The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.10, pp.492~496

eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.10.492

A Deep Learning System for Emotional Cat Sound Classification and Generation

Joo Yong Shim† · SungKi Lim†† · Jong-Kook Kim†††

ABSTRACT

Cats are known to express their emotions through a variety of vocalizations during interactions. These sounds reflect their emotional states, making the understanding and interpretation of these sounds crucial for more effective communication. Recent advancements in artificial intelligence has introduced research related to emotion recognition, particularly focusing on the analysis of voice data using deep learning models. Building on this background, the study aims to develop a deep learning system that classifies and generates cat sounds based on their emotional content. The classification model is trained to accurately categorize cat vocalizations by emotion. The sound generation model, which uses deep learning based models such as SampleRNN, is designed to produce cat sounds that reflect specific emotional states. The study finally proposes an integrated system that takes recorded cat vocalizations, classify them by emotion, and generate cat sounds based on user requirements.

Keywords : Audio Classification, Audio Generation, Animal Emotion Recognition, SampleRNN, Deep Learning System

감정별 고양이 소리 분류 및 생성 딥러닝 시스템

심 주 용† · 임 성 기†† · 김 종 국†††

요 약

반려동물, 특히 고양이는 인간과의 상호작용에서 다양한 소리를 통해 감정을 표현하는 것으로 알려져 있다. 고양이의 소리는 그들이 느끼는 감정 상태를 반영하며, 이를 이해하고 해석하는 것은 반려동물과의 소통을 더욱 원활하게 하는 데 중요한 요소이다. 최근 인공지능 기술의 발전으로 감정 인식과 관련된 연구가 활발히 진행되고 있으며, 특히 딥러닝 모델을 활용한 음성 데이터 분석이 주목받고 있다. 본 연구는 이러한 배경에서 출발하여, 고양이의 소리를 감정별로 분류하고 생성하는 딥러닝 시스템을 개발하는 것을 목표로 한다. 분류 모델은 고양이 소리를 감정별로 정확하게 분류하기 위해 학습되며, 소리 생성 모델은 SampleRNN과 같은 딥러닝 기법을 활용하여 특정 감정을 표현하는 고양이 소리를 생성할 수 있도록 설계된다. 마지막으로, 학습된 두 모델을 통합하여 고양이 소리를 녹음하고 이를 감정별로 분류한 결과 및 사용자의 요구에 따른 고양이 소리를 생성하여 제공할 수 있는 시스템을 제안한다.

키워드 : 소리 분류, 소리 생성, 동물 감정 인식, SampleRNN, 딥러닝 시스템

1. 서 론

반려동물과의 커뮤니케이션 수단으로서 소리는 매우 중요한 역할을 한다. 개나 고양이 등 반려동물은 다양한 방식으로 의사소통을 시도하지만, 그 중에서도 소리를 통한 커뮤니케이션은 가장 핵심적인 방법 중 하나이다[1]. 그러나 반려동물이 내는 소리를 인간이 정확하게 이해하는 것은 어려우며, 반려동물의 감정을 잘못 해석할 경우, 반려동물의 건강과 행동에 부정적인 영향을 미칠 수 있다. 따라서 반려동물이 내는 소리를 분석하고 이를 기반으로 감정을 예측하며 함께 소통할 수 있는 기술이 필요하다. 이러한 기술을 통해 반려동물의 감정을 보다 정확하게 이해할 수 있으며, 반려동물의 행동 문제를 예방하거나 해결하는 데 도움이 될 수 있다.

본 논문에서는 딥러닝 기반 반려묘의 감정별 소리 분석 및 생성 시스템을 제안한다. 기존의 딥러닝 기반 소리 분석 및 합성 기술들은 대부분 인간의 감정 혹은 환경, 악기 소리 등에 초점을 맞추어 진행되어 왔다[2-5]. 그러나, 동물 소리, 특히 반려묘에 대한 연구는 매우 부족하며, 기존 반려동물에 관한 시스템에서는 감정을 정확하게 분류해 내지 못하거나 소통을 위해 녹음된 소리를 사용하는 방식이 대부분으로 다양한 소리를 표현하기가 어려웠다[6, 7]. 이러한 한계를 극복하기 위해,

† 준 회 원 : 고려대학교 정보통신기술연구소 박사후연구원
†† 비 회 원 : ㈜애니멀보이스 대표이사
††† 종신회원 : 고려대학교 전기전자공학과 교수
Manuscript Received : August 29, 2024
Accepted : September 6, 2024
* Corresponding Author : Jong-Kook Kim(jongkook@korea.ac.kr)

딥러닝을 이용해 소리를 분석하고 분류하여 복잡한 패턴 인식과 학습을 통해 반려동물의 다양한 감정과 의사를 정확하게 구분해 내며, 또한 반려묘가 상황과 감정에 따라 내는 유사한 소리를 생성할 수 있는 시스템을 제안한다. 이를 통해 반려동물과 인간 사이의 커뮤니케이션을 개선하고, 보다 나은 상호작용을 위한 기술적 기반을 마련하고자 한다.

