효과음 자막 생성을 위한 딥러닝 기반의 다중 사운드 분류

정현영*, 김규미*, 김현희*
*동덕여자대학교 정보통계학과
gusdud4573@gmail.com, kyume9@gmail.com, heekim@dongduk.ac.kr

A Multiclass Sound Classification Model based on Deep Learning for Subtitles Production of Sound Effect

Hyeonyoung Jung*, Gyumi Kim*, Hyon Hee Kim*
*Dept. of Statistics and Information Science, Dongduk Women's University

요 약

본 논문은 영화에 나오는 효과음을 자막으로 생성해주는 자동자막생성을 제안하며, 그의 첫 단계로써 다중 사운드 분류 모델을 제안하였다. 고양이, 강아지, 사람의 음성을 분류하기 위해 사운드 데이터의 특징벡터를 추출한 뒤, 4가지의 기계학습에 적용한 결과 최적모델로 딥러닝이 선정되었다. 전처리 과정 중 주성분 분석의 유무에 따라 정확도는 81.3%와 33.3%로 확연한 차이가 있었으며, 이는 복잡한 특징을 가지는 사운드를 분류하는데 있어 주성분 분석과 넓고 깊은 형태의 신경망이 보다 개선된 분류성과를 가져온 것으로 생각된다.

1. 서론

배리어프리버전 영화(이하 배리어프리영화)란, 기존의 영화에 화면을 음성으로 설명해주는 해설과 화자 및 대사, 음악, 소리정보를 알려주는 자막을 넣어 모든 사람이 함께 즐길 수 있도록 만든 영화이다. 기존의 영화에서는 배우의 대사를 자막으로 생성하지만 효과음, 음악 소리, 동물 소리와 같은 대사 이외의 다양한 소리는 자막으로 제공되지 않는다. 따라서 배리어프리 영화가 일반화되면 청각 장애인도 보다 풍부한 소리를 자막으로 서비스 받을 수 있게 될 것이다.

현재의 자막 생성기술인 음성 인식(Speech To Text, STT) 기술은 대사만을 자막으로 생성해 낸다는 점에서, 대사 이외의 음향효과와 같은 소리 정보를 알리는 자막이 필요한 배리어프리영화에 적용하기엔 부족한 점이 있다. 따라서 본 논문은 화면에 나타나지 않은 소리정보도 자막으로 나타내는 사운드 기반의 자막 생성을 위한 다중 사운드 분류 모델을 제안하였다.

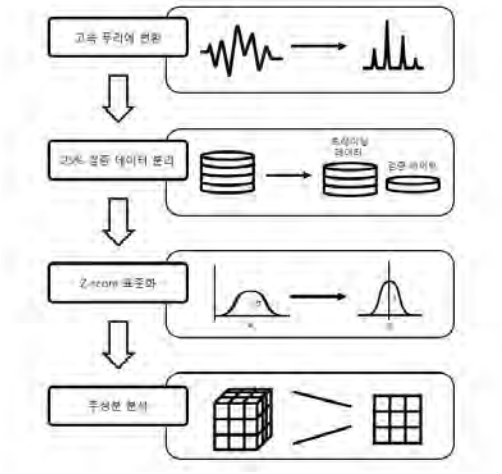
본 연구에선 강아지, 고양이, 사람의 사운드 데이터를 수집 후 고속 푸리에 변환(Fast Fourier Transform)을 적용하고 주성분 분석(PCA)을 통해 사운드의 특징을 추출하였다. 다음으로 다중 클래스 사운드 분류에 적절한 기계 학습 모델을 찾아내기 위해서, 가우시안 나이브 베이즈, 로지스틱 회귀, 랜덤 포레스트, 딥러닝까지 총 4가지 기계학습 모델을 적용하였다. 네 가지 모델의 분류 정확도를 비교한 결과, 주성분 분석을 적용한 딥러닝 모델이 81.3%의 정확도로 가장 높은 성능을 내는 것을 알 수 있었다.

제안한 모델을 활용하여 분류된 효과음을 자막으로 생성하면, 동물의 소리도 자막으로 볼 수 있어 청각 장애인들을 위한 자막 서비스가 보다 생동감있게 제공될 수 있을 것으로 기대된다.

