대체어 생성과 Zero-Shot TTS에 기반한 비속어 음성 필터링 시스템

신현서○, 박상근
경희대학교 소프트웨어융합학과
shs2677@khu.ac.kr, sk.park@khu.ac.kr

Profanity Speech Substitution System Based on Alternative Word Generation and Zero-Shot TTS

Hyeonseo Shin○, Sangkeun Park
Department of Software Convergence, Kyung Hee University

요 약

온라인 동영상 스트리밍 서비스가 대중화되면서, 사람들은 유튜브 등의 동영상 플랫폼에서 영상 시청에 많은 시간을 소비하고 있다. 이러한 동영상 플랫폼에서 수익과 조회수를 위한 선정적이고 폭력적인 콘텐츠가 증가하고 있다. 영상에서의 선정성 및 폭력성을 제거하기 위한 방법 중 하나로, 비속어 감지 및 필터링 관련한 다양한 연구가 수행되었다. 하지만 기존 연구는 주로 비속어를 묵음 처리하거나 효과음으로 대체하는 방식을 사용해서 본 영상의 문맥 손실을 야기하고 시청 경험을 저하시키는 한계가 있다. 본 논문에서는 온라인 동영상에 포함된 비속어 음성을 맥락에 맞는 자연스러운 대체 음성으로 합성하여 시청자의 몰입도를 유지하는 새로운 방법을 제안한다.

1. 서 론*

온라인 동영상 스트리밍 서비스가 대중화되면서, 사람들은 동영상 시청에 많은 시간을 사용하고 있다. 2023년 12월, 유명 온라인 동영상 서비스인 유튜브(YouTube)의 모바일 앱의 월간 활성 이용자(MAU)가 카카오톡을 제치고 국내 1위를 달성하기도 했으며, 2024년 1월 기준으로 1인당 월평균 유튜브 영상 시청 시간이 처음으로 월 40시간을 넘겼다[1]. '유튜버', '인플루언서'가 초등학생 희망 직업 상위권에 오를 정도로 동영상 플랫폼의 인기가 치솟고 있지만[2], 수익과 조회수를 위한 유튜브 콘텐츠의 선정성과 폭력성이 지속적으로 문제가 되고 있다[3]. 특히, 미디어 내의 무분별한 욕설 등의 비속어 사용은 시청자로 하여금 불쾌감을 줄 뿐만 아니라 청소년 및 아동의 언어 습관에도 치명적인 영향을 미칠 우려가 있다[4].

미디어 상의 비속어로 인한 불쾌감 및 악영향을 방지하기 위해, 비속어를 감지하고 적절히 처리하기 위한 다양한 연구들이 행되었다. 비속어가 감지되면 해당 음성을 다른 효과음으로 대체하거나 묵음으로 변환, 또는 다른 음성으로 대체하는 방법 등이 활용되었다[5, 6, 7]. 그러나 이러한 기존의 비속어 처리 방식은 영상 및 음성의 자연스러운 연결을 손실시킬 수 있어서 시청자의 시청 경험을 감소시키게 된다는 한계가 있다.

본 논문에서는 비속어를 미리 지정된 효과음으로 대체하거나 묵음처리하는 대신, 비속어가 포함된 문장의 맥락에서 해당 비속어를 대체할 수 있는 자연스러운 대체 음성을 찾아 합성하는 새로운 방법을 제안한다. 시청자는 비속어 음성 대신 최대한 영상에서의 화자 목소리와 비슷한 형태로 합성된 대체 텍스트 음성을 듣기 때문에 보다 나은 시청 경험을 할 수 있을 것으로 기대된다.

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2024년도 SW중심대학사업의 결과로 수행되었음 (2023-0-00042)

2. 관련연구

온라인 영상에서 비속어 사용을 감지하고 이를 필터링 및 차단하기 위한 연구가 꾸준히 진행되어 왔다. 영상과 음성을 포함하는 미디어에서의 비속어 처리 방식은 주로 비속어를 묵음화하거나 특수 효과음을 삽입하는 방식으로 구현되어 왔다. 한유림 et al.[5]은 비속어를 감지하면 문장 전체에 효과음이나 묵음 처리를 적용하였고, 김혜영 et al.[6]은 음성 데이터를 텍스트로 변환한 후 비속어를 감지하여 묵음 처리하는 방식으로 비속어 필터링을 구현하였다. Visutsak et al.[7]은 SVM을 활용하여 비속어를 탐지한 후, 묵음 처리를 통해 해당 비속어를 차단하는 방식을 구현했다. 이러한 기존의 비속어 처리 방식은 비속어 표현을 특정 데이터로 대체함으로써 시청자가 느낄 수 있는 불쾌감이나 악영향을 감소시킨다는 장점은 있지만, 비속어 대체 과정에서 문맥의 자연스러움이 손실되어 시청자의 집중도 및 영상 이해도에 부정적인 영향을 끼칠 수 있다는 한계가 있다[8].

