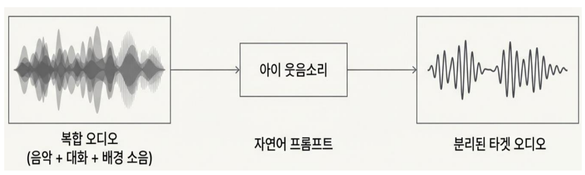
EventSep: 사전 학습 모델 결합 기반의 선택적 음원 분리

박경빈
Department of Electrical Engineering, Korea University, 145 Anam-ro, Seongbuk-gu Seoul 02841, Republic of Korea (overjoy1008@korea.ac.kr)

INTRODUCTION

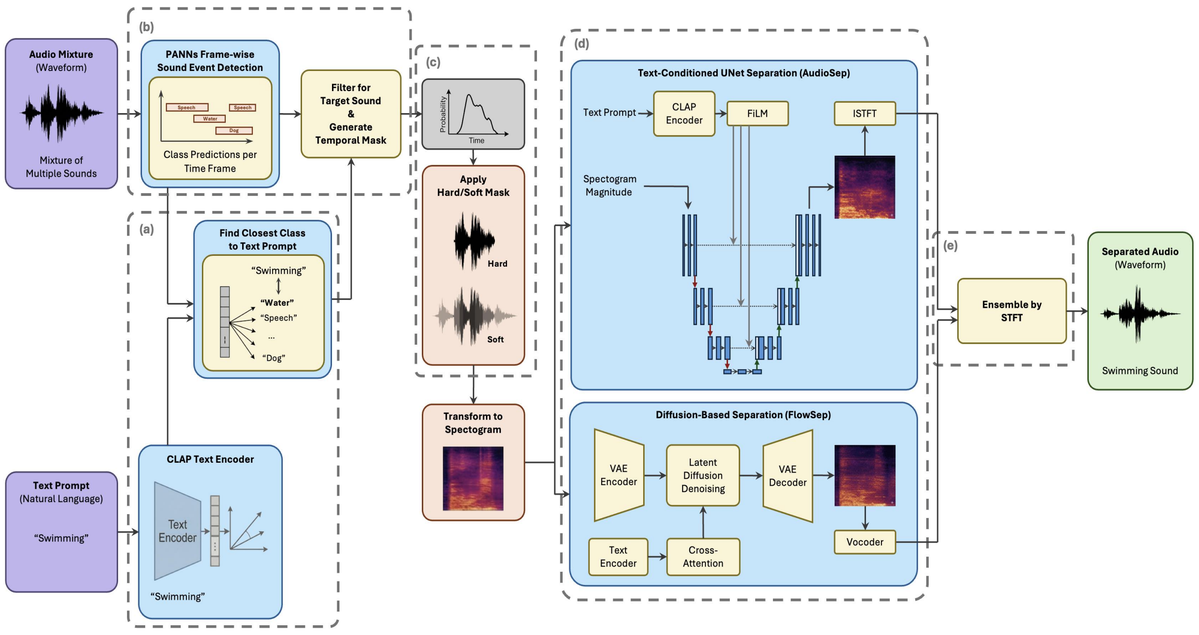


텍스트 조건 음원 분리 (Text-guided Source Separation)

- 여러 소리가 섞인 복합 오디오에서 사용자가 자연어로 지정한 오디오만을 추출하는 기술
  - 예) '아이의 웃음소리만 남겨줘', '바이올린 연주 부분만 추출해줘'
- 자연어 프롬프트는 무슨(WHAT) 대상을 분리해야 할지 정보는 존재하나, 언제(WHEN) 존재하는지의 정보 부재
- 기존의 텍스트 조건 음원 분리 연구는 시간적 비정렬, Over-Separation, Hallucination 등의 한계가 존재

