상용 API 의 감정에 따른 음성 인식 성능 비교 연구

양장훈 [1]
[1] 서울미디어대학원대학교 인공지능응용소프트웨어학과 교수

jhyang@smit.ac.kr

A Study on the Comparison of the Commercial API for Recognizing Speech with Emotion

Janghoon Yang[1]
[1]Dept. of AI Software Engineering, Seoul Media Institute of Technology

요 약

최근 인공지능 기술의 발전에 따라서 다양한 서비스에서 음성 인식을 활용한 서비스를 제공하면서 음성 인식에 대한 중요성이 증가하고 있다. 이 논문에서는 국내에서 많이 사용되고 있는 대표적인 인공지능 서비스 API 를 제공하는 구글, ETRI, 네이버에 대해서 감정 음성 관점에서 그 차이를 평가하였다. AI Hub 에서 제공하는 감성 대화 말뭉치 데이터 셋의 일부인 음성 테스트 데이터를 사용하여 평가한 결과 ETRI API 가 문자 오류율 (1.29%)과 단어 오류율(10.1%)의 성능 지표에 대해서 가장 우수한 음성 인식 성능을 보임을 확인하였다.

1. 서론

인공지능 스피커의 상용화와 스마트 폰에서 인공지능 에이전트가 도입되면서, 음성 인식 기술에 많은 발전이 있어왔다. 특히, 딥러닝 기반 언어 모델에 사용되는 트랜스포머 (Transformer)는 음성 인식을 위해서 주요한 특징을 뽑아내는 방법으로 사용되었을 때에 음성 인식 성능이 크게 향상 될 수 있음이 확인되었다. [1]

딥 러닝을 활용한 기존의 음성 인식은 사용되는 데이터 셋에 따라서 성능의 차이를 보이고 있다. wav2vec 2.0[2]의 사전 학습 모델을 이용하여 음성 인식을 사용하는 모델은 LibriSpeech 데이터 셋에 대해서 단어 오류율 (word error rate, WER)을 1.4%을 달성하였다 [3]. 시퀀스를 처리하는 전통적인 신경망 구조인 LSTM 과 Confomer 를 결합하여 Swichboard 300 데이터 셋에 대해서 단어 오류율를 5.0%를 달성하였다 [4]. [4]와 유사한 형태의 구조를 갖는 인경망 구조에 대해서 다양한 복수의 데이터 셋을 가지고 10 억개의 파라미터를 갖는 거대 모델을 학습시킬 경우 WSJ 데이터 셋에 대해서 단어 오류율 1.3%를 달성하였다 [5]. 기존 연구 결과를 살펴 볼 때에 다양한 데이터 셋에 대해서, wav2vec 2.0 과 Conformer 를 활용한 딥러닝 알고리즘이 우수한 성능을 갖는 것을 확인할 수 있다.

음성 인식 기술의 발달에 따라서 음성 인식의 활용도가 증가하면서, 다양한 음성 인식을 위한 API 을 다수의 기업에서 제공하고 있다. [6]에서는 공용 방송의 10 개 분야의 뉴스 음성 데이터를 수집하여 네이버, 카카오, ETRI, 구글, 아마존, IBM, MS 에 대해서 비교 테스트를 수행하였고, 정확도 성능 지표에 대해서 카카오 API 가 가장 우수한 성능을 보였다. [7]에서는 보다 현실적인 상황에서의 API 의 성능을 평가하기 위해서 ETRI 에서 제공하는 다채널 잡음 처리 기술 개발 및 평가용 데이터를 이용하여 아마존, 아주어, 구글, 카카오, 네이버에 대해서 비교 평가한 결과 구글이 가장 우수한 53% 단어 오류율을 제공함을 확인하였다.

[6]과 [7]의 결과를 통해 확인할 수 있는 사항은 어떤 데이터 셋을 사용하는지, 그 데이터 셋이 생성되는 녹음 환경은 어떤 상황인지 등에 따라 달라지고, API 를 제공하는 회사가 얼마나 자주 최신 기술을 적용해서 API 의 성능을 개선하는지에 따라서 성능이 달라질 수 있음을 확인할 수 있다. 이 연구에서는 감정 인식 서비스 시장이 향후 크게 성장할 것을 고려하여 감정 인식 데이터에 대해서 2023 년도에 사용되고 있는 상용 API 의 성능이 발화되는 감정에 따라서 어떤 성능의 차이를 갖는지를 확인하고자 한다.

