한국어 발화 문장에 대한 비언어 표현 정보를 자동으로 생성하는 모델

김재윤[1○], 장진예[1], 김산[1], 정민영[1], 강현욱[2], 신사임[1]
[1]한국전자기술연구원 인공지능연구센터, [2]전남대학교 기계공학부
{ovm2316, jinyea.jang, kimsan0622, minyoung.jung, sishin}@keti.re.kr, kanghw@jnu.ac.kr

A Model to Automatically Generate Non-verbal Expression Information for Korean Utterance Sentence

Jaeyoon Kim[1○], Jinyea Jang[1], San Kim[1], Minyoung Jung[1], Hyunwook Kang[2], Saim Shin[1]
[1]Korea Electronics Technology Institute & [2]Chonnam National University

요 약

자연스러운 상호작용이 가능한 인공지능 에이전트를 개발하기 위해서는 언어적 표현뿐 아니라, 비언어적 표현 또한 고려되어야 한다. 본 논문에서는 한국어 발화문으로부터 비언어적 표현인 모션을 생성하는 연구를 소개한다. 유튜브 영상으로부터 데이터셋을 구축하고, Text to Motion의 기존 모델인 T2M-GPT와 이종 모달리티 데이터를 연계 학습한 VL-KE-T5의 언어 인코더를 활용하여 구현한 모델로 실험을 진행하였다. 실험 결과, 한국어 발화 텍스트에 대해 생성된 모션 표현은 FID 스코어 0.11의 성능으로 나타났으며, 한국어 발화 정보 기반 비언어 표현 정보 생성의 가능성을 보여주었다.

주제어: 멀티 모달, Text to Motion, T2M-GPT, 한국어 발화 및 비언어 표현의 관계

1. 서론

대화나 담화에서 화자가 전달하고자 하는 내용은 음성 텍스트인 언어 정보와 목소리 톤, 어조, 제스처 등의 비언어적 정보의 조합으로 표현된다 [1]. 이러한 복합 표현은 사용자와 인공지능 에이전트 간의 자연스러운 상호작용에 중요한 요소이며, [2, 3] 기술 구현에 있어 멀티모달 관점의 접근이 필요하다.

최근 자연어 처리와 컴퓨터 비전 분야에서는 서로 다른 모달리티 간의 전환에 대한 연구 방향이 제시되고 있다 [4, 5, 6]. 그중, 사람의 동작에 대한 텍스트 설명으로부터 모션을 생성하는 연구는 이루어져 왔지만, 한국어 및 발화 정보로부터 모션 표현을 생성하는 연구 사례는 찾아보기 어렵다.

본 연구에서는 한국어 발화 텍스트로부터 발화의 비언어적 정보 표현인 모션을 생성하는 연구를 진행한다. 발화 텍스트와 비언어 모션의 두 모달리티 간 연관성을 모델링하여 한국어 발화에 대한 자연스러운 모션 표현을 생성하는 새로운 방법을 시도해 보고, 자동으로 생성된 비언어 표현의 활용 가능성을 살펴본다.

본 논문의 구성으로는 2장에서 관련 연구를 요약하고, 3장에서는 활용한 생성 모델의 구조와 동작 원리를 설명한다. 4장에서는 사용된 데이터에 대한 설명과 실험 과정 및 결과를 서술하고, 5장 결론 부분에서는 본 연구의 한계점과 향후 연구 계획을 정리하며 마무리한다.

2. 관련 연구

텍스트 설명으로부터 비언어 동작 모션을 생성하는 모델은 MotionDiffuse [7], T2M-GPT [8] 등이 있다.

MotionDiffuse 모델은 텍스트 입력을 트랜스포머(Transformer) [9] 기반의 아키텍처를 사용하여 임베딩한 뒤, 해당 임베딩 벡터와 랜덤한 노이즈 벡터를 입력으로 받아, 디퓨전 모델 [10]을 통해 다양한 모션을 생성할 수 있다.