대체어 생성과 관련된 다양한 텍스트 분석 연구가 존재한다. 지승현 & 이수원[9]은 부적절한 대체어를 효과적으로 필터링하기 위해 대체 토큰 감지 모델을 활용한 대체어 생성 방식을 제안했다. Arefyev et al.[10]은 타겟 단어와 문맥 정보를 결합한 대체어 생성 방식을 제안하였으며, 다양한 언어 모델을 비교해 높은 품질의 대체어를 생성했다. Zhou et al.[11]은 목표 단어를 부분적으로 마스킹하는 임베딩 드롭아웃 기법을 활용함으로써 단어의 의미와 문맥을 균형있게 고려하여 대체어를 생성하는 방법을 제안했다. 그러나 일반적인 단어가 아닌 비속어의 경우, 단순히 해당 단어의 의미가 중요하기보다는 감정을 강화시키거나, 감탄사로 사용되기도 하는 등, 다양한 상황에서 여러 의미로 활용되기 때문에 단어 자체의 의미를 명확히 규정하기 어렵다는 한계가 존재한다. 따라서 비속어의 대체어는 단어 자체의 의미보다 감정적 뉘앙스와 상황적 맥락이 중요하기 때문에, 의미 유사성보다는 문맥에 맞는 적절한 표현을

찾는 것이 더 효과적이다.

본 연구에서는 비속어 음성을 단순 유사어가 아닌 문맥에 자연스럽게 어울리는 단어의 음성으로 대체하여, 시청자가 불쾌감을 느끼지 않고 건전하게 동영상을 시청할 수 있는 시청 경험을 제공하고자 한다.

3. 시스템 설계 및 구현

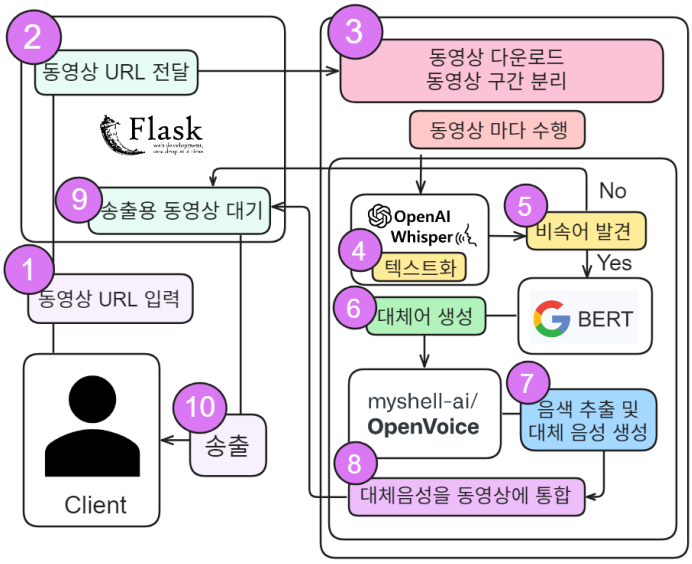

그림 1. 서비스 아키텍처

본 연구에서는 유튜브에서 동영상 시청시 영상에 존재하는 음성 비속어를 검출하고, 이를 문맥에 맞는 적절한 단어의 음성으로 대체하는 시스템을 개발한다. 사용자가 시청을 원하는 유튜브의 URL을 입력하면, Radford et al.[12]의 Whisper 모델을 활용해 유튜브 동영상의 음성을 텍스트로 변환한 후, 본 연구진이 사전 정의한 비속어 텍스트(예: 존나, 새끼, 시발 등)가 존재하는지 확인한다. 비속어가 검출되면 BERT 모델을 사용해 문맥에 맞는 적절한 대체어 텍스트를 찾는다. 이 대체어 텍스트는 OpenVoice v2[1]의 Zero-Shot TTS를 통해 해당 비속어를 말한 화자의 목소리를 흉내낸 음성으로 새롭게 합성되고, 해당 대체여 음성이 기존의 비속어 음성을 대체하게 된다. 이 과정이 끝나면 시청자는 비속어 음성이 제거되고 대체어 음성이 합성된 유튜브 영상을 시청할 수 있다. 시스템 아키텍처는 [그림 1]과 같다.