2. 실험 방법 및 결과

비교 대상: 구글, 네이버, ETRI 에서 API 형태로 제공

하는 음성 인식 성능

데이터 셋: AI Hub 에서 제공하는 감성 대화 말뭉치 데이터 셋의 일부인 음성 테스트 데이터를 사용하였다. 이 데이터 셋은 남성과 여성 음성을 각각 5000 개로 구성하고 있으며 음성의 평균 문자수는 27.38 개 평균 단어수는 9.5 개로 구성되어 있다. 6 개의 감정인 상처, 기쁨, 불안, 당황, 분노, 슬픔에 대해서 성별로 동일한 숫자의 데이터를 제공하며, 감정별로 평균 833.3 개의 데이터와 19.96 개의 편차를 가지고 있어서 감정별 데이터 수의 차이는 크지 않다.

성능 지표: 음성 인식의 성능에 일반적으로 많이 사용되는 문자 오류율(CER)과 단어 오류율(WER)을 사용하여 성능을 평가한다. 각각의 오류율은 다음과 같이 정의된다. [7]

$$CER = \frac{C_S + C_D + C_I}{C_C} \quad (1)$$

$$WER = \frac{W_S + W_D + W_I}{W_C} \quad (2)$$

위 식에서 C_S, C_D, C_I, C_C 는 각각 인식한 음성을 목표 문서와 동일하게 만들기 위해서 필요로 하는 치환된 문자 수, 제거된 문자 수, 새롭게 추가된 문자 수, 목표 문서에서의 문자 수를 나타내고, W_S, W_D, W_I, W_C 는 동일한 맥락에서 단어 단위로 정의되는 수를 나타낸다.

실험 결과: 표-1 에서는 감정과 성별에 따라서 고려한 API 성능을 CER 관점에서 측정한 결과를 정리하였다. 구글, ETRI, 네이버 API 에서 감정과 성별에 대해서 모두 ETRI API 가 우수한 성능을 보이고 네이버 API 상대적으로 열악한 성능을 보이고 있음을 확인할 수 있다. 표-2 에서 정리한 WER 성능에 대해서 각 API 의 성능의 우위는 동일하게 나타나고 있음을 확인할 수 있다. 네이버 API 에 대한 주요한 오류의 유형 사례를 표-3 에서 정리하였다. 주로 음성의 뒷 부분을 처리하거나 띄어쓰기에 있어서 오류가 발생하고, 이는 침묵 구간 검출 부분이나 음성 데이터 표현과 같은 부분에서 문제가 있는 것으로 추정된다. 구글과 ETRI API 에서는 주요한 오류의 유형이 치환된 문자나 단어인 반면에 네이버의 경우에는 제거된 문자의 비중과 치환된 문자가 주요한 오류의 유형이면서 그 비중이 비슷한 것을 확인할 수 있었다.

네이버 API 의 성능의 제한 때문에 감정과 성별에 따른 API 의 감정 인식에 대한 분석은 구글과 ETRI API 성능 결과 위주로 분석한다. CER 성능 관점에서는 불안 감정을 갖는 음성을 가장 잘 인식하고 상처 감정에 대해서 음성 인식 성능이 가장 열화됨을 보인다. 가장 좋은 경우와 가장 나쁜 경우의 성능의 차이가 각각 49%와 266%정도의 차이를 갖는다는 것을 고려시 감정 상태가 API 의 음성 인식의 차이를 만들 수 있다고 추정할 수 있다. 또한 WER 성능 관점에서는 CER 과 마찬가지로 불안의 감정을 갖는 음성에 대해서 가장 우수한 성능을 보이는 반면에 구글은 분노의 감정, ETRI 는 기쁨의 감정을 갖는 음성에 대해서 가장 열악한 성능을 보이면서 그 차이가 각각 13.5%와 13.9%로서 여전히 의미있는 차이를 보이는 것으로 추정된다. 가장 좋은 성능이나 가장 열화된 성능을 보이는 감정들이 CER 과 WER 에 있어서 차이를 갖는 것은 감정 상태에 따라서 말하는 속도, 높이, 발음의 정확도 등에 따라서, 문자의 인식도와 단어의 인식에 있어서 차이를 갖기 때문으로 추정된다. 또한 남성과 여성의 음성 인식 성능의 차이에 있어서 구글 API 는 성별에 따른 성능의 차이가 크지 않으나, ETRI API 는 상대적으로 성별에 따른 성능의 차이가 상당한 것으로 보여지고 있다. 이는 API 에 사용된 인공지능 모델을 학습할 때에 사용된 데이터의 특징에 의해서 이러한 성능의 차이가 발생한 것으로 추정할 수 있다.