2. 데이터 전처리

[그림 1]은 데이터 전처리 과정을 나타낸 것이다. 고양이 167개, 강아지 112개, 사람 100개의 사운드 데이터를 트레이닝 데이터로 수집했으며, 테스트 데이터로는 각 클래스별로 50개씩의 사운드 데이터를 「내 어깨 위 고양이, 밥」, 「화이트 갓」 등 다양한 영화로부터 추출하여 활용했다. 이때, 사람의 사운드 데이터는 여성 50%, 남성 50%로 동일하게 수집하였으며, 연령별 사운드의 차이를 고려하여 각

아이 15%, 성인 20%, 노인 15%로 연령대를 고르게 수집하였다.



[그림 1] 데이터 전처리 프로세스

다음으로 고속 푸리에 변환을 사용하여 데이터로부터 시간적 흐름의 소리 정보를 주파수의 흐름으로 변환하였다. 고속 푸리에 변환 기법[1]을 이용하면 임의의 신호를 수학적 변수로 변환할 수 있기 때문에 현재 음성분석, 지진파 분석 등 신호 분석에서 널리 사용되고 있다.

```
array([[ 15.30645979,  26.27110801,  30.61643619, ..., -54.33958484,
        -54.34032627, -54.33370308],
       [  1.94797151,  29.3065174 ,  32.65448653, ..., -61.55720458,
        -61.59933568, -61.58510577],
       [  8.62795442,  22.80452313,  22.9281394 , ..., -67.47438591,
        -67.46644165, -67.48636664],
       ...,
       [ -1.01850621,   1.22566043,  -5.30150667, ..., -43.20566552,
        -38.67296109, -42.82643031],
       [ 28.73957227,  35.47679534,  36.72878961, ..., -44.69949798,
        -44.70262766, -44.7064108 ],
       [ -8.40316883, -21.84247719,  -8.28662848, ..., -56.40225397,
        -56.3708293 , -56.36798461]])
```

[그림 2] 고속 푸리에 변환한 형태의 데이터프레임

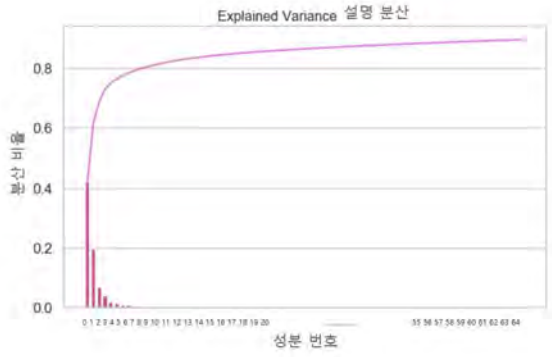
[그림 2]는 고속 푸리에 변환을 통해 계산된 사운드 데이터이다. 고속 푸리에 변환을 통해 사운드 데이터를 샘플링 된 특징값으로 추출하여 기계학습에 적용하였다.

[표 1] 데이터 분리 표

	트레이닝	검증	테스트
고양이	125	41	50
강아지	83	29	50
사람	75	25	50

[표 1]은 고속 푸리에 변환된 트레이닝 데이터의 각 분류별 25%를 검증 데이터로 분리해 준 전체 데이터의 개수이다. 테스트 데이터에 적용하기 전, 모델을 검증하기 위한 검증용 데이터와 트레이닝 데이터는 Z-score 표준화와 주성분 분석을 각각(따로) 진행하였다.

데이터를 효과적으로 분류하기 위하여 Z-score 표준화(Standard Scaler)한 후 주성분 분석(Principal Component Analysis)을 통해 차원을 축소하였다.



[그림 3] 트레이닝 데이터의 설명 분산

[그림 3]은 각각의 주성분 벡터가 이루는 축에 투영한 결과의 분산의 비율과 누적비율을 나타낸다. 이를 통해 방향벡터가 큰 6개의 성분(component)을 주성분으로 선택하여 입력 데이터로 사용하였다.