세부적으로, 본 논문에서는 고양이의 감정과 의사를 나타내는 소리 데이터를 학습시켜 고양이의 소리를 분석하고 유사한 소리를 생성하여 소통할 수 있는 시스템을 개발하는 연구를 총 3가지 단계에 걸쳐 수행한다. 먼저, 첫 번째 단계에서는 소리 파일 분석 및 분류를 위한 AI 모델 개발에 집중한다. 반려동물의 감정별 소리 파형을 추출하고, 소리의 특징을 잘 살릴 수 있도록 STFT(Short-Time Fourier Transform)와 Mel-spectrogram 등의 방법을 통해 소리를 시각적으로 표현한다. 또한, 소리 특징을 학습하여 감정별 특징 추출(feature extraction)을 진행하는 동시에 분류(classification)를 통하여 감정 및 의사별로 분류할 수 있도록 한다. 두 번째 단계에서는 소리 합성 및 생성을 위한 모델 개발을 목표로 한다. 샘플 단위로 소리를 생성해 내는 SampleRNN[6]을 활용하여 소리 생성 모델을 구축하고, 고양이 소리에 적합한 데이터 전처리와 튜닝을 통해 감정별 소리 생성 학습을 진행하며, 최종적으로 생성된 소리를 평가한다. 마지막으로, 세 번째 단계에서는 모든 기능을 통합한 시스템을 구축한다. 이 과정에서는 실제 고양이 소리를 녹음하여 인식하고, 소리를 감정별로 분류하고 생성된 소리를 적절하게 선택하여 결과를 반환할 수 있도록 모델들을 통합한다.

2. 본 론

2.1 감정별 소리 분석 및 종류

감정별 소리 분석 및 분류 과정을 진행하기 위해, Fig. 1과 같이 모델을 구성하였다. 사용된 모델은 소리 데이터의 파형(waveform)을 입력으로 받아, 시간적 및 주파수적 특징을 잘 반영할 수 있는 Mel-spectrogram 표현으로 변환한다. 이렇게 처리된 소리 데이터를 기반으로, 고양이 소리를 카테고리별로 효과적으로 구분 할 수 있는 단순한 여러 개의 CNN 레이어로 구성된 Feature Extraction 모델을 설계하여 Classification을 하도록 학습하여 소리를 감정 별로 분류할 수 있도록 하였다.

2.2 생성 모델 설계

감정별 소리 생성을 위해서는 SampleRNN 모델을 사용하였다. SampleRNN은 이름에서 알 수 있듯이 오디오 신호를 샘플 단위로 예측하는 RNN 기반의 모델이다. 오디오 신호는 1초에 수천 개의 샘플로 구성된다. SampleRNN은 이러한 샘플 단위의 매우 세밀한 시간 단위에서 작동하기 때문에 고양이가 내는 매우 짧은 소리를 처리하기에 적합하다. 구체적으로, SampleRNN은 계층적인 RNN 구조를 활용하여 이전 샘플

