나의 클래스를 선택한다.

$$c^* = \arg \max_{c_i \in \mathcal{C}} \cos(z_{text}, z_{c_i})$$

예를 들어, “swimming”, “splashing water”, “water flowing”과 같은 프롬프트는 대체로 water 클래스에 매핑되고, “female narrator”, “person talking”과 같은 표현은 speech 클래스에 매핑된다. 자동차 소음과 관련된

표현은 engine 또는 vehicle 계열 라벨로, 악기 관련 표현은 violin, cello, flute 등의 악기 라벨로 대응되는 식이다.

이렇게 선택된 c^* 은 이후 단계에서 frame-wise 확률 $p[k]$ 의 타깃 차원을 지정하는 역할을 한다.

3.2 SED 기반 시간 선택적 마스킹

선택된 타깃 클래스 c^* 에 대해 PANNs는 프레임 단위 존재 확률 $p[k]$ 를 출력한다. EventSep은 이 확률을 이용해 시간축에서 타깃 이벤트가 존재하는 구간만 강조하는 마스크를 생성한다. 이 마스크는 텍스트 기반 분리 모델이 존재하지 않는 구간까지 억지로 분리하려는 과분리를 구조적으로 억제한다.

EventSep은 hard mask와 soft mask 두 가지 방식을 지원한다. 추가적으로 상대 정규화 기반 soft mask와 가우시안 필터를 적용한 soft mask 변형도 구현되어 있으나, 주요 실험과 분석은 hard와 soft 두 방식에 초점을 맞춘다.

하드 마스크는 가장 간단한 형태로 threshold τ 를 기준으로 이진 값을 갖는다.

$$m_{\text{hard}}[k] = \begin{cases} 1, & p[k] \geq \tau, \\ 0, & p[k] < \tau. \end{cases}$$

소프트 마스크는 확률 값을 선형적으로 반영하여 threshold 미만에서는 감쇠, threshold 이상에서는 통과를 의미한다.

$$m_{\text{soft}}[k] = \min\left(\frac{p[k]}{\tau}, 1\right)$$

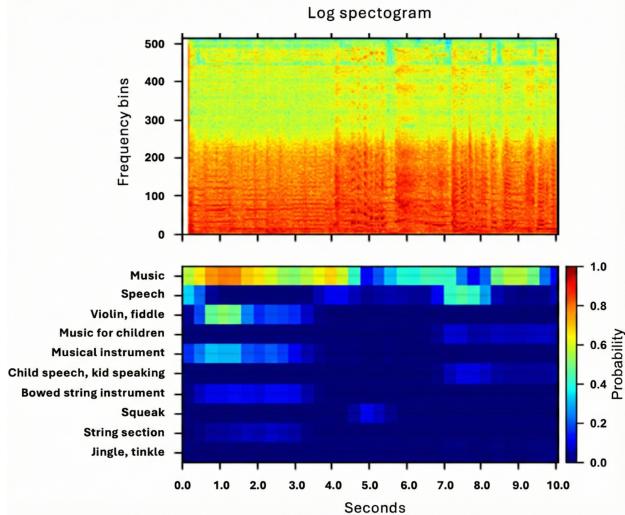


Figure 2: Frame-wise SED 예시. 상단은 입력 오디오 mixture의 로그 스펙트로그램이고, 하단은 SED 모델이 예측한 프레임별 사운드 존재 확률을 카테고리별 heatmap 형태로 나타낸 것이다. 음악, 말소리, 바이올린, 어린이 음성 등 서로 다른 소리가 다양한 시간 구간에서 등장하는 패턴을 명확하게 확인할 수 있다.

하드 마스크는 시간 선택성이 뚜렷하지만 threshold 근처에서 프레임이 불연속적으로 꺼졌다 켜지는 현상이 발생하여, 순간적인 오디오 단절로 인한 잡음을 유발하기 쉽다. 소프트 마스크는 확률 분포의 모양을 연속적으로 반영하므로, 특정 오디오의 존재 여부가 불확실한 경우(어떤 샘플인지 명확히 식별이 어렵거나, 시작·종료 시점이 모호한 경우 등)에도 훨씬 부드럽고 안정적인 출력을 제공한다.