3. 성능비교 및 최적모델 선정

[표 2] 모델 별 정확도 비교

모델	검증 정확도	테스트 정확도
가우시안 나이브 베이즈	0.874	0.513
랜덤 포레스트	0.874	0.593
로지스틱 회귀	0.916	0.773
딥러닝	0.884	0.813

[표 2]는 가우시안 나이브 베이즈, 랜덤 포레스트, 로지스틱 회귀, 딥러닝 총 4가지 모델에 트레이닝 데이터와 테스트 데이터를 넣은 후 정확도를 보여준다. 로지스틱 회귀 모델에서 검증 정확도가 91%를 넘겨 가장 최적모델로 보이는 듯 했으나, 테스트 데이터에 적용하였을 때는 77%로 나타났다.

반면 검증 정확도에서 두 번째로 높은 성능을 보인 딥러닝 모델의 경우, 테스트 데이터에 적용하였을 때 81%의 정확도가 나타남으로써 이를 최적모델로 선정하였다.

4. 제안한 딥러닝 모델

[표 3] 제안한 딥러닝 모델의 분류별 인식률

정확도 점수 :
Accuracy Score:
0.813

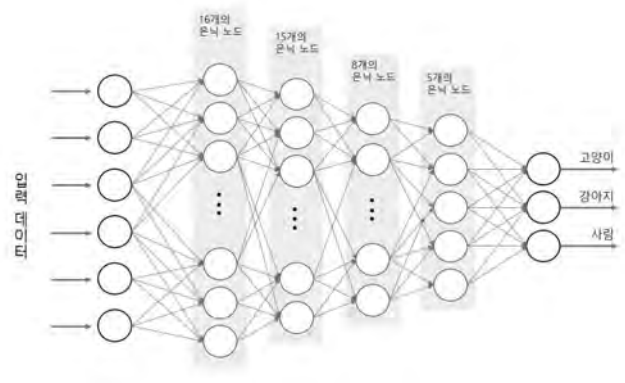
	정밀도 precision	재현율 recall	F1 점수 f1-score	support
고양이 cat	0.70	0.80	0.75	50
강아지 dog	0.87	0.78	0.82	50
사람 people	0.90	0.86	0.88	50
정확도 accuracy			0.81	150
macro avg	0.82	0.81	0.82	150
weighted avg	0.82	0.81	0.82	150

[표 3]는 테스트 데이터를 최적모델에 적용한 결과를 클래스별로 정밀도(precision), 재현율(recall), F1-score, 그리고 지지도(support)를 나타낸 표이다.

강아지와 사람은 각 82%와 88%로 높은 F1-score를 보였지만, 고양이는 75%로 비교적 낮은 F1-score를 보였다.

또한 추가적인 연구에 고양이와 강아지 소리가 동시에 나오는 테스트 데이터를 넣어 주었을 때, 강아지로 인식하는 것으로 보아, 강아지의 소리가 인식이 더 잘 되는 것을 알 수 있었다.

사람의 사운드 데이터의 경우, 트레이닝 데이터의 언어가 전부 영어로 이루어져 있으며, 테스트 데이터의 언어도 영어로 이루어져 있어 높은 F1-score를 보였다. 하지만 다른 언어가 테스트 데이터로 들어왔을 때는 낮은 F1-score를 보였다. 이로 인해 사람의 사운드 와 다른 사운드를 분류를 할 때 다양한 언어로 학습하는 것이 중요한 요소임을 알 수 있었다.



[그림 4] 제안하는 딥러닝 모델의 구조

[그림 4]는 다중 사운드 분류를 위한 5층으로 구성된 완전연결구조의 딥러닝 모델의 학습과정을 나타낸다. 고속 푸리에 변환(FFT) 함수로 계산하여 주성분 분석을 적용한 데이터는 6개의 입력 노드를 거친다. 은닉층의 노드가 16, 15, 8, 5로 구성된 본 모델은 활성화 함수로 'relu'를 사용하고 있으며, 가중치 최적화를 위한 함수는 'adam'을 적용하였고, 고양이, 강아지, 사람을 분류하기 위해 3개의 출력 노드를 가지고 있다. 최대 학습 횟수는 200회로 설정하였으며, L2 규제를 위한 매개 변수인 알파(alpha)값은 0.015로 설정해주었다.