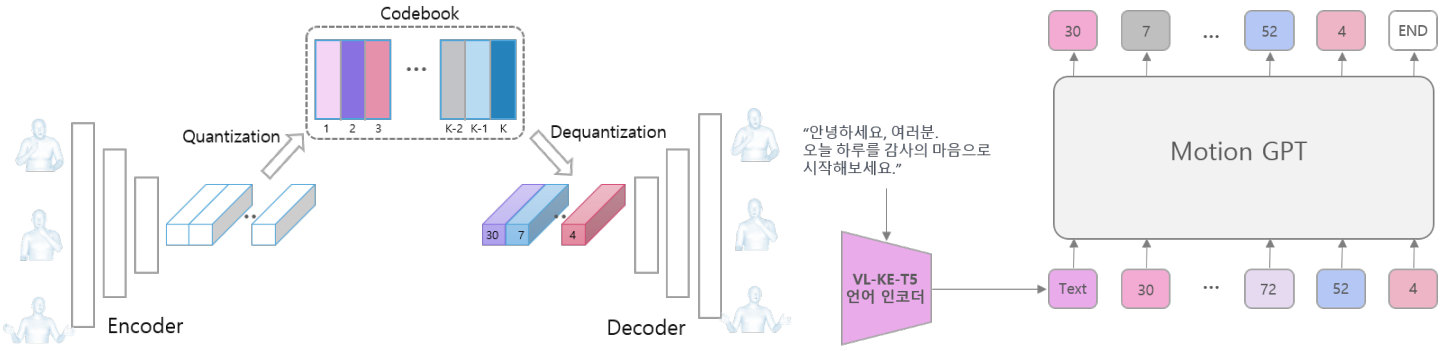
T2M-GPT는 두 개의 모듈로 구성되어 있으며, 먼저 VQ-VAE [11]를 통해 모션의 연속성 정보를 담은 코드북을 생성한다. 그다음, 학습된 코드북과 트랜스포머 아키텍처에 기반해서 입력 텍스트로부터 모션 시퀀스를 생성한다. 이는 다시 VQ-VAE를 통해 디코딩되어 모션으로 표현된다.

3. 모델

본 연구에서는 Text to Motion의 기존 모델들 중 T2M-GPT를 활용하여 발화 텍스트에 대한 비언어 동작을 생성 모델을 구현하였다. 앞서 언급한 바와 같이, 해당 모델은 그림 1과 같이 두 개의 큰 모듈로 구성되어 있으며 본 장에서는 각 모듈에 대해서 설명한다.

3.1 VQ-VAE

그림 1 (a)의 VQ-VAE는 연속적인 모션 데이터들로부터 이산적인 표현의 코드북을 생성한다. VQ-VAE는 변이형 오토인코더와 이산 표현을 결합한 생성 모델로 VQ-VAE의 인코더의 입력으로 모션 X = [x1, x2, x3 ...]이 들어가면, 이를 디코더 단에서 복원하는 태스크를 학습하면서 임베딩 공간 상에 코드북을 구축한다. 구축된 코드북은 그림 1 (b) 모듈에서 모션 시퀀스 예측에 사용된다. 예측된 모션 시퀀스로부터 모션을 역으로 생성 시에도 VQ-VAE를 사용한다.



(a) VQ-VAE (b) VL-KE-T5 언어 인코더 & GPT

그림 1. 한국어 발화로부터 모션 표현을 생성하는 모델 구조

3.2 VL-KE-T5 언어 인코더 & GPT

그림 1 (b)는 모션 GPT로 발화 텍스트 입력에 대한 비언어 모션 시퀀스를 예측하는 모듈이다. 기존 T2M-GPT 모델에서는 OpenAI에서 개발한 CLIP [12]의 언어 인코더를 사용하는데, 본 연구에서는 한국어 발화 텍스트 입력 처리를 위해 해당 부분을 VL-KE-T5[1]로 대체하였다. VL-KE-T5는 한국전자기술연구원에서 개발한 KE-T5 [13]와 구글에서 개발한 ViT(Vision Transformers) [14]의 임베딩 벡터들을 영상-언어 병렬 데이터를 이용하여 정렬한 모델이다.

모듈의 학습 과정을 설명하면 다음과 같다. 발화 텍스트가 입력으로 들어가면, VL-KE-T5 언어 인코더는 텍스트의 임베딩 값값을 추출한다. 추출된 임베딩 값은 트랜스포머 디코더 모델의 입력으로 들어가 첫 번째 모션 코드를 예측하게 된다. 순차적으로 추론되는 모션 코드 시퀀스는 로그 가능도(Log Likelihood)가 최대화되도록 학습되며, 아래의 수식(1)와 같이 손실 함수를 표현할 수 있다.

$$L_{\text{gpt}} = \mathbb{E}_{S\sim p(S)}[-\log p(S|c)] \quad (1)$$

$p(S|c)$는 주어진 발화 텍스트 c에 대한 시퀀스 S의 확률을 나타낸다. 해당 확률을 최대화함으로써, 트랜스포머 디코더 모듈의 발화 텍스트에 해당하는 모션을 높은 확률로 생성하도록 학습된다.

4. 데이터셋 구축 및 실험

4.1 학습 데이터 구성

본 연구에 활용된 데이터는 1인 유튜버 영상으로부터 자체적으로 구축한 데이터셋이다. 유튜브 영상으로부터 모델 학습에 필요한 한국어 발화 텍스트 및 모션 데이터를 얻기 위한 처리 과정을 아래와 같이 기술한다.