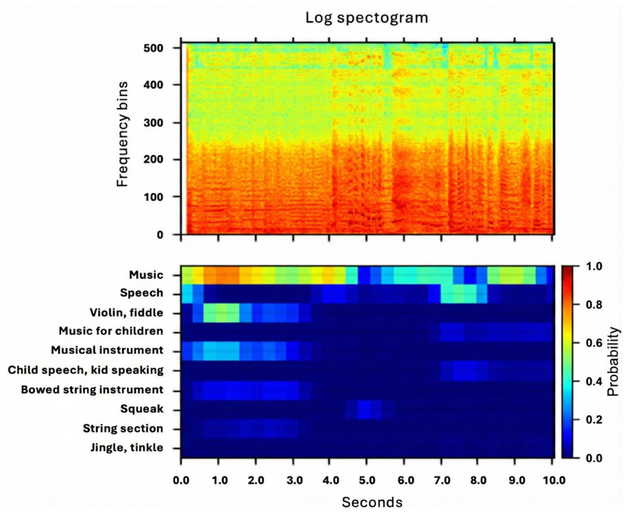
METHODS

Architecture



(a) 텍스트 임베딩과 타깃 클래스 간 매핑
(b) PANNs 기반 프레임단위 SED (Sound Event Detection)
(c) SED 확률을 이용한 시간 선택적 마스킹
(d) Text-Conditioned UNet Separation, Diffusion-Based Separation의 병렬 분리
(e) STFT 기반 신호 결합 단계

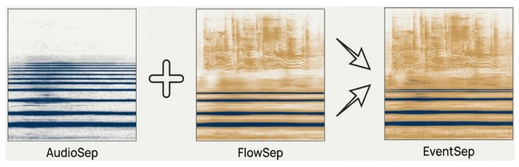
PANNs Frame-wise SED (Sound Event Detection)



- SED는 오디오 신호 속에서 어떤 소리가 포함되어 있는지, 그리고 그 소리가 언제 등장하고 사라지는지를 추정하는 기술
- Frame-wise SED: 각 시간 프레임 단위로 소리의 존재 확률을 추정
- 본 연구에서는 PANNs-Cnn14 모델을 사용
- PANNs는 정해진 AudioSet 527 클래스(speech, dog, water, footsteps, applause, musical instrument 등)에 대해서만 확률 예측이 가능하므로, Text Encoder로 텍스트 프롬프트-클래스 간 내적을 계산하여 가장 유사한 클래스 지정
- 타깃 클래스의 프레임 단위 존재 확률을 얻어 입력 오디오에 선택적 마스킹을 적용

텍스트 기반 & 확산(Diffusion) 기반 모델의 병렬 구조

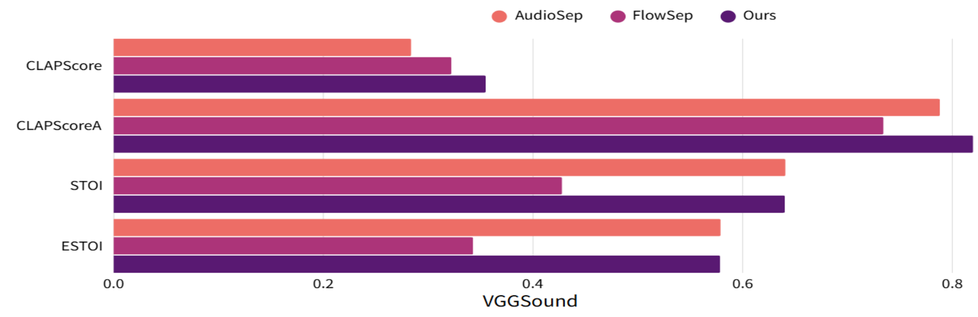
- **Text-Conditioned UNet Separation** (텍스트 기반 음원 분리)
  - STFT(Short-Time Fourier Transform) 기반의 UNet 내에서 주어진 오디오를 주파수별로 쪼갠 뒤, 텍스트 프롬프트를 각 단계별로 적용하여 이와 일치하는 주파수만을 통과시키는 방식
  - 프롬프트의 의미를 정밀하게 해석하며, 매우 선명한 음질을 출력
  - <u>무엇을(WHAT)</u> 분리할 것인가에 특화
- **Diffusion-Based Separation** (확산 기반 음원 분리)
  - VAE, Latent Diffusion Denoising, VAE Decoder, Vocoder로 구성
  - 스펙토그램 이미지를 인코딩하여 확산(Diffusion) 과정을 거친 뒤 디코딩하는 방식을 통해 분리된 음원의 스펙토그램을 복원해내는 방식
  - 음원의 음색, 고주파 성분, 미세한 질감을 자연스럽게
  - <u>어떻게(HOW)</u> 분리할 것인가에 특화
- **STFT 기반 신호 결합**
  - 두 신호를 리샘플링(16k로 통일), 시간축 정렬, RMS로 정규화한 후 STFT를 통해 결합
  - AudioSep의 의미적 분리력과 FlowSep의 고주파 복원력을 안정적으로 단일 파형에 통합