SED가 출력하는 마스크는 약 100 fps 해상도를 가지므로, 분리 모델 입력인 파형 길이에 맞추기 위해 선형 보간을 수행한다. 파형 길이를 L 이라 할 때, 보간된 마스크 $m[n]$ 은 다음과 같이 정의된다.

$$m[n] = \text{Interp}(m[k]), \quad n = 0, \dots, L - 1$$

여기서 $\text{Interp}(\cdot)$ 는 구현 상 선형 보간 함수를 의미한다.

최종적으로 입력 파형 $x(n)$ 에 대해 다음과 같은 마스킹된 신호가 생성된다.

$$x'(n) = x(n) m[n]$$

마스크 합이 거의 0에 가까워 전체가 소거되는 경우에는 안정성을 위해 전체를 1로 복원하는 보정이 적용된다.

3.3 듀얼 모델 기반 분리 구조

SED 마스킹을 거친 신호는 텍스트 기반 음원 분리 모델(Text-Conditioned UNet Separation)과 확산 기반 음원 분리 모델(Diffusion-Based Separation)에 병렬로 입

력된다. EventSep은 두 모델의 상보적 특성을 동시에 활용하기 위해 이중 분리 구조를 채택하였다.

먼저 Text-Conditioned UNet Separation은 CLAP 텍스트 임베딩을 조건으로 사용하는 STFT 기반 UNet(ResUNet[17]) 구조의 텍스트 조건부 분리 모델이며, 본 연구에서는 AudioSep을 기본 모델로 사용한다. 마스킹된 파형 $x'(n)$ 은 스펙트로그램으로 변환한 뒤 UNet의 입력으로, 텍스트 프롬프트는 CLAP 텍스트 인코더로 각각 전달된다. 텍스트 임베딩은 FiLM 방식으로 UNet 내부의 각 특징 흐름에 적용되어, 프롬프트 내용에 따라 특징을 선택적으로 강조하거나 억제하도록 조정한다. UNet은 이를 기반으로 타깃 소스에 해당하는 마스크를 예측하고, 이 마스크가 입력 스펙트로그램에 곱해진 뒤 ISTFT를 통해 파형이 복원된다. AudioSep은 의미적 정렬 능력이 높지만 시간 정보는 포함하지 않으므로, SED 기반 마스킹이 불필요한 구간의 과분리를 억제하는 핵심 역할을 한다.

반면 Diffusion-Based Separation은 VAE 인코더와 latent diffusion을 결합한 텍스트 조건부 오디오 생성 모델이며, 본 연구에서는 FlowSep을 기본 모델로 사용한다. 마스킹된 신호는 먼저 로그 멜 스펙트로그램으로 변환된 뒤, VAE 인코더를 거쳐 저차원 latent로 압축된다. 텍스트 프롬프트는 별도의 텍스트 인코더에서 임베딩되어 확산 UNet의 cross-attention 조건으로 주입되며, UNet은 초기 잡음 상태의 latent를 점진적으로 디노이징하여 프롬프트에 부합하는 latent를 생성한다. 생성된 latent는 VAE 디코더를 통해 멜 스펙트로그램으로 복원되고, 마지막으로 Vocoder를 통해 파형으로 변환된다.

3.4 STFT 기반 신호 결합

AudioSep은 32 kHz 환경에서 전 구간에 대해 마스크 기반 분리를 수행하는 반면, FlowSep은 16 kHz 환경에서 약 10초 길이의 세그먼트를 무작위로 잘라 확산 기반으로 재생성하는 구조이기 때문에 두 모델은 입력 해상도, 시간 범위, 출력 레벨에서 불일치가 발생한다. 따라서, STFT 결합 전에 시간축 정렬과 레벨 보정을 통해 두 신호를 공통 기준으로 맞추는 절차가 필요하다.