[1] https://github.com/AIRC-KETI/VL-KE-T5

4.1.1 발화 전사 및 발화 단위의 클립 분할

발화 텍스트 전사는 OpenAI에서 개발한 ASR(Automatic Speech Recognition) 모델인 Whisper [15]의 large-v2 모델을 사용하였다. ASR 결과로 함께 출력되는 발화 시간 정보를 이용하여 발화 단위로 영상 클립 분할을 진행하였다.

4.1.2 자세 추정 모델 기반의 SMPL 파라미터 추출

본 연구에서는 수집 영상에 포함된 발화자의 움직임 정보를 추출하기 위해 3D 자세 추정이 가능한 OSX [16] 모델을 활용하였으며, 해당 모델을 통해 클립 영상 속 사람의 관절과 자세를 SMPL-X [17] 파라미터 형식으로 추정하였다. 여러 부위에 대한 추정 값 중 최종적인 모델 학습을 위해 골반(root), 몸체(body), 손(lhand, rhand)에 해당하는 52개 관절의 파라미터만을 추출하였다. 이때, 대부분의 영상 속 발화자는 하반신이 비추어지지 않기 때문에 골반부터 하반신에 해당되는 파라미터는 서있는 자세를 표현하는 값으로 대치하고, 상반신에 해당하는 몸체와 손의 움직임만이 고려되었다.

4.1.3 최종 데이터셋(KTubeUM) 구성

발화 텍스트의 경우, 토큰화와 품사 태깅이 필요하여 KoNLPy [18]패키지의 MeCab을 사용해서 진행하였다. 최종적으로 구성된 **KTubeUM**(Korean Youtube Utterance and Motion) 데이터셋의 훈련 및 평가 데이터의 분포는 표 1과 같다.

표 1. 데이터셋 정보

	KTubeUM
훈련용 데이터	20,732 (80%)
검증용 데이터	1,350 (5%)
평가용 데이터	4,020 (15%)
총합	26,102

4.2 실험

본 실험은 VQ-VAE 모듈을 두 가지 데이터셋으로 각각 학습시킨 후, 각각의 VQ-VAE에 기반하여 KTubeUM 데이터셋을 모션 GPT 모듈 학습에 적용하고 모션 생성의 결과를 정량적으로 확인하였다. 표 2는 두 가지 모듈 학습에 공통적으로 적용된 주요 하이퍼 파라미터를 정리한 표이다.

표 2. 학습 시 설정 파라미터

	VQ-VAE	GPT
batch size	256	256
learning rate	2e-4	0.0001
total iter	50000	100000
weight decay	0.0	1e-6
optimizer	AdamW	AdamW

4.2.1 VQ-VAE 학습

본 실험에서는 자체 구축한 KTubeUM 데이터셋과 기존 모델에서 사용한 HumanML3D [4] 데이터셋으로부터 각각 모션에 대한 코드북을 구성하는 모델의 학습을 진행하여 모션 생성에서의 성능을 비교한다. 두 모델의 구조와 관련한 파라미터는 동일하게 적용하였다. KTubeUM 데이터셋의 모션이 갖는 표현은 상반신 및 손의 움직임으로 한정되어 있기 때문에 모션 토큰의 임베딩 차원은 기존 T2M-GPT에서 설정한 512보다 작은 128로 설정하였다. 또한, VQ-VAE의 코드북 차원과 VL-KE-T5의 임베딩 차원이 반드시 동일한 크기일 필요는 없지만, VL-KE-T5의 임베딩 차원과 동일하도록 코드북 차원을 768로 설정하여 학습을 진행하였다. GPT 모듈 학습에 사용되는 VQ-VAE는 FID(Fréchet Inception Distance) 스코어를 기준으로 가장 좋은 성능을 내는 모델로 선정하였다.

4.2.2 GPT 학습

두 번째로, 앞서 구축한 두 종류의 VQ-VAE에 기반하여 GPT 학습을 진행하였다. 3장에서 언급한 바와 같이, VQ-VAE 모듈을 통해 영상 클립의 모든 모션의 코드 시퀀스가 생성되고, VL-KE-T5 언어 인코더로부터 추출된 텍스트의 임베딩 값이 입력으로 들어가면 트랜스포머 아키텍처에 기반하여 발화 텍스트로부터 모션 코드 시퀀스의 생성을 학습하게 된다. 사전 학습된 VL-KE-T5 언어 인코더는 768 차원으로 텍스트를 임베딩하기 때문에 이를 고려하여 GPT의 임베딩 차원을 1536으로 설정하였다. 또한, 9개의 층과 16개의 어텐션 헤드를 사용하도록 트랜스포머 아키텍처를 구성하였다.

4.2.3 결과

표 3은 두 종료의 데이터셋으로 구축한 코드북에 기반하여 생성된 비언어 모션 표현에 대한 정량적인 평가 결과이다. 아래 세 가지 평가 지표로 비교하였다.

