멀티모달 생성형 인공지능에서 사운드 기반 생성의 문화적 편향 분석

이혜진[1O], 진솔[1], 김형일[2]
[1]전남대학교 컴퓨터정보통신공학과, [2]전남대학교 전자컴퓨터공학부
{ilseeu01, hyungil.kim}@hanguk.ac.kr

요 약

본 연구는 멀티모달 생성형 인공지능에서 사운드 기반 생성에서 발생하는 문화적 편향을 분석한다. 특히 sound-to-image 및 text-to-sound 생성 태스크를 대상으로, 서양권 데이터로 학습된 모델이 비서구 문화적 맥락을 어떻게 반영하는지를 실험적으로 평가한다. 한국 문화 요소가 포함된 입력을 구성하고, 생성 결과를 품질과 의도 정합성 관점에서 분석한 결과, 모델이 문화적 특성을 충분히 반영하지 못하고 서구 중심적 표현으로 치환하는 경향을 확인하였다. 본 연구는 사운드가 포함된 멀티모달 생성 환경에서의 문화적 편향 문제를 정리하고, 향후 공정한 멀티모달 생성 모델 설계를 위한 기초 분석을 제공한다.

1. 서론

최근 대규모 딥러닝 모델의 발전과 함께 이미지, 오디오, 텍스트를 동시에 처리할 수 있는 멀티모달 생성형 인공지능(multimodal generative AI) 모델이 빠르게 발전하고 있다. 이러한 모델들은 이미지 생성, 사운드 합성, 텍스트 기반 음성 생성, 그리고 서로 다른 모달리티 간 변환(text-to-image, image-to-audio 등)과 같은 다양한 응용 분야에서 활용되며, 범용 생성 모델로서의 가능성을 확대하고 있다.

그러나 현재 널리 사용되는 멀티모달 생성형 AI 모델들은 주로 서양권 중심의 대규모 데이터셋을 기반으로 학습됐다는 한계를 가진다. 데이터의 양은 방대하지만, 문화적·지역적 다양성은 상대적으로 충분히 반영되지 못하는 경우가 많으며, 특히 비서구권 문화나 저자원(low-resource) 환경에 대한 표현은 제한적인 경향을 보인다. 이에 따라 모델이 생성하는 이미지나 음향, 또는 이들 간의 변환 결과가 입력의 문화적 맥락을 정확히 반영하지 못하거나, 특정 문화적 요소를 서구적 관점으로 왜곡하여 표현하는 문제가 발생할 수 있다.

이러한 현상은 단순한 생성 품질 저하를 넘어, 멀티모달 생성형 AI가 내포하는 문화적 편향(cultural bias) 문제로 이어질 수 있다. 특히 문화적 의미가 강하게 내포된 텍스트 입력이나 소리 기반 입력의 경우, 모델이 이를 보편적이거나 서구 중심적인 표현으로 치환하여 생성하는 경향은 실제 활용 환경에서 심각한 해석 오류를 유발할 가능성이 있다. 최근에는 멀티모달 생성형 AI에서 문화적 편향성을 분석하고자 하는 연구들이 수행되고 있으며, 생성 결과에 내재한 문화적 왜곡이나 특정 문화권에 대한 편향적 표현이 어떻게 발생하는지에 대한 논의가 확대되고 있다.

본 연구에서는 서양권 데이터로 학습된 멀티모달 생성형 AI 모델을 대상으로, 소리로부터 이미지를 생성하는(sound-to-image) 과제와 텍스트로부터 소리를 생성하는(text-to-sound) 과제에 집중하여 문화적 편향을 분석한다. 특히 문화적 맥락이 비교적 명확하게 드러나는 입력을 구성하고, 생성된 이미지 및 사운드가 입력의 문화적 의미를 얼마나 충실히 반영하는지를 중심으로 분석한다.

이를 통해 본 연구는 멀티모달 생성 과정에서 사운드가 포함될 때 발생하는 문화적 편향의 특성을 분석하고, 텍스트-이미지 중심에서 벗어나, 사운드 기반 멀티모달 생성 환경에서의 문화적 편향 문제를 이해하는데, 기초 자료로 활용될 수 있을 것이다.