두 방식으로 각 출력의 시간 범위 불일치와 레벨 차이를 해결하기 위해, EventSep은 먼저 AudioSep 출력(32 kHz)을 FlowSep 출력에 맞추어 16 kHz로 리샘플링하고, FlowSep이 생성한 세그먼트의 무작위 시작 위치에 해당하는 AudioSep 구간을 절단하여 두 신호의 시간 축을 정렬한다. 이어서 두 신호의 RMS 값을 계산하여 FlowSep 신호를 AudioSep 세그먼트와 동일한 에너지 수준으로 정규화함으로써, 확산 기반 생성 과정에서 발생하는 과도한 진폭 차이를 보완한다.

정렬과 정규화가 완료된 두 신호는 STFT를 적용하여 시간 주파수 도메인으로 변환된 뒤, EventSep은 이 공간에서 두 결과의 magnitude만을 주파수별로 결합한다. 위상은 일관성을 위해 AudioSep의 위상을 사용하며, 결합된 복소 스펙트럼은 ISTFT를 통해 파형으로 복원된다. 상대 구간은 AudioSep 16 kHz 신호 위에 overlap-add 방식으로 삽입되며, 최종적으로 전체 파형을 32 kHz

로 되돌려 출력한다. 이를 통해, 두 모델의 출력이 시간·에너지·주파수 축에서 정확히 대응되도록 만들기 위한 절차이며, 이를 통해 AudioSep의 의미적 분리력과 FlowSep의 고주파 복원력을 안정적으로 단일 파형에 통합할 수 있다.

4 실험 구성

4.1 데이터셋

EventSep은 fine-tuning 없이 사전 학습된 모델(AudioSep, FlowSep, PANNs)을 단순 조합하여 동작하므로, 모델의 일반화 능력을 정확히 평가하기 위해 상호 다른 분포를 갖는 세 가지 벤치마크를 사용하였다.

4.1.1 VGGSound [18]

VGGSound는 200개 클래스, 18,928개의 10초 오디오-비디오 클립으로 구성되며, 대표적인 소리로는 “blowing candles”, “playing saxophone”과 같은 일상적 행동 기반 이벤트가 포함된다. 샘플은 실제 행동(action) 기반으로 수집되어 도메인 노이즈가 크며, AudioSet(527 classes) 중심으로 학습된 PANNs와 클래스 구조가 다르다. 이러한 특성은 EventSep이 프롬프트와 AudioSet 클래스 간의 의미적 연결을 얼마나 안정적으로 수행하는지 평가하는 데 적합한 조건을 제공한다.

4.1.2 MUSIC Dataset [19]

MUSIC dataset은 총 718개의 10초 단일 악기(solo) 연주 클립으로 구성되며, 피아노·바이올린·기타 등 명확한 harmonic 구조와 spectral coherence가 뚜렷한 음악 도메인으로 구성되어 있다. EventSep은 “female vocal”, “acoustic guitar”과 같은 음악 프롬프트 입력 시 SED, AudioSep, 그리고 FlowSep의 조합이 조화적 구조 보존에 어떤 영향을 주는지를 평가하기 위해 사용하였다.

4.1.3 ESC-50 [20]

ESC-50은 50개 환경음 클래스(총 2000개, 5초 길이)로 구성된 단일 이벤트 기반(monophonic) 데이터셋으로, “dog bark”, “door knock”과 같은 구체적 환경음을 포함한다. Background가 거의 없어 SED 기반 시간 선택성(time-selective masking)이 분리 품질에 어떤 기여를 하는지를 직접적으로 관찰할 수 있다.