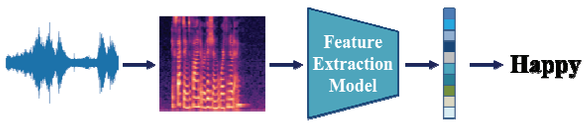


Fig. 1. Classification Model

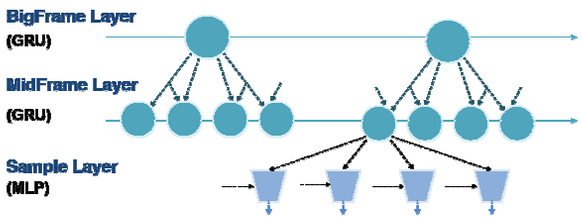


Fig. 2. Audio Generation Model

들로부터 다음 샘플 값을 예측하고, 이러한 예측을 통합하여 최종적인 오디오 시퀀스를 생성한다. 이 계층 구조는 여러 시간 해상도에서 패턴을 학습하고 이를 결합하여 자연스러운 오디오 신호를 생성을 할 수 있도록 한다. 각 계층 레이어는 상위 레이어의 출력을 입력으로 사용하며, 점점 더 세밀한 시간 범위에서 패턴을 포착하고 예측을 수행한다. 고양이 소리 생성을 위해 SampleRNN은 i) BigFrame, ii) MidFrame, iii) Sample 레이어의 3가지 계층으로 구성한다. BigFrame 레이어는 가장 상위에 위치하며, 긴 시간 범위의 패턴을 포착한다. BigFrame 레이어는 입력 데이터를 큰 단위로 묶어 처리하며, 오디오의 전반적인 구조나 장기적인 패턴을 학습한다. MidFrame 레이어는 BigFrame 레이어의 출력을 받아, 더 작은 단위로 데이터를 처리하여 중간 수준의 패턴을 학습한다. 이 레이어의 출력은 샘플 단위를 처리하는 Sample 레이어로 전달되어 개별 오디오 샘플을 예측하고, 예측된 샘플들은 순차적으로 통합되어 최종 오디오 시퀀스를 구성하게 된다. Sample 레이어는 최종 예측을 위하여 RNN 대신 MLP 구조를 사용한다. 각 계층은 자신이 담당하는 범위 내에서 데이터를 처리하여 점진적으로 세밀한 정보를 추가하여, 최종적으로 높은 해상도의 자연스러운 소리를 생성한다.

2.3 전체 시스템 모델

전체 시스템에서는 두 가지 기능을 통합하여 사용할 수 있도록 한다. 서버에 학습된 분류 및 생성 모델을 저장하여 요청에 따라 응답할 수 있는 시스템을 제안한다. 제안된 시스템에서는 학습된 분류 모델을 이용해 실제 고양이 소리를 녹음하여 인식 후 감정을 예측하고 이 예측된 값을 다시 반환한다. 또, 감정 혹은 의사 표현과 관련된 텍스트 파일을 입력으로 넣어주게 되면, 서버에서는 텍스트 파일의 태그(Tag) 값을 확인하여 적절한 음성 파일을 생성하여 다시 반환하여 준다.

3. 실 험

제안된 시스템을 검증하기 위하여 감정별 소리 분류 모델

Table 1. Number of Audio Files used for Training Classification and Generation Models

	Angry	Defense	Fighting	Happy	Hunting Mind	Mating	Mother Call	Paining	Resting	Warning
Classification	300	291	300	297	289	301	296	291	296	300
Generation	600	582	600	594	578	602	592	583	592	600

및 생성 모델에 대한 실험을 진행하였고, 고전적인 방식과 비교하여 적합성을 입증한다. 실험에 사용된 데이터 셋[7]의 감정별 파일 수는 Table 1에 명시 되어있다.

3.1 소리 데이터의 분류

감정별 소리를 분류할 때에는 고전적인 Clustering 방식이 종종 사용 된다. 본 논문에서는 가장 대표적인 k-means clustering을 진행해보았다. 10개의 감정에 대해서는 clustering이 어려워 비교적 쉬운 대표 6개의 감정에 대해 k-means clustering 진행하였다. 결과는 Fig. 3의 pie chart처럼 제대로 clustering 되지 못하는 모습을 보였다. 안쪽의 파이 차트가 바깥쪽의 숫자로 분류된 파일의 비율을 나타낸다. 파란색은 Angry, 초록색은 Happy, 빨간색은 Mating, 청록색은 MotherCall, 보라색은 Paining, 라임색은 Resting을 의미한다. 다른 기존의 clustering 방식도 시도하였으나 단순한 방식으로는 고양이 소리를 감정별로 제대로 분류해 낼 수 없었다. 그러나 제안된 모델의 경우, 6가지 감정에 대하여 83.6%의 정확도로, 10가지 감정에 대하여 78.9%의 정확도로 감정을 분류할 수 있었다. 또한 분류된 모델로 특징을 추출하여 T-SNE 시각화를 진행해 보았다. Fig. 4에서 그 결과를 확인할 수 있다. 각 색깔은 다른 Class를 의미하며, 감정별로 경계가 뚜렷하게 나타나며 Clustering이 되는 것을 보여준다. 따라서 분류 모델을 통해서 감정별 특징을 잘 학습하였음을 확인할 수 있다.

