미디어에서의 오디오 메타데이터 최적화 추출 및 분류 방안에 대한 연구

윤민희* · 박효경 · 문일영

한국기술교육대학교

A Research of Optimized Metadata Extraction and Classification of in Audio

Min-hee Yoon* · Hyo-gyeong Park · Il-Young Moon

Korea University of Technology and Education

E-mail : oliver93@koreatech.ac.kr / sjjshk@koreatech.ac.kr / iymoon@koreatech.ac.kr

요 약

최근 미디어의 시장의 급격한 성장과 그에 따른 사용자들의 기대감이 증가하고 있다. 이 연구에서는 미디어에서 추출한 오디오를 통하여 다양한 태그를 추출하고 인공지능을 활용하여 특정 카테고리로 분류한다. 이 카테고리는 감정에 대한 종류이며 기쁨, 분노, 슬픔, 즐거움, 사랑, 증오, 욕망 등이 있을 수 있다. 해당 연구를 수행하기 위하여 Jupyter Notebook 프로그램을 사용하며, Jupyter Notebook 내에서 LiBROSA 라이브러리를 이용하여 음성데이터를 분석하고 Keras와 계층 모델을 이용하여 Neural Network를 학습한다.

ABSTRACT

Recently, the rapid growth of the media market and the expectations of users have been increasing. In this research, tags are extracted through media-derived audio and classified into specific categories using artificial intelligence. This category is a type of emotion including joy, anger, sadness, love, hatred, desire, etc. We use JupyterNotebook to conduct the corresponding study, analyze voice data using the LiBROSA library within JupyterNotebook, and use Neural Network using keras and layer models.

키워드

오디오 추출, 음성데이터, 메타데이터, AI

Ⅰ. 서 론

정보통신 기술의 발달로 영상 콘텐츠의 수가 기하급수적으로 증가하면서 빅데이터화 되고 있다. 영상 데이터는 비정형 데이터로써, 두 가지 다른 콘텐츠인 오디오와 이미지로 구성될 뿐 아니라, “흐르는” 콘텐츠 유형이기에 메타데이터를 쉽게 식별하기 어렵다는 특징이 있다. 급증하고 있는 영상 데이터는 빅데이터의 특징으로 3V라고 불리는 다양성(Variety), 데이터의 양(Volume), 데이터 생성 속도(Velocity)에 부합한다고 볼 수 있다[1]. 다양한 종류의 데이터가 융복합 환경에서 매우 빠른 속도로 생산되고 있고, 이에 따라 대용량의 데이터를 빠르게 처리하고 분석해야 한다는 것이다. 이는 현재의 영상 콘텐츠의 증가 추세와도 관련이 있다. 기존 영상 콘텐츠는 사람이 수작업으로 메타데이터를 입력하여 추천 알고리즘을 생성하였다. 본 연구에서는 자동으로 태깅을 하여 음성데이터를 분류하는 과정을 설명한다.

* corresponding author

II. 환경설정

2.1 환경설정

연구를 위하여 Backend.AI의 환경을 세팅하여 진행하였다. Backend.AI는 클라우드 리소스 관리 플랫폼이다. 가상 플랫폼인 Backend.AI는 개발자, DevOps 엔지니어, 엔터프라이즈 등 다양한 사용자들을 위하여 개발됐다, Backend.AI의 부분적인 GPU 자원은 요구에 따라 효율적으로 확장할 수 있도록 지원한다. 연구는 Python 3.6버전, Tensorflow 1.15 버전, CPU 8코어, RAM 61GB를 사용하여 진행되었다.

2.2 음성 특징 추출

음성이 가지고 있는 특징의 추출은 해당 음성을 분석하고 찾는 데 매우 중요한 부분이다. 분석하지 못하는 형식의 특징 추출을 직접 변환시키는 모델을 사용함으로써 데이터를 사용한다. 음성의 분류, 예측 및 권장 알고리즘에는 특징 추출이 필요하다.