4.2 평가 지표 및 실험 환경

본 연구에서는 EventSep의 핵심 특성인 텍스트 기반 선택적 음원 분리를 평가하기 위해 텍스트-오디오 의미적 정렬도를 측정하는 CLAPScore[4]와 음성 명료도(intelligibility)를 평가하는 STOI 및 ESTOI[21, 22]를 중심 지표로 사용하였다.

전통적 분리 지표인 SDR 및 SI-SDR[8, 23]은 정답 소스와의 시간 정렬(time-alignment)에 매우 민감하므로, 텍스트 제어 및 확산 기반 구조를 갖는 EventSep에서는

해석이 왜곡될 수 있다. 수 밀리초 수준의 미세한 시점 이동만으로도 점수가 크게 변동하며, 텍스트 기반 분리에서는 타깃 신호가 하나로 고정되지 않아 지표 해석 자체가 모호해진다. 따라서 본 연구에서는 SDR 계열 지표를 참고적 수준으로만 활용하였다.

대신, 분리 결과가 프롬프트와 얼마나 의미적으로 일치하는지를 평가하는 CLAPScore 및 CLAPScoreA[4]를 핵심 기준으로 삼았다. CLAPScore는 텍스트 임베딩과 오디오 임베딩을 동일한 표현 공간에서 비교하여 두 벡터가 얼마나 유사한 방향을 가리키는지를 측정하며, 분리된 신호가 프롬프트 의미와 어느 정도 대응하는지를 직접적으로 평가한다. CLAPScoreA는 프롬프트 대신 정답 오디오(GT)의 임베딩을 기준으로 계산된다. 특히 이를 지표는 시간 정렬에 의존하지 않으므로, 정답 파형을 엄밀하게 추종하기보다는 프롬프트와 의미적으로 유사한 음향 패턴을 생성하는 FlowSep의 출력 특성을 평가할 때도 안정적으로 적용된다는 점에서 EventSep 분석에 적합하다.

한편, STOI와 ESTOI[21, 22]는 음성 명료도(intelligibility)를 측정하는 지표로, 음성 프롬프트가 포함된 실험에서 분리된 신호의 이해도를 보조적으로 확인하는 용도로 사용하였다. 고주파 자음 성분의 보존이나 왜곡 정도 등, 복원 과정이 음성 명료도에 미치는 영향을 진단하는 데 유효하다.

실험은 NVIDIA RTX 5080 (Blackwell, sm_120) GPU가 장착된 WSL Ubuntu 22.04 환경에서 Docker 기반으로 수행하였으며, PyTorch 2.10 Nightly와 CUDA 12.8 조합을 사용하였다.

5 실험 결과

본 장에서는 EventSep의 전체적인 분리 성능을 다양한 데이터셋(VGGSound, MUSIC, ESC-50)을 대상으로 정량적으로 평가한다. 텍스트 기반 분리 모델(AudioSep), 확산 기반 모델(FlowSep), 그리고 제안 기법(EventSep)을 동일한 실험 조건에서 비교함으로써, EventSep이 제공하는 의미적 정렬 성능과 청감 품질 향상이 어떤 방식으로 나타나는지 살펴본다. 이를 위해 CLAPScore 및 CLAPScoreA를 중심으로 분리 결과의 의미적 유사도를 측정하고, STOI와 ESTOI를 통해 음성 프롬프트가 포함된 구간에서의 명료도 변화를 분석한다.

Table 1은 세 데이터셋에 대한 전체 성능 비교 결과를 요약한다. VGGSound 기준으로 AudioSep 대비 EventSep(SED soft mask 적용)은 CLAPScore가 약 25% 향상되었으며(0.2838에서 0.3549로 증가), 이는 텍스트 기반 분리 모델이 가지는 대표적 한계인 시간적 비정렬(temporal misalignment) 문제가 SED 기반의 시간 선택적 필터링(time-selective filtering)을 통해 효과적으로 완화되었음을 의미한다. 특정 소리가 존재하지 않는 구간에서 발생하는 과분리(over-separation)를 프레임 단위로 억제함으로써 의미적 정렬 성능이 안정적으로 개선된 것이다.