2. 실험방법

2.1 분석 대상

본 연구에서는 사운드 기반 생성 모달리티를 중심으로 분석을 수행한다. 구체적으로, 소리로부터

이미지를 생성하는(sound-to-image, S2I) 태스크와 텍스트로부터 소리를 생성하는(text-to-sound, T2S) 태스크를 대상으로 하며, 각 모달리티별로 대표적인 생성 모델 3종씩 선정하여 총 6개의 모델을 비교 분석 하였다.

S2I 생성 모델로 환경음 또는 상황음을 입력으로 해당 소리가 발생하는 장면을 이미지로 생성하는 Sound2Scene [1] 및 Sound2Vision [1] 모델, 그리고 확산 모델 기반 사운드-이미지 생성 모델로 SonicDiffusion [2]를 채택하였다. T2S 생성 모델로는 텍스트 조건을 기반으로 오디오를 생성하는 확산 모델인 AudioLDM [3], 텍스트와 오디오 간의 멀티모달 사전 학습을 기반으로 한 생성 모델인 Make-An-Audio [4], 그리고 Meta에서 공개한 오디오 생성 프레임워크인 AudioCraft [5]를 채택하였다.

2.2 평가지표

본 연구에서는 멀티모달 생성형 AI에서 나타나는 문화적 편향을 정량적으로 분석하기 위해, 품질(quality), 의도 정합성(intent alignment)의 두 가지 평가 축을 중심으로 평가를 수행한다.

품질 평가는 생성 결과가 시각적/청각적으로 얼마나 자연스럽고 일관적인지를 측정하는 지표로 특정 문화적 표현이 왜곡되거나 과도하게 단순화되어 나타나는지를 간접적으로 확인하기 위한 기준으로 활용한다. S2I 생성 결과에 대해서는 FID를 사용하였으며, 이를 통합적으로 평가하기 위해 EvalGIM [6] 라이브러리를 활용하였다. T2S의 경우 OpenL3 [7] 기반 Fréchet Audio Distance (FAD) [7] 을 사용하여 음향의 자연성과 분포 유사성을 평가했다. 문화적 편향이 존재할 경우, 특정 문화적 입력에 대해 생성 품질이 비정상적으로 저하되거나 특정 표현 양식으로 수렴하는 현상이 나타날 수 있으며, 본 평가지표는 이러한 경향을 정량적으로 확인하는 데 활용된다.

의도 정합성 평가는 입력에 포함된 문화적 의미가 생성 결과에 얼마나 정확히 반영되었는지를 측정하기 위한 핵심 지표이다. 문화적 편향은 단순한 오류가 아니라, 입력된 문화적 맥락이 다른 문화권의 표현으로 치환되거나 일반화되는 방식으로 나타나는 경우가 많다. 이를 분석하기 위해 S2I 태스크에서는 한국 전통 소리 데이터를 입력으로 사용하고, 생성된 이미지와 입력 음향 간의 의미적 일치도를 CLIP Score [7]로 측정한다. T2S 태스크에 대해서는 CLAP Score [7]를 활용하여 텍스트 프롬프트와 생성된 음향 간의 의미 정합성을 평가하였다. 이 과정에서 특정 문화적 입력이 반복적으로 서구적 이미지나 음향으로 치환되는 경우, 이는 모델이 문화적 의도를 정확히 반영하지 못하고 있음을 의미하며, 문화적 편향의 한 형태로 해석된다.

3. 실험 결과 및 분석

3.1 데이터셋 구성

표 1. 텍스트 입력 프롬프트

Pair	Contemporary	Traditional
악기	the sound of a piano	the sound of a gayageum
가사	the sound of a washing machine	the sound of pounding laundry with wooden beaters
종교	the sound of a church bell	the sound of a church bell
쓰기	the sound of a typewriter	the sound of grinding an ink stick on an inkstone

문화적 편향을 측정하기 위해 한국 문화 요소가 포함된 입력 데이터와 서양 문화 요소가 포함된 입력 데이터를 쌍으로 구성하였다.