<표 1> API 의 감정과 성별에 따른 CER 성능

성별		상처	기쁨	불안	당황	분노	슬픔	평균
남성	구글	0.0576	0.0383	0.0455	0.0431	0.0530	0.0425	0.0467
	ETRI	0.0085	0.0072	0.0078	0.0072	0.0090	0.0073	0.0078
	네이버	0.1169	0.0982	0.0975	0.1003	0.1053	0.0992	0.1029
여성	구글	0.0556	0.0551	0.0305	0.0402	0.0321	0.0568	0.0450
	ETRI	0.0289	0.0266	0.0063	0.0112	0.0083	0.0264	0.0180
	네이버	0.0906	0.0953	0.0712	0.0721	0.0791	0.0935	0.0836
평균	구글	0.0566	0.0467	0.0380	0.0416	0.0426	0.0496	0.0459
	ETRI	0.0187	0.0169	0.0070	0.0092	0.0087	0.0169	0.0129
	네이버	0.1037	0.0968	0.0843	0.0862	0.0922	0.0964	0.0933

<표 2> API 의 감정과 성별에 따른 WER 성능

성별		상처	기쁨	불안	당황	분노	슬픔	평균
남성	구글	0.2238	0.1922	0.2118	0.2093	0.2350	0.2044	0.2128
	ETRI	0.0861	0.0859	0.0823	0.0873	0.0902	0.0967	0.0881
	네이버	0.3379	0.3170	0.3452	0.3237	0.3504	0.3219	0.3327
여성	구글	0.2119	0.2187	0.1733	0.1844	0.2020	0.1988	0.1982
	ETRI	0.1276	0.1289	0.1064	0.1064	0.1049	0.1143	0.1148
	네이버	0.3224	0.3217	0.2919	0.2860	0.3164	0.3217	0.3100
평균	구글	0.2178	0.2055	0.1925	0.1969	0.2185	0.2016	0.2055
	ETRI	0.1069	0.1074	0.0943	0.0968	0.0975	0.1055	0.1014
	네이버	0.3302	0.3193	0.3186	0.3048	0.3334	0.3218	0.3213

<표 3> 네이버 API 의 음성 인식 오류 사례

경우		문장
1	추정	우리 아빠는 나한테 제대로 된 선물 한번 준 적 없어
	정답	우리 아빠는 나한테 제대로 된 선물 한 번 준 적 없으셔.
2	추정	아무도 내 생일은 안
	정답	아무도 내 생일은 안 챙겨줘.
3	추정	내가 네 생각을 말하려고 할때마다 아빠는 나를 때리 죠
	정답	내가 내 생각을 말하려고 할 때마다 아빠는 나를 때리셨어.

3. 결론

이 논문에서는 대표적인 음성 인식 상용 API 인 구글, ETRI, 네이버 API 에 대해서 감정을 표현하는 음성 데이터 셋에 대해서 성능을 평가하였다. ETRI API 가 가장 우수한 성능을 보이는 반면에 네이버 API 는 음성의 후반 처리에 있어서의 오류에 의해서 상대적으로 열화된 성능을 가짐을 확인하였다. 구글과 ETRI 의 성능의 차이를 고려할 때에 ETRI 에 사용된 인공지능 모델에 본 실험에 사용된 데이터 셋이 사용될 가능성 있기 때문에 최신의 다양한 데이터 셋에 대해서 평가를 해야 보다 정확하게 API 의 성능의 차이를 확인할 수 있을 것으로 예상된다. 또한, 상용 API 의 성능을 보완하기 위해서 각 API 의 출력에 대해서 추가적인 후처리 인공지능을 도입할 때에 최종적인 성능이 어떻게 달라지는 보는 것도 의미있는 연구가 될 것으로 예상된다.

감사의 글

본 연구는 문화체육관광부 "관광서비스 혁신성장 연구개발사업" (R202202015)의 지원에 의해서 수행되었음

참고문헌

[1] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov,, and A. Mohamed "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units", arXiv:2106.07447v1 [cs.CL], 2021.

[2] A. Baevski, H.. Zhou, A. Mohamed, and M. Aul, ". wav2vec 2.0: A framework for self-supervised learning of speech representation,". arXiv preprint arXiv:2006.11477, 2020.

[3] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition," arXiv:2010.10504, 2020.

[4] Z. Tüske, G. Saon, and B. Kingsbury, "On the limit of English conversational speech recognition," arXiv:2105.00982v1, 2021. ·

[5] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi, "SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network," arXiv:2104.02133v3, 2021.

[6] 유현재, 김명화, 박상길, 김광용, "클라우드 기반의 음성인식 오픈 API 의 응용 분야별 한국어 연속음성인식 정확도 비교 분석," 한국통신학회논문지 제 45 권 제 10 호, 1,793 - 1,803, 2020.

[7] 이건희, 이상화, 지수환, 김아욱, 임현승, "소음 환경에서 상용 음성인식 API 의 성능 비교," 전기학회논문지 제 71 권 제 9 호, 1,266 - 1,273, 2022.