AudioSep은 세 데이터셋 모두에서 가장 높은 음성 명료도 지표(STOI, ESTOI)를 기록하며, 특히 MUSIC

Model	VGGSound				MUSIC				ESC-50			
	CLAP↑	CLAP-A↑	STOI↑	ESTOI↑	CLAP↑	CLAP-A↑	STOI↑	ESTOI↑	CLAP↑	CLAP-A↑	STOI↑	ESTOI↑
AudioSep	0.2838	<u>0.7881</u>	0.6408	0.5790	0.3497	<u>0.8830</u>	0.6973	0.6325	<u>0.4426</u>	<u>0.7700</u>	<u>0.6976</u>	<u>0.5998</u>
FlowSep	<u>0.3222</u>	0.7344	0.4278	0.3429	0.2083	0.7108	0.4534	0.3615	0.3545	0.6332	0.4569	0.3290
Ours	0.3549	0.8199	0.6403	<u>0.5786</u>	<u>0.3494</u>	0.8832	<u>0.6969</u>	<u>0.6321</u>	0.4433	0.7705	0.6977	0.5999

Table 1: 전체 모델 성능 비교 (VGGSound, MUSIC, ESC-50). 모든 모델은 동일한 입력 오디오, 동일한 텍스트 프롬프트, 동일한 평가 지표를 기준으로 평가하였다. 각 지표별 최고 성능은 굵은 글씨, 차상위 성능은 밑줄로 표시하였다.

과 ESC-50에서는 CLAPScore 역시 EventSep과 거의 동등한 수준으로 유지된다. 반면 FlowSep은 생성 기반 구조로 인해 일부 도메인(VGGSound)에서 CLAPScore 가 상승하지만, 나머지 지표에서는 일관된 성능을 보이지 못하며, 타깃 신호를 염밀하게 분리하기보다는 유사한 음향 패턴을 생성하는 방식으로 작동하여 도메인에 따라 의미 정렬이 향상되기도 하지만 분리 선명도나 명료도는 불안정해지는 한계를 가진다.

이에 비해 EventSep은 AudioSep 대비 의미 정렬이 보다 정확하게 유지되면서도, FlowSep 대비 음성 명료도 저하가 거의 발생하지 않는다. 이는 AudioSep이 제공하는 명확한 텍스트 기반 분리 특성과 FlowSep의 고주파 복원 및 스펙트럼 매끄러움이 결합되며, SED 기반 시간 선택성이 두 신호의 조합을 안정적으로 지지하기 때문이다. 결과적으로 EventSep은 의미적 정합성과 지각적 자연스러움 사이의 균형점을 형성하여, 분리 품질 전반에서 두 모델을 단순히 별별로 사용하는 경우보다 일관적이고 해석 가능한 출력을 제공한다.

6 Ablation 연구

6.1 SED 마스킹 방식

마스킹 자체의 영향을 분리해 관찰하기 위해 AudioSep 단독 모델에 대해서만 mask와 threshold를 변화시키는 방식으로 실험을 진행하였다. Table 2는 임계값(threshold) 변화에 따른 Hard/Soft Masking 비교 결과이다. 본 실험에서는 threshold $\tau \in \{0.55, 0.60, 0.65, 0.70\}$ 범위를 대상으로 하였으며, masking 방식은 hard, soft 중에서 평가하였다. Threshold가 낮을수록 프레임 단위 검출의 재현률이 증가하고, threshold가 높을수록 오탐율을 억제하여 정밀도가 높아지는 구조이므로, EventSep의 시간 선택성 안정성은 이 선택에 직접적으로 영향을 받는다.