3.1 대체어 후보군 생성 및 평가

사용자가 비속어 음성 대체를 원하는 유튜브 영상의 URL을 입력하면, 해당 영상은 STT(Speech to Text) 모델인 Whisper를 통해 20초 단위마다 분할되어 텍스트로 변환된다 [표1]. 20초 단위로 분할된 각 텍스트마다 비속어가 있는지 확인하고, 비속어가 검출된다면 해당 비속어 자리를 [MASK]라는 토큰으로 대체한다. BERT 모델을 활용해 [MASK] 처리된 해당 비속어를 대체하기에 가장 적절한 대체어 후보 단어 5개를 생성한다. 원래 문장의 [MASK] 자리에 대체어 후보 단어 5개를 하나씩 삽입해보면서 문장을 재구성하고, 재구성된 문장마다 해당 후보 단어가 문맥에 적합할 확률을 계산한다 [표2]. 각 확률을 로그화하여 점수를 산정하고, 이 점수가 가장 높은 후보가 최종 대체어로 선택된다. [표1]은 20초 단위로 분할된 텍스트 예시이며, 이 중 "존나"라는 비속어가 포함되어있음을 확인할 수 있다. [표2]와 같이, "존나"를 대체할 대체어 후보 5개 중 "그렇게"가 문맥상 가장 적절한 단어임을 확인할 수 있다.

표 1. Whisper를 통한 Speech to Text 결과

타임스탬프	텍스트
00:00.000 --〉 00:02.600	그러면 네가 직접 가든가
00:02.600 --〉 00:03.300	그만해
00:03.300 --〉 00:05.900	야! 네가 제일 나빠
00:05.900 --〉 00:07.240	위하는 척은 지 혼자 다 하면서
00:07.240 --〉 00:09.140	존나 멀뚱멀뚱 가만히 있잖아
...	...
00:18.780 --〉 00:19.980	야, 이은유

표 2. 대체어 후보 평가 과정

대체어 후보	너는	나만	그냥	그렇게	옆에서
점수	7.724	-1.611	7.747	8.282	7.977

3.2 Zero-Shot TTS 기반 대체 음성 생성

비속어를 대체하기 위한 단어가 선택되면, 해당 단어의 음성을 생성하기 위해 Zero-Shot TTS 오픈소스인 OpenVoice v2를 활용했다. Zero-Shot TTS는 사전에 학습하지 않은 새로운 화자의 음성을 입력받아 별도의 추가적인 훈련과정 없이 음색과 억양을 모방하여 텍스트를 음성으로 변환하는 기술이다. OpenVoice v2는 짧은 시간 내에 적은 양의 데이터만으로도 고품질의 음색을 추출하고 적용할 수 있어, 짧은 시간 동안 반복적으로 대체 음성을 생성하는 본 연구에 적합한 모델이다.

비속어가 포함된 20초 구간의 음성을 OpenVoice v2에 입력하여 화자의 음색을 추출한다. 추출된 음색을 활용해 비속어를 대체할 대체어의 음성을 새롭게 생성한다. 생성된 대체 음성을 원본 음성의 타임스탬프와 일치하도록 합성하면 최종적으로 비속어 대신 대체어로 변환된 새로운 음성 파일이 완성된다. 완성된 음성 파일은 원본 비디오에 다시 합성되어 사용자에게 송출된다.

[1] https://github.com/myshell-ai/OpenVoice

3.3 서비스 개발

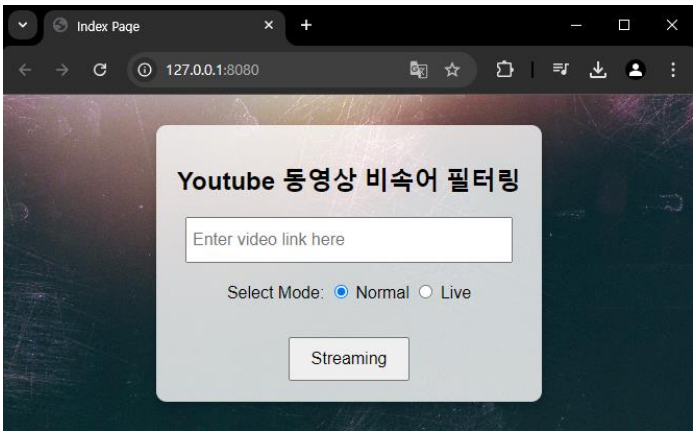

그림 2. 서비스 메인 화면

그림 3. 동영상 송출 화면 (〈스위트홈〉 예고편[2])