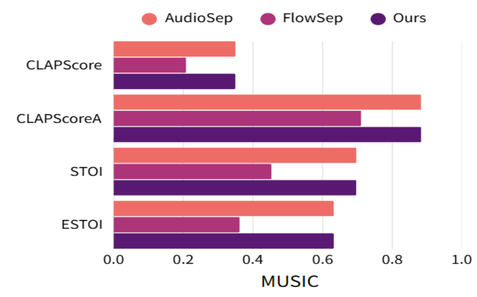


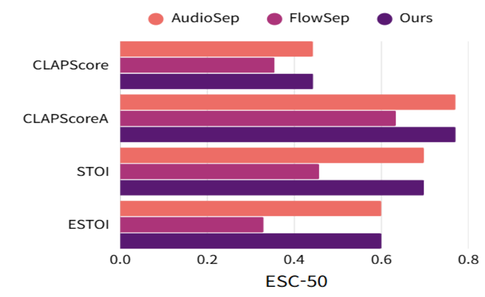
RESULTS

Quantitative comparisons

- 모델 성능 비교
  - 텍스트 기반 분리 모델(AudioSep 단독), 확산 기반 분리 모델(FlowSep 단독), Ours(EventSep)
  - 데이터셋: VGGSound(일상적 행동 기반 오디오 클립), MUSIC(악기 연주 기반), ESC-50(환경음 기반)
  - 평가 지표: CLAPScore & CLAPScoreA(텍스트와 오디오 간 의미적 유사도), STOI & ESTOI(음성 명료도)



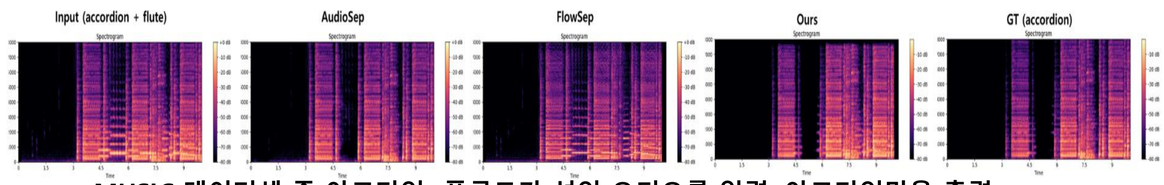




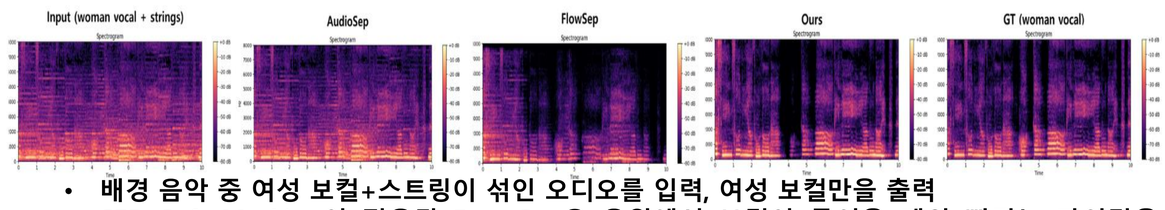
- VGGSound에서 AudioSep 대비 <u>EventSep의 CLAPScore 25% 향상 (0.2838→0.3549)</u>
- AudioSep 대비 의미 정렬이 보다 정확하게 유지되면서도, FlowSep 대비 음성 명료도 저하가 거의 발생하지 않음