Hard masking은 프레임 단위 확률이 임계값을 넘는지 여부만으로 이진 마스크를 구성하므로, threshold 변화에 매우 민감하며 확률 분포가 희소한 경우 특정 구간이 완전히 제거되는 문제가 발생하기 쉽다. 반면 soft masking은 확률값의 크기를 연속적으로 반영하는 방식이기 때문에 작은 변화도 점진적으로 표현되며, threshold가 달라져도 마스크의 형태가 급격히 무너지지 않는다. 이러한 구조적 차이로 인해 soft masking은 threshold가 0.55일 때 SDRi(8.901)와 SI-SDR(8.975)이 모두

Hard Masking	SDRi↑	SI-SDR↑
0.70	6.698	4.753
0.65	5.980	3.738
0.60	5.609	2.974
0.55	8.876	8.847
Soft Masking	SDRi↑	SI-SDR↑
0.70	7.252	8.602
0.65	7.813	8.714
0.60	8.386	8.846
0.55	8.901	8.975

Table 2: Hard vs. Soft Masking (VGGSound).

가장 높게 나타나며, EventSep의 시간 선택적 조건부 처리를 가장 안정적으로 구현하는 방식으로 확인되었다.

본 실험에서는 hard와 soft masking 모두 동일한 방식으로 SED 확률을 시간축에 맞게 확장하여 적용하였으며, threshold 변화가 EventSep의 시간 선택성 안정성에 어떤 영향을 미치는지 확인하기 위해 동일한 조건을 유지하였다.

6.2 Ensemble 방식에 따른 성능 분석

본 연구에서는 AudioSep과 FlowSep이 분리 단계에서 드러나는 서로 다른 강점(의미적 정확도, 청각적 자연스러움)에 기반하여, 각 모델에서 출력된 오디오의 STFT 크기(magnitude)를 조합하는 다섯 가지 앙상블 구조를 설계하였다. 모든 방식에서 위상은 AudioSep의 위상 $\angle S_A(f, t)$ 을 사용하고, 결합은 오디오 크기(magnitude)에 대해서만 이루어진다.

앙상블 단계 이전에, AudioSep 출력은 32 kHz에서 16 kHz로 리샘플링되어 FlowSep 출력과 동일한 해상도에서 정렬된다. FlowSep이 생성하는 파형은 무작위 시작점 s 에서 길이 L 만큼 생성되므로 AudioSep 파형의 동일 구간 $x_A[s : s + L]$ 을 절단하여 시간축을 맞춘다. 또한 RMS 에너지 차이를 보정하기 위해

$$\text{flow}_{\text{norm}} = \text{flow} \cdot \frac{\text{RMS}(x_A)}{\text{RMS}(\text{flow})}$$

을 수행한 뒤, 두 신호에 대해 STFT를 계산한다.

$$S_A(f, t) = \text{STFT}(x_A)$$

$$S_F(f, t) = \text{STFT}(\text{flow}_{\text{norm}})$$

Table 3는 앞서 설계한 다섯 가지 Ensemble 방식에 대해 VGGSound 기준으로 CLAPScore와 CLAPScoreA를 비교한 결과를 제시한다.

Method	CLAPScore↑	CLAPScoreA↑
Rate-weighted	0.3396	0.7979
Band-split	0.3396	0.8009
Reverse Band-split	0.3394	0.8011
Progressive	0.3398	0.7977
Reverse Progressive	0.3387	0.8009

Table 3: Ensemble 성능 비교 (VGGSound).

본 실험을 통해 다섯 가지 Ensemble 방식이 서로 다른 형태의 주파수별 trade-off를 달성을 확인할 수 있었다. 먼저 Rate-weighted 방식은 AudioSep과 FlowSep의 비율을 고정 가중으로 조절함으로써, 두 모델의 semantic-perceptual balance를 결정하는 기본적인 기준점 역할을 한다. 이에 비해 Band-split 및 Reverse Band-split 방식은 주파수 대역을 명확히 분리하여 결합하는 전략으로, 전자는 저주파에서 AudioSep의 안정적인 재현력을 유지하고 고주파에서 FlowSep의 세밀한 질감을 활용하는 반면, 후자는 그 반대의 조합을 통해 상보적 특성을 다른 방식으로 배치한다. 마지막으로 Progressive와 Reverse Progressive 방식은 주파수 상승에 따라 두 모델의 기여도가 점진적으로 변하는 구조를 갖추고 있으며, 이러한 점진적 혼합이 특정 구간의 잡음을 부드럽게 완화하는 효과를 제공함을 관찰할 수 있었다.