S2I 생성 모델을 평가하기 위해 AI Hub의 '한국 전통 소리 및 생활 소리 데이터' [8] 을 기반으로 평가용 데이터를 구성하였다. 본 데이터셋에는 타자기, 먹 갈기, 교회 종, 사찰 종, 세탁기 사용, 다듬이질 등 한국 고유의 전통적인 소리를 포함하였다. 비교를 위해 서양 문화권의 유사한 현대적 맥락의 소리도 함께 수집하였다. 또한 Youtube에서 수집한 피아노 및 가야금 연주 음원을 추가로 포함하였다. 각 텍스트 프롬프트에 대해 random seed를 1부터 50까지 변화시켜 총 50장의 이미지를 생성하였으며, 생성 결과의 예시는 그림 1과 그림 2에 제시한다. 정답 이미지 세트는 Google 이미지 크롤링을 통해 동일 수량으로 수집하였다.

T2S 평가를 위한 데이터셋은 문화적 맥락이 명확하게 드러나는 텍스트 프롬프트로 표 1과 같이 구성하였다. 프롬프트는 악기, 가사 노동, 종교적 소리, 쓰기 도구 등으로 분류되며, 각 범주에서 한국 문화와 서양 문화 요소를 대응시키는 방식으로

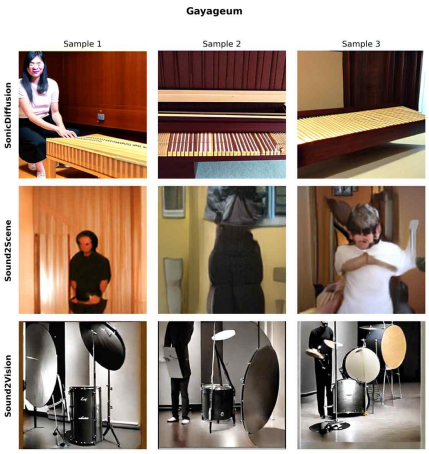

그림 1 가야금 프롬프트를 넣은 output

그림 2 피아노 프롬프트를 넣은 output

설계하였다. 이때 비교 기준이 되는 정답 음원 세트는 AI Hub의 한국 전통 소리 및 생활 소리 데이터를 활용하였다.

3.2 품질 평가 결과

T2S 품질 평가는 OpenL3 기반 Fréchet Distance를 사용하여 수행하였으며, 그 결과는 그림 3에 제시하였다. SonicDiffusion이 전반적으로 가장 낮은 거리를 보여 실제 음향 분포와의 유사도가 가장 높았다. 전반적으로 모델들은 Traditional 범주에서 높은 거릿값을 보여 음향 품질의 불안정성이 관찰되었다.

S2I 품질 평가는 FID를 통해 수행하였고, 결과는 그림 4와 같다. SonicDiffusion이 대부분의 소리 범주에서 가장 낮은 FID 값을 기록하여 전반적으로 우수한 시각적 생성 품질을 보였다. 반면 다른 모델들은 특히 전통적 소리 범주에서 상대적으로 높은 FID와 변동성을 나타냈다.