3.2 소리 데이터의 생성

소리 생성은 모든 감정별로 각각 모델을 학습하였으며, 감정별로 적절한 epoch수를 설정하여 학습을 완료하였다. 각각 감정별로 소리를 생성하기 위해서는 분류 작업보다 더 많은 데이터로 학습을 진행해야 했다. 따라서, 생성 모델 학습을 위해서 데이터 증강을 진행한 후 학습을 진행하였다. 데이터 증강 후의 파일 수는 Table 1에 명시되어있듯 감정 별로 기존 데이터의 약 2배, 각각 약 600개의 데이터를 사용하였다. 모든 감정별로 적절하게 SampleRNN을 이용하여 학습하고, 학습된 모델의 test과정을 거쳐 각각 감정별로 소리가 생성됨을 확인하였다. 이렇게 감정별로 생성된 소리를 보다 엄격한 기준을 통해 걸러내기 위해, 추가적으로 생성된 파일들에 classification 과정을 더하여, 감정별로 소리가 제대로 된 파일만 추출하도록 학습하였다. 이렇게 생성된 파일들의 예시

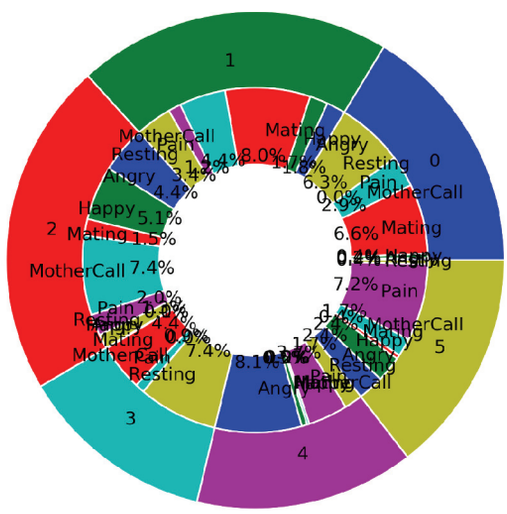


Fig. 3. K-means Clustering Results

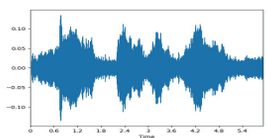
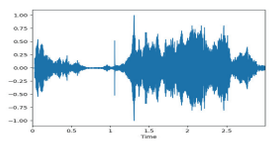
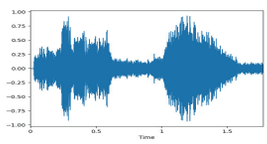
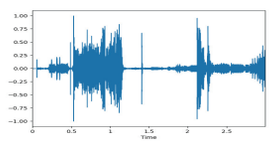
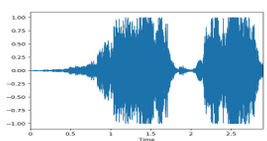
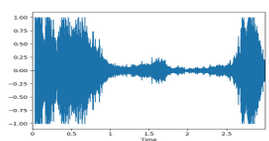
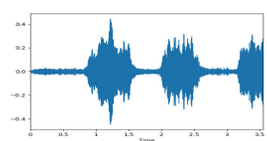
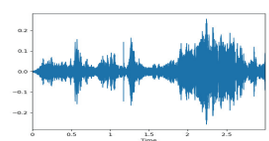
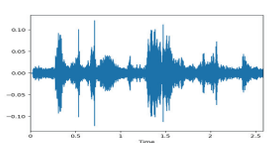
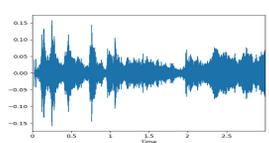
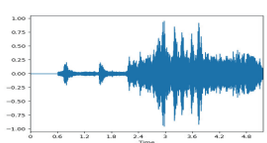
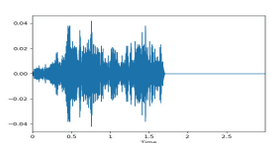
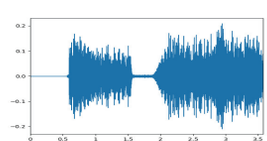
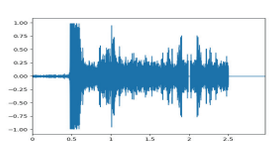
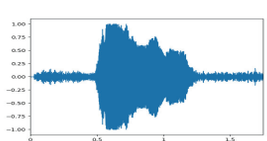
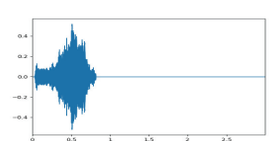
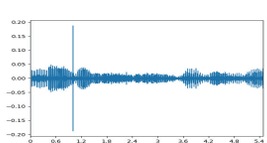
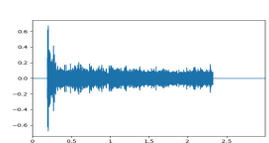
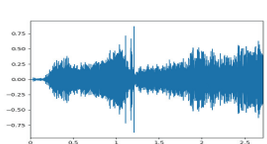
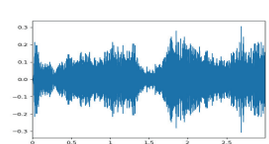
Fig. 4. T-SNE Results