결론적으로 Ensemble은 성능을 크게 바꾸는 요소는 아니지만, 특정 대역에서 발생하는 잡음을 완화하여 보다 자연스럽고 안정적인 출력을 만드는 데 기여한다.

종합하면, EventSep은 의미적 정확도, 시간적 일관성, 청각적 자연스러움의 세 측면을 동시에 확보하는 텍스트 기반 선택적 음원 분리 모델임이 실험적으로 확인되었다.

7 논의

EventSep은 텍스트 의미 기반 분리(AudioSep), 프레임 단위 시간 정보(SED), 그리고 확산 기반 복원(FlowSep)을 결합함으로써 기존 텍스트 기반 분리가 겪는 시간적 비정렬과 과분리 문제를 구조적으로 완화하였다. 특히 SED soft masking은 타깃 이벤트가 존재하지 않는 구간에서의 불필요한 분리를 억제해 의미적 정렬 성능을 크게 개선하는 핵심 요소로 확인되었다.

다만, EventSep은 PANNs-SED의 AudioSet 527개 라벨 체계에 구조적으로 의존하기 때문에, 세분화된 악기 구분이 필요한 MUSIC 등 도메인 특화 환경에서는 시간 선택성의 정밀도가 제한된다. AudioSet 라벨은 “Musical instrument”처럼 상위 범주 중심의 구성이

많아, 실제 데이터에 존재하는 다양한 악기(예: violin, cello, flute, clarinet, trumpet 등)가 모두 동일한 상위 라벨로 통합되는 경우가 많다. 이처럼 하나의 AudioSet 라벨이 여러 실제 악기 클래스와 1:다 관계를 이루기 때문에, SED가 특정 악기를 과도하게 활성화하거나 반대로 놓치는 사례가 발생할 수 있다. 또한, 프롬프트를 AudioSet의 단일 클래스에 매핑하는 과정에서 미세한 의미 차이가 축약되며, 복합적 의미를 포함한 프롬프트에서는 정보 손실이 불가피하다.

또한, FlowSep과의 결합은 고주파 복원과 지각적 자연스러움을 향상시키지만, 확률적 생성 특성으로 인해 원 신호에 없던 미세 신호가 생성되는 경우가 존재한다. 이러한 현상은 전체적인 청감 품질을 높여 주는 한편, 특정 상황에서는 음질의 일관성이 완전히 유지되지 않을 수 있다. 이는 EventSep이 의미적 정렬과 지각적 자연스러움 사이에서 균형을 달성하고 있음에도 불구하고, 복잡한 프롬프트 처리나 세밀한 조건 기반 분리에서는 여전히 해결해야 할 과제가 남아 있음을 보여준다.

마지막으로 본 연구에서는 주로 단일 개념 중심 프롬프트에 초점을 두었으며, 부정적 프롬프트(negative prompt)나 문맥 기반 복합 프롬프트(complicated context prompt) 같은 다양하고 복잡한 언어적 조건을 충분히 다루지 못했다. 향후 LLM 기반으로 다양한 프롬프트 변형을 생성하고, 프롬프트-오디오 쌍을 대량으로 증강(Data Augmentation)하는 방향으로 확장할 경우, 텍스트 기반 분리의 제어력과 일관화 성능을 크게 향상시킬 수 있을 것이다.