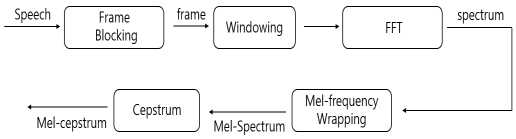

그림 1. MFCC를 활용한 특징 추출 알고리즘

III. 사용 기술

3.1 **MFCC**

Spectral 특징 추출에 가장 널리 사용되고 있는 방법은 MFCC를 이용하는 것이다. MFCC(Mel-Frequency Cepstral Coefficient)란 MFC를 집합적으로 구성하는 계수이다. 즉, MFCC는 신호의 고속 푸리에 변환으로부터 파생된 짧은 시간의 신호를 Window를 통하여 주파수로 바꿀 때의 계수를 뜻한다. 일반적인 Cepstrum과 멜 주파수 Cepstrum의 차이는 MFCC에서 주파수 대역이 멜 스케일에서 같은 간격으로 존재한다는 것이며, 이는 일반적인 Cepstrum에서 사용되는 선형 간격 주파수 대역보다 인간 청각 시스템의 반응을 더 가깝게 근사하게 한다[2].

3.2 **SVM(Support Vector Machine)**

SVM은 기계 학습의 분야 중 하나로 패턴 인식, 자료 분석을 위한 지도 학습 모델이며, 주로 분류와 회귀 분석을 위해 사용한다. 두 카테고리 중 어느 하나에 속한 데이터의 집합이 주어졌을 때, SVM 알고리즘은 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진 선형 분류 모델을 만든다. 만들어진 분류 모델은 데이터가 사상된 공간에서 경계로 표현되는데 SVM 알고리즘은 그중 가장 큰 폭을 가진 경계를 찾는 알고리즘이다. SVM은 선형 분류와 더불어 비선형 분류에서도 사용될 수 있다[3].

3.3 **Zero-Crossing Rate**

ZCR(Zero-Crossing Rate)은 신호가 양의 값에서 0으로, 음의 값에서 양의 값으로, 또는 음의 값에서 양의 값으로 변하는 속도이다. 이 값은 음성 인식과 음악 정보 검색 모두에서 널리 사용되어 왔다[4].

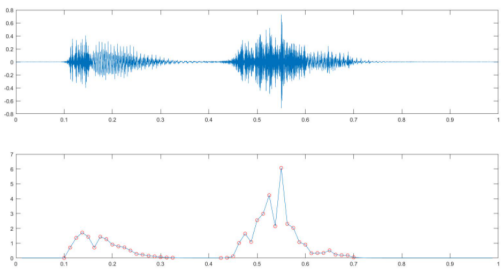

그림 2. ZCR을 통한 음성 신호 처리

IV. 실험 결과

4.1 **Python에서의 LibROSA 라이브러리**

Python의 표준 라이브러리는 매우 광범위하며 다양한 기능을 제공한다. 라이브러리에는 일상적인 프로그래밍에서 발생하는 많은 문제에 대한 표준적인 해결책을 제공하는 Python으로 작성된 모듈이 포함된다. 이 모듈 중 일부는 플랫폼 관련 사항을 플랫폼 중립적인 API들로 추상화시킴으로써, Python 프로그램의 이식성을 권장하고 개선하도록 명시적으로 설계되었다[5].

LibROSA는 Python의 수많은 라이브러리 중 하나이다. 사용자들은 LibROSA를 통하여 음악과 오디오 분석을 할 수 있다. 해당 라이브러리를 통하여 음악 정보 검색 시스템을 만드는 데 필요한 구성요소를 받을 수 있다. 연구 진행에 있어서 LibROSA는 음성 분석의 중요한 역할을 하였다.

4.2 **Keras**

Keras란 Python으로 작성된 고수준 신경망 API로 TensorFlow, CNTK, 혹은 Theano와 함께 사용할 수

있으며 빠른 실험에 중점을 두고 있다.

Keras를 사용하면 사용자 이용성, 모듈성, 확장성을 통해 빠르고 간편한 프로토타이핑을 할 수 있다. 또한, 컨볼루션 신경망, 순환 신경망, 그리고 둘의 조합까지 모두 지원된다. CPU뿐 아니라 GPU에서도 매끄럽게 실행된다[6].