파형은 Table 2에 제공되었다. 파형을 보면 감정별로 원본 데이터와 유사하게 소리를 생성해 냄을 확인 할 수 있다.

4. 결 론

본 논문에서는 감정별 고양이 소리 분류 및 생성 딥러닝 시스템을 제안하고 검증하였다. 이 시스템은 반려동물과의 상호작용을 더욱 풍부하게 만들 수 있는 가능성을 열어주며 감정 인식 기술의 발전에 기여할 수 있을 것으로 기대된다.

Table 2. Generation Result Samples

Categories	Real Audio Sample	Generated Audio Sample
Angry		
Defense		
Fighting		
Happy		
Hunting Mind		
Mating		
MotherCall		
Paining		
Resting		
Warning		

References

[1] G. R. Farley, S. M. Barlow, R. Netsell, and J. V. Chmelka, "Vocalizations in the cat: behavioral methodology and spectrographic analysis," *Experimental Brain Research*, Vol.89, pp.333-340, 1992.

[2] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proceedings of the International Conference on Neural Information Processing Systems*, Spain, pp.892-900, 2016.

[3] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," in *Proceedings of the International Conference on Platform Technology and Service*, Korea(South), pp.1-5, 2017.

[4] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to Sound: Generating Natural Sound for Videos in the Wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, USA, pp. 3550-3558, 2018.

[5] J . Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio *synthesis," arXiv preprint arXiv:1902.08710*, 2019.

[6] E. Kucukkulahli and A. T. Kabakus, "Towards Understanding Cat Vocalizations: A Novel Cat Sound Classification Model Based on Vision Transformers," *Applied Acoustics*, Vol.226, 110218, 2024.

[7] Y. R. Pandeya and J. Lee, "Domestic cat sound classification using transfer learning," *The International Journal of Fuzzy Logic and Intelligent Systems*, Vol.18, pp.154-160, 2018.

[8] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, Aaron, and Y. Bengio, "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model," in *Proceedings of the International Conference on Learning Representations*, 2017.

심 주 용

https://orcid.org/0000-0003-2828-9232
e-mail : shimjoo@korea.ac.kr
2019년 고려대학교 전기전자전파공학부 (학사)
2024년 고려대학교 전기전자공학과(박사)
2024년 ~ 현 재 고려대학교 정보통신기술연구소 박사후연구원
관심분야 : Multi-Modal & Generative Models

임 성 기

https://orcid.org/0009-0003-7793-4215
e-mail : lsk1101@gmail.com
1996년 고려대학교 전자공학과(학사)
1998년 서울대학교 전기공학부(석사)
2012년 MITSloan 경영대학원(MBA)
1998년 ~ 2001년 삼성전자 컴퓨터/디스플레이사업부 선임연구원
2001년 ~ 2002년 GON테크놀로지 선임연구원
2003년 ~ 2003년 LG전자 모바일커뮤니케이션사업부 선임연구원
2004년 ~ 2020년 KBS 한국방송 팀장
2021년 ~ 현 재 ㈜애니멀보이스 대표이사
관심분야 : Human-AnimalCommunication & Deep Learning

김 종 국

https://orcid.org/0000-0003-1828-7807
e-mail : jongkook@korea.ac.kr
1998년 고려대학교 전자공학과(학사)
2000년 Purdue University 전기 및 컴퓨터공학(석사)
2004년 Purdue University 전기 및 컴퓨터공학(박사)
2005년 ~ 2007년 삼성 SDS IT R&D 센터 선임연구원
2007년 ~ 현 재 고려대학교 전기전자공학과 교수
관심분야 : 딥러닝(Deep Learning) & 이기종 분산 컴퓨팅 (Heterogeneous Distributed Computing)