- **FID (Fréchet Inception Distance)**: 생성된 모션과 실제 모션 사이의 분포 거리를 계산하여 측정된다. 따라서 값이 낮을수록 좋은 표현 모션의 생성을 의미한다.
- **Diversity**: 생성된 모션 데이터 중 300쌍을 무작위로 추출하여 이들 간 평균 유클리드 거리를 계산하여 측정된다. 얼마나 다양한 모션을 생성하였는지에 대한 척도가 되며, 원본 모션에서의 측정값과 생성된 표현 모션의 측정값의 차이가 크지 않을수록 좋은 표현 모션의 생성을 의미한다.
- **Multimodal Distance**: 입력 텍스트 임베딩값과 생성된 모션 사이의 평균 유클리드 거리를 계산하여 측정된다. 원본 모션에서의 측정값과 생성된 표현 모션의 측정값의 차이가 크지 않을수록 좋은 표현 모션의 생성을 의미한다.

표 3. 실험 결과

데이터셋	FID	MM-Dist	Diversity
HumanML3D	15.80	14.47(±0.6)	3.99(±2.48)
KTubeUM	0.11	13.89(±0.1)	1.44(±0.07)

HumanML3D 데이터셋 기반의 코드북을 사용했을 때, 보다 다양한 모션 표현이 생성되었지만, KTubeUM 데이터셋 기반의 코드북을 사용했을 때, 원본 모션과 더욱 유사한 모션 표현이 생성되었다. 이는 비언어적 발화 표현 정보에 특화된 코드북 사용이 모션 표현의 정교한 예측에 도움이 되었음을 보여준다.

5. 결론

본 연구에서는 유튜브 영상으로부터 자체적으로 데이터셋을 구축하여 한국어 발화 텍스트로부터 자연스러운 비언어 모션 표현을 생성하는 모델을 구현하고 실험을 진행하였다. 두 종류의 코드북에 기반한 모션 표현의 생성 결과가 정량적으로 어떠한 차이를 보이는지 비교 실험을 진행함으로써 발화의 언어 정보를 기반으로 한 비언어 표현 정보의 자동적 생성 가능성을 보여주었다.

발화의 언어적 표현과 비언어적 표현은 복잡한 요소와 맥락에 의해 조절되므로 본 연구는 개선의 여지가 존재한다. 보다 자연스러운 비언어적 표현을 위해 또 다른 모달리티를 추가 적용해 볼 수 있다. 또한, 어절 및 형태소 단위로 모션 표현과의 관계를 학습하기 위해 발화문 내에 스페셜 토큰을 추가하여 보다 정교한 모션 표현의 생성을 위한 연구를 진행할 예정이다.

감사의 글

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원(No. 2022-0-00608)과 정보통신산업진흥원(S0102-23-1008)의 지원을 받아 수행된 연구임

참고문헌

[1] 김한샘, "한국어 화자의 비언어 행위 연구," *한국언어문화교육학회 학술대회*, pp. 67–71, 2015.

[2] 최미란, "감성 기반 멀티모달 대화 시스템의 기술 및 표준화 현황," *한국통신학회지 (정보와통신)*, Vol. 40, No. 3, pp. 51–58, 2023.

[3] 김기락, 연희연, 은태영, 정문열 *et al.*, "감정에 기반한 가상인간의 대화 및 표정 실시간 생성 시스템 구현," *Journal of the Korea Computer Graphics Society*, Vol. 28, No. 3, pp. 23–29, 2022.

[4] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022.

[5] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "Motionclip: Exposing human motion generation to clip space," *European Conference on Computer Vision*, pp. 358–374, 2022.

[6] H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, D. Tao, and M. J. Black, "Generating holistic 3d human motion from speech," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 469–480, 2023.

[7] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *arXiv preprint arXiv:2208.15001*, 2022.

[8] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen, "T2m-gpt: Generating human motion from textual descriptions with discrete representations," *arXiv preprint arXiv:2301.06052*, 2023.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, Vol. 30, 2017.

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, Vol. 33, pp. 6840–6851, 2020.

[11] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, Vol. 30, 2017.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," *International conference on machine learning*, pp. 8748–8763, 2021.

[13] 신사임, 김산, and 서현태, "KE-T5: 한국어-영어 대용량 텍스트를 활용한 이중언어 사전학습기반 대형 언어모델 구축," *제 33회 한글 및 한국어 정보처리 학술발표 논문집*, pp. 419–422, 2021.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *International Conference on Machine Learning*, pp. 28 492–28 518, 2023.

[16] J. Lin, A. Zeng, H. Wang, L. Zhang, and Y. Li, "One-stage 3d whole-body mesh recovery with component aware transformer," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21 159–21 168, 2023.

[17] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10 975–10 985, 2019.

[18] E. L. Park and S. Cho, "Konlpy: Korean natural language processing in python," *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, Vol. 6, pp. 133–136, 2014.