[표 4] 주성분 분석 적용에 따른 딥러닝 모델 성능비교

딥러닝 모델	검증 정확도	테스트 정확도
주성분 분석을 적용한 딥러닝 모델	0.884	0.813
주성분 분석을 적용하지 않은 딥러닝 모델	0.432	0.333

주성분 분석을 적용하지 않은 딥러닝 모델과 성능을 비교하였을 때, 전처리 단계에서 주성분 분석을 거친 모델과 확연한 성능차이가 있었다. 이는 고속 푸리에 변환 된 수많은 값을 동일한 가중치로 넣어주는 것 보다 특징벡터를 분석한 후 입력 데이터로 넣어 주는 것이 더 효과적으로 보여 진다.

5. 다중 사운드 분류의 응용

본 연구에서 제안한 다중 사운드 분류 모델을 자막생성에 활용한다면 효과음도 자막으로 생성해 낼 수 있다. 또한 본 연구의 모델은 고양이, 강아지, 사람으로 이루어진 3개의 클래스로만 분류를 해냈지만, 향후 데이터의 확대를 통해 더 다양하고 세밀한 효과음까지 자막으로 표현할 수 있으리라 기대한다.

[그림 5] 다중 사운드 분류의 응용 예시

[그림 5]는 다중 사운드 분류모델을 영화에 적용하였을 때의 상황을 나타낸 것이다. 대사만을 자막

으로 나타내는 것이 아닌, 고양이나 강아지 소리도 자막으로 나타내면서 청각장애인의 영화 이해도를 높일 수 있으며, 풍성한 상황해설을 할 수 있다.

6. 결론 및 기대효과

본 논문은 대사 뿐 아니라 효과음도 자막으로 나타낼 수 있는 사운드 기반 자동 자막 생성을 제안하며, 이의 첫 단계로 다양한 사운드를 분류해 낼 수 있는 다중 사운드 분류 모델을 연구하였다.

사운드 데이터의 경우 복합한 특징벡터를 가지고 있기 때문에 단순한 기계학습보다 딥러닝 모델에서 더 좋은 분류 결과를 보였으며, 주성분 분석 과정을 통하여 분류에 큰 영향을 미치는 성분을 추출하였다. 이를 딥러닝 모델에 적용했을 때, 주성분 분석을 적용하지 않은 모델에 비해 월등이 좋은 성능을 나타냈다는 점에서, 사운드 데이터를 분류하는데 특징벡터의 분석과정이 큰 의미가 있음을 알 수 있었다.

본 연구의 다중 사운드 분류 모델은 81% 정확도라는 결과를 냈으나, 영어가 아닌 언어의 데이터를 넣었을 때 성능이 크게 떨어졌다는 것을 고려하면 데이터셋의 확대가 이루어졌을 때 더 좋은 성능을 낼 수 있을 것이라 기대된다. 또한 향후 자동 자막 생성 기술와 접목된다면 대사 뿐 아니라 화면에 나타나지 않는 사운드까지 자막으로 나타낼 수 있다는 점에서 배리어프리영화에 적용하여 원활한 영화 공급 및 취약계층이 문화적 권리를 향유하는데 큰 도움이 될 것이라 기대된다.

참고문헌

[1] 김형석, 김인태 “고속푸리에변환을 이용한 부식강재의 응력집중계수 산출”, 대한토목학회 정기학술대회, 2018, pp.569-570

[2] 최재승 “남녀성별 분류를 위한 화자종속 음성인식 알고리즘”, 한국정보통신학회논문지, Vol.17, No.4, pp.775-780, 2013

[3] 박대서, 방준일, 김화종, 고영준 “CNN을 이용한 음성 데이터 성별 및 연령 분류 기술 연구”, 한국정보기술학회논문지, Vol.16, No.11, pp.11-21, 2018

[4] 김지은, 이인성 “MFCC를 이용한 GMM 기반의 음성/혼합 신호 분류”, 전자공학회논문지, Vol.50 No.2, pp.185-192, 2013

[5] 금지수, 임성길, 이현수 “스펙트럼 분석과 신경망을 이용한 음성/음악 분류”, 한국음향학회지, Vol.26, No.5, pp.207-213, 2007

[6] 정명범, 고일주 “오디오의 파형과 FFT 분석을 이용한 대표 선율 검색”, 정보과학회논문지 : 소프트웨어 및 응용, Vol.34, No.12, pp.1037-1044, 2007