Qualitative comparisons

- 오디오 스펙토그램 비교



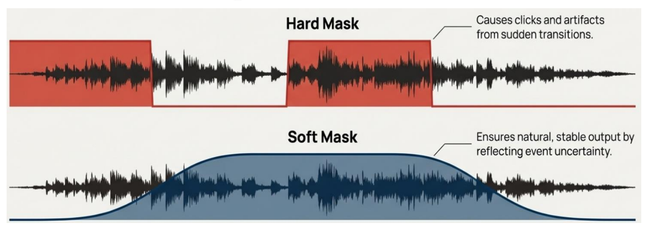
- MUSIC 데이터셋 중 아코디언+플루트가 섞인 오디오를 입력, 아코디언만을 출력
- 아코디언이 없는 부분을 잘 분리한 AudioSep(2번째), 고음질에서 비슷한 형태를 구현한 FlowSep(3번째), 두 모델의 장점을 결합한 EventSep(4번째)의 스펙토그램



- 배경 음악 중 여성 보컬+스트링이 섞인 오디오를 입력, 여성 보컬만을 출력
- SED Soft Masking이 적용된 EventSep은 음원에서 보컬이 들어올 때와 빠지는 타이밍을 비교적 잘 포착

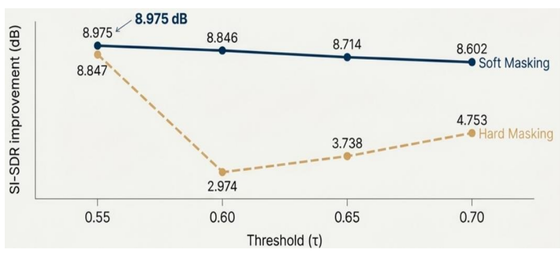
Ablation Study

- SED Masking 기법에 따른 영향 분석



Hard Mask
- SED에서 얻은 클래스 존재 확률이 Threshold보다 크면 1, 작으면 0

Soft Mask
- SED에서 얻은 클래스 존재 확률값에 따라 0~1까지 연속적이고 부드러운 마스킹 적용



- Soft Masking & Threshold=0.55 일 때 SDRi(8.901)와 SI-SDR(8.975)가 가장 높게 측정 – 소리의 시작과 끝이 모호한 경우에도 안정적인 결과 출력
- Hard Masking은 Threshold를 기준으로 신호를 켜고 끄기 때문에, 오디오 단절로 인한 잡음 유발 가능성이 높음

- **Ensemble 기법**
  - 텍스트 기반 및 확산 기반 음원 분리 모델의 두 결과를 합치는 기법에 따른 영향 분석
  - 정해진 비율대로 섞는 Rate-weighted, 주파수 대역을 고/저주파로 분리하여 각 모델이 담당하는 Band-Split, 주파수 상승에 따라 두 모델의 기여도가 점진적으로 변하는 Progressive
  - Ensemble은 성능을 크게 바꾸는 요소는 아니지만, 특정 대역에서 발생하는 잡음을 완화하여 보 다 자연스럽고 안정적인 출력을 만드는 데 기여

CONCLUSIONS

- 텍스트 기반 음원 분리의 한계(모호성·시간 부재·환각)를 해결하기 위해 텍스트 기반 의미 표적화, SED 기반 시간 선택성, 텍스트 기반 및 확산 기반 모델의 병렬 구조, 신호 결합과 Ensemble로 구성된 EventSep 프레임워크 제안
- 텍스트만으로는 알 수 없었던 WHEN(오디오가 언제 등장하는가)의 정보를 SED를 통해 보완
- 사전 학습 모델의 구조적 설계만으로 의미적 정확도, 시간적 일관성, 청각적 자연스러움을 동시에 확보하는 안정적인 오디오 분리가 가능함을 입증