표 1. Classification Report

	Precision	recall	F1 score	support
0	0.86	0.87	0.86	180
1	0.69	0.73	0.71	124
2	0.84	0.77	0.80	267
3	0.84	0.78	0.81	266
4	0.84	0.90	0.87	249
5	0.80	0.80	0.80	268
6	0.82	0.87	0.85	189
7	0.85	0.86	0.86	191
accuracy			0.82	1734

위의 [표 1]은 각 특징을 분류한 후의 분류성과지표를 나타낸 것이다. Precision(정밀도), Recall(재현율), F1 Score, Accuracy(정확도)를 각 분류마다 나타낸 것이다. 0부터 7까지는 감정을 숫자로 치환한 것이다. 치환한 값은 다음과 같다. {중립:0, 차분함:1, 행복함:2, 슬픔:3, 화남:4, 공포:5, 혐오감:6, 놀라움:7}

각 감정 특징 중 가장 높은 F1 Score가 나타난 것은 '4번, 화남'이었으며, 가장 낮은 값은 '1번, 차분함'이었다.

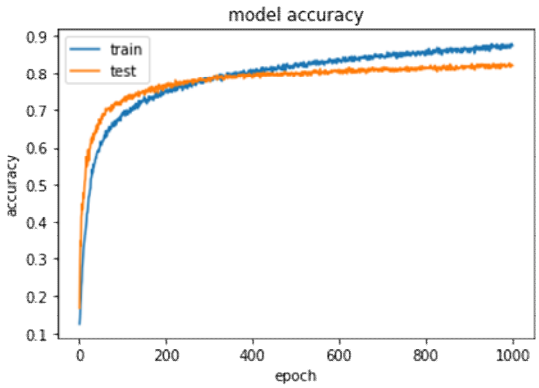

그림 3. Model Accuracy

[그림 3]은 모델의 Accuracy에 대한 그래프이다. 모델에 대한 Accuracy 값은 82.24%로 나타났다.

실험을 진행함에 따라 위의 식에서 epsilon 값을 1e-07로 대입하여 진행하였다.

opt = keras.optimizers.rmsprop(lr=0.00001, rho=0.9, epsilon=1e-07, decay=0.0)

이전의 다른 연구에서 epsilon 값은 기존 논문 코드에서 None으로 되어 있었다. 그 결과 default 값으로 K.epsilon() 함수가 호출되었던 것이다. 따라서, 본 실험을 진행하며 epsilon 값을 임의로 조정할 때마다 결과값과 차트의 모양이 다르게 도출되었다.

Ⅴ. 결 론

실험을 통하여 아날로그 신호인 소리를 디지털 신호로 전환되어 특징을 추출하고 분류하였다. 이는 사용자의 기분 파악, 미디어 추천 등 다양한 분야에 응용할 수 있을 것이다. 본 실험에서 쓰인 데이터보다 더 많은 양의 데이터를 수집하여 학습시킨다면, 더 높은 정확도 기대해볼 수 있을 것이다. 특징 추출에 있어서 정확도와 태깅의 세부화를 거친다면 미디어 이용자들에게 더욱 정밀한 추천 서비스를 제공할 수 있을 것이다.

Acknowledgement

이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업 (No. 2018R1D1A3B07049722) 과제 지원에 의하여 연구되었음.

References

[1] C.Nalini, A.R.Arunachalam, "A STUDY ON PRIVACY PRESERVING TECHNIQUES IN BIG DATA ANALYTICS" *International Journal of Pure and Applied Mathematics*, Vol. 116 No. 10, 2017.

[2] K.S.R. Murty, B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition", *IEEE* SIGNAL PROCESSING LETTERS, VOL. 13, NO. 1, JANUARY 2006.

[3] Derek A.PisnerDavid M.Schnyer, *Machine Learning,* Academic Press, pp 101-1211, 2020.

[4] Bachu R.G., Kopparthi S., Adapa B., Barkana B.D. "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal ", *asee*, 2008.

[5] https://librosa.org/doc/latest/index.html

[6] https://keras.io/ko/