[그림 2]와 같이, 비속어 음성 대신 대체어 음성이 합성된 새로운 동영상을 시청하기 위한 서비스를 개발했다[3]. 사용자가 비속어없이 시청하고 싶은 유튜브 영상의 URL을 입력하고 Streaming 버튼을 클릭하면 [그림3]과 같이 비속어 음성이 제거되고 적절한 단어의 음성으로 대체된 영상을 시청할 수 있다.

4. 결론 및 향후 연구

본 연구에서는 유튜브 동영상에서 비속어 음성을 검출하고, 이를 문맥에 맞는 자연스러운 단어의 음성으로 대체해서 영상을 시청할 수 있는 서비스를 개발했다. 기존 연구들이 주로 비속어 검출 후 묵음 처리나 효과음 삽입에 초점을 두어 비속어의 감정을 억제하는 데 그쳤다면, 본 연구는 대체어 생성과 음성 합성을 통해 비속어를 보다 자연스러운 방식으로 대체했다는 장점이 있다. 특히, 비속어 대체어 선정 시 사전적 의미의 유사성보다는 문맥에 맞는 새로운 단어를 선택하고, 이를 화자의 음성을 기반으로 합성해서 시청자에게 자연스러운 시청 경험을 제공한다는데 그 의의가 있다. 본 연구는 BGM의 소실 및 대체음성 어투 유사도가 한계점으로 존재한다. 향후 연구로서는 BGM 분리 기술 및 고성능 Zero-Shot TTS 기술을 도입하여 더욱 발전한 대체음성을 생성할 수 있을 것으로 기대된다.

5. 참고문헌

[1] 계현우, “한국인 1인당 유튜브 월평균 40시간 사용 돌파”, KBS뉴스, http://surl.li/iobpgg, 2024.03.04.
[2] 최은경, “초등생 장래희망 직업서 의사 2위로 껑충… 그럼 1위는?”, 조선일보, http://surl.li/ezggay, 2023.11.27.
[3] 김재희, “무단침입에 자작극까지… 도 넘은 ‘유튜브 콘텐츠 경쟁’”, 동아일보, http://surl.li/fjfwbt, 2020.09.02.
[4] 곽희양, “아무 영상이나 보고 욕 배운 ‘AI 어린이’”, 경향신문, http://surl.li/zqgnuz, 2020.06.12.
[5] 한유림, 이소아, 장현지, 조성연, 김명주. “딥러닝 기반 영상 속 유해언어 차단 시스템 개발,” 한국컴퓨터교육학회 학술발표대회논문집, 26, 2, 157-160, 2022.
[6] 김혜영, 이영우, 서지헌, 박성민, 김현우, 박영진. “양방향 LSTM 기반 한국어 음성 비속어 필터링”, 멀티미디어학회논문지, 27, 1, 126-133, 2024.
[7] Porawat Visutsak, Apiwut Wijitemee, Akaphon Mahaphon, Orawan Chaowalit. “Machine-Learning-Based Profanity Detection and Removal in Cartoons Videos”, International Conference on Control and Robotics (ICCR), 2023.
[8] Francielle Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, Thiago Pardo. “Contextual-Lexicon Approach for Abusive Language Detection”, Proceedings of the International Conference on Recent Advances in Natural Language Processing, 1438-1447, 2021.
[9] 지승현, 이수원. “대체 토큰 감지 모델을 통한 대체어 추출”, 정보과학회논문지, 50, 4, 321-328, 2023
[10] Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, Alexander Panchenko. “Always Keep your Target in Mind: Studying Semantics and Improving Performance of Neural Lexical Substitution”, Proceedings of International Conference on Computational Linguistics, 1242 - 1255, 2020.
[11] Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. "BERT-based Lexical Substitution", Proceedings of the Annual Meeting of the Association for Computational Linguistics, 3368 - 3373, 2019.
[12] Alec Radford, Jong Wook Kim, Tao Xu1 Greg Brockman, Christine McLeavey, Ilya Sutskever. “Robust Speech Recognition via Large-Scale Weak Supervision”, International Conference on Machine Learning, 28492-28518, 2023.

[2] https://www.youtube.com/watch?v=B5IQqZDSRjk
[3] https://youtu.be/D_N0HM708a8