8 결론

본 연구는 텍스트 기반 음원 분리가 갖는 구조적 한계로 인해 안정적이고 정밀한 선택적 분리가 어려운 문제를 다루었다. 이러한 문제를 해결하기 위해 본 논문은 SED 기반 시간 선택성, 텍스트 기반 및 확산 기반 음원 분리 모델의 병렬 구조 기반의 EventSep을 제안하였다. EventSep은 추가 학습 없이도 텍스트 기반 의미 표적화, SED 기반 시간 선택성, 확산 기반 지각적 복원을 통합하여 기존 AudioSep 대비 더 높은 의미적 정확도와 자연스러운 청감 품질을 제공하였다. 특히 SED soft masking은 텍스트 기반 분리의 구조적 한계였던 temporal misalignment와 과분리를 효과적으로 억제함으로써 EventSep의 성능 향상에 핵심적 역할을 하였다.

References

- [1] Qiuqiang Kong, Yin Cao, Tariq Iqbal, Yuxuan Wang, and Mark D. Plumley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2020.
- [2] Haohe Liu, Yi Zhu, Qiuqiang Kong, Xubo Li, Wenwu Wang, and Mark D. Plumley. Audiosep:

- Open-domain audio source separation via text prompting. *arXiv preprint arXiv:2308.04136*, 2023.
- [3] Yichong Zhang, Yuchen Li, Yatong Du, Xubo Chen, and Qiuqiang Kong. Flowsep: Latent diffusion models for audio source separation. *arXiv preprint arXiv:2402.03578*, 2024.
- [4] Benjamin Elizalde, Shrey Deshmukh, and et al. Clap: Learning audio concepts from natural language supervision. *arXiv preprint arXiv:2206.04769*, 2022.
- [5] Alec Radford, Wei Shang, Erick Gutierrez, and et al. Clipsep: Learning to separate audio with clip-based audio-text instruction. *arXiv preprint arXiv:2309.00908*, 2023.
- [6] Min Xu, Haohe Liu, Qiuqiang Kong, and et al. Language-audio source separation. *arXiv preprint arXiv:2305.09598*, 2023.
- [7] Xubo Li, Xubo Chen, Qiuqiang Kong, and Wenwu Wang. Sepdrafter: Draft-and-revise for text-guided audio source separation. *arXiv preprint arXiv:2401.02959*, 2024.
- [8] Stacey Quackenbush et al. Objective measures for speech quality assessment. *Signal Processing*, 1992.
- [9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.
- [10] Haohe Liu, Xubo Chen, Wei Wu, Wenwu Wang, Mark D. Plumbley, and Yuxuan Lu. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.
- [11] Haohe Liu, Xubo Chen, and et al. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2401.01039*, 2024.
- [12] Daniel Stoller, Thomas Richards, Santiago Castro, and et al. Stable audio: Versatile audio generation via latent diffusion models. *arXiv preprint arXiv:2311.15090*, 2023.
- [13] Ke Chen, Qiuqiang Kong, Yatong Du, Jian Yang, and Xubo Chen. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- [14] Khaled Koutini, Jan Schlüter, and Gerhard Widmer. Passt: Efficient audio classification with patchout spectrogram transformer. *arXiv preprint arXiv:2110.05069*, 2021.
- [15] Jort F. Gemmeke, Daniel P. W. Ellis, and et al. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.
- [16] Wenhui Wang and Furu Wei. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*, 2020.
- [17] Zhihua Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.
- [18] Honglie Chen, Weidi Xie, and Andrea Vedaldi. Vggssound: A large-scale audio-visual dataset. *arXiv preprint arXiv:2007.00274*, 2020.
- [19] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh H. McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. MUSIC dataset is released as part of the project repository.
- [20] Karol J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018, 2015.
- [21] Cees H. Taal et al. A short-time objective intelligibility measure for time-frequency weighted signals. *ICASSP*, 2011.
- [22] Jesper Jensen et al. An extended short-time objective intelligibility measure. *IEEE Transactions on Audio, Speech, and Language Processing*, 2015.
- [23] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Scale-invariant signal-to-distortion ratio: Fast computation and optimization. *IEEE ICASSP*, 2019.