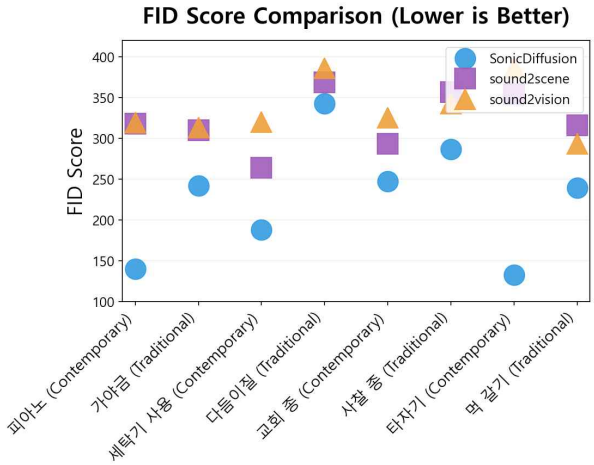

그림 3 S2I 모델의 FID 비교값

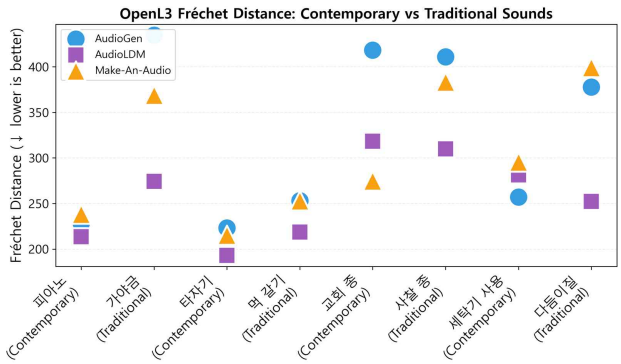

그림 4 T2S모델의 FID 비교값

3.3 정합성 평가 결과

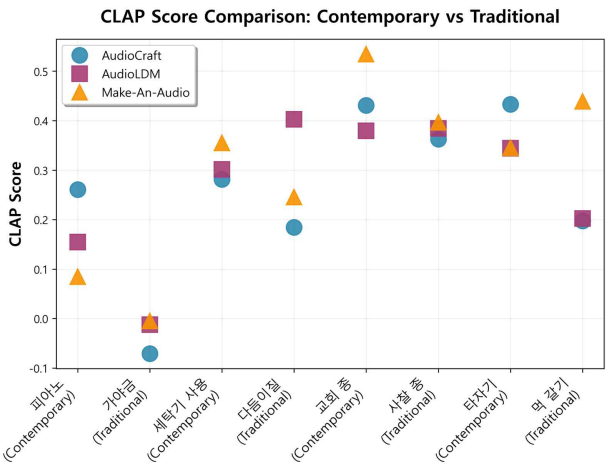

그림 5 T2S 모델의 CLAP 점수

T2S 정합성 평가는 CLAP score를 통해 수행하

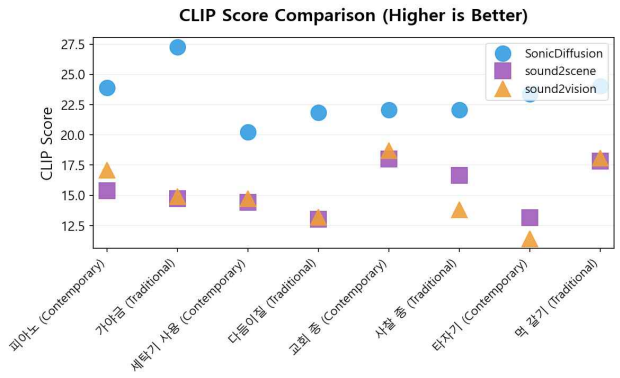

그림 6 I2S 모델의 CLIP 점수

였다. 그림 5에서 확인할 수 있듯이, AudioCraft 모델이 전반적으로 높은 CLAP 점수를 기록하여 텍스트-오디오 의미 정합성이 가장 우수한 것으로 나타났다. 전반적으로 모델들은 Traditional 범주에서 점수 하락이 관찰되었다.

S2I 정합성 평가는 CLIP score를 사용하여 수행하였으며, 그 결과는 그림 6에 제시하였다. SonicDiffusion이 대부분의 범주에서 가장 높은 CLIP 점수를 보여 오디오 입력과 생성 이미지 간 의미적 일치도가 가장 높았다. 다른 모델들은 특정 범주에서 상대적으로 낮은 점수를 보여 정합성의 변동성이 확인되었다.

일부 예외적인 경우를 제외하면, 서구적이고 현대적인 프롬프트일수록 실제 이미지와 더 높은 유사도를 보인다.

본 연구는 멀티모달 생성형 AI의 문화적 편향을 정량적으로 평가하기 위한 프레임워크를 제안하였다. 기존 연구가 T2I에 집중됐지만, 본 연구는 S2I, T2S의 두 가지 모달리티를 동일한 평가 체계로 분석함으로써 멀티모달 AI의 편향 문제를 더욱 폭넓게 조망할 수 있는 기반을 마련하였다.

실험 결과, 서양 중심 데이터로 학습된 모델들이 한국 문화 요소가 포함된 입력에 대해 상대적으로 낮은 의도 정합성을 보이는 것을 확인하였다. 이는 현재 멀티모달 생성 모델들이 비서양 문화권의 데이터를 충분히 학습하지 못했음을 시사하며, 글로벌 환경에서 AI 서비스를 제공할 때 문화적 공정성 문제가 발생할 수 있음을 보여준다. 특히 동아시아 문화권 내에서도 한국, 일본, 중국의 문화적 차이를 세밀하게 구별하지 못하는 경향이 관찰되었으며, 이는 모델이 비서양 문화를 하나의 범주로 뭉뚱그려 처리할 가능성을 제기한다.

4. 결론

본 연구의 한계점으로는 평가 대상 모델의 수가 제한적이라는 점, 그리고 한국 문화에 초점을 맞추어 다른 문화권에 대한 분석이 부족하다는 점을 들 수 있다. 또한 문화적 편향의 원인을 학습 데이터의 불균형으로 추정하였으나, 모델 구조나 학습 방법론 등 다른 요인의 영향에 대해서는 추가적인 분석이 필요하다.

향후 연구에서는 ImageBind와 같은 멀티모달 임베딩 모델을 활용하여 모달리티 간 편향을 직접 측정하는 방법을 탐색할 예정이다. 또한 I2S 등 다른 모달리티로 평가 범위를 확장하고, 동남아시아, 중동, 아프리카 등 다른 비서양 문화권에 대한 편향 분석도 수행할 계획이다. 나아가 편향을 완화하기 위한 방법론으로서 문화 특화 데이터 증강, 모델 미세조정, 프롬프트 엔지니어링 등의 접근법을 검토하고 그 효과를 검증하는 연구도 후속 과제로 진행할 예정이다.

감사의 글

본 연구는 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원 및 전남대학교 학술연구비(과제번호: 2025-0332-01)지원에 의하여 연구되었음

참고문헌

[1] Sung-Bin Kim, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh, “Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6430–6440, 2023.

[2] Burak Can Biner, Farrin Marouf Sofian, Umur Berkay Karakaş, Duygu Ceylan, Erkut Erdem, and Aykut Erdem, “SonicDiffusion: Audio-Driven Image Generation and Editing with Pretrained Diffusion Models,” arXiv preprint arXiv:2405.00878, 2024.

[3] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley, “AudioLDM: Text-to-Audio Generation with Latent Diffusion Models,” arXiv preprint arXiv:2301.12503, 2023.

[4] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao, “Make-An-Audio 2: Temporal-Enhanced

Text-to-Audio Generation," arXiv preprint arXiv:2305.18474, 2023.

[5] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "AudioGen: Textually Guided Audio Generation," arXiv preprint arXiv:2209.15352, 2022.

[6] Melissa Hall, Oscar Mañas, Reyhane Askari-Hemmat, Mark Ibrahim, Candace Ross, Pietro Astolfi, Tariq Berrada Ifriqi, Marton Havasi, Yohann Benchetrit, Karen Ullrich, Carolina Braga, Abhishek Charnalia, Maeve Ryan, Mike Rabbat, Michal Drozdzal, Jakob Verbeek, and Adriana Romero-Soriano, "EvalGIM: A Library for Evaluating Generative Image Models," arXiv preprint arXiv:2412.10604, 2024.

[7] Stability AI, "stable-audio-metrics: Evaluation Metrics for Audio Generation Models," GitHub repository, 2023. [Online]. Available: https://github.com/Stability-AI/stable-audio-metrics

[8] 한국지능정보사회진흥원 AI Hub, "이미지·사운드 매칭 데이터," AI Hub, 2025. Online: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&searchKeyword=이미지%20사운드%20매칭%20데이터&aihubDataSe=data&dataSetSn=71602, accessed Jan. 2026.

[9] R. Naik and B. Nushi, "Social Biases through the Text-to-Image Generation Lens," Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), pp. 1–23, 2023.