딥러닝 기반 종단형 음성 코덱의 학습 안정화

안성환, 김정훈, 김민찬, 김세민, 문성환, 이동준, 정명훈, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실

{shahn, jhkim, mckim, smkim21, shmun, djlee, mhjeong}@hi.snu.ac.kr, nkim@snu.ac.kr

Towards Stable Training of Deep Learning-Based End-to-end Speech Codec

Sunghwan Ahn, Jeunghun Kim, Minchan Kim, Semin Kim, Sunghwan Mun,
Dongjune Lee, Myeonghun Jeong, Nam Soo Kim
Human Interface Laboratory,
Department of Electrical and Computer Engineering and INMC,
Seoul National University

요 약

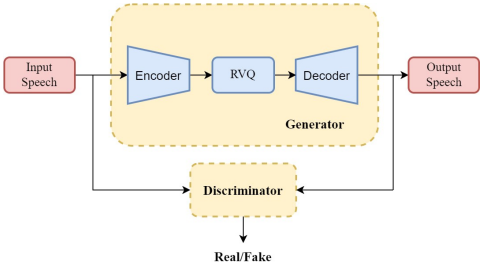
본 논문은 기존의 딥러닝 기반 종단형 음성 코덱 모델에서 학습을 안정화하는 방법을 제안한다. 한국어 및 영어 음성 데이터셋에 대하여 본 방법을 적용하였을 때 기존과는 달리 안정적으로 학습이 되는 것을 확인할 수 있었다.

Ⅰ. 서론

최근 딥러닝 연구의 활성화로 인해 음성코덱 분야에서도 딥러닝 기반 종단형 모델이 많은 연구가 되고 있다. 하지만 기존의 모델을 실제로 학습해 보면 학습이 매우 불안정한 모습을 확인할 수 있다. 본 논문에서는 기존 모델에서 학습이 불안정한 이유를 분석한 후, 해당 문제를 해결할 수 있는 기법을 제안한다. 영어 데이터와 한국어 데이터 각각에 대해 본 기법을 적용하여 학습한 결과 안정적으로 학습이 되는 것을 실험적으로 보였다.

Ⅱ. 본론

음성 코덱은 음성을 압축하는 Encoder, 양자화 하는 Quantizer, 음성을 복원하는 Decoder 로 이루어져 있다. 인코더는 음성을 최대한 압축하고 디코더는 원본 음성과 최대한 비슷하게 복원하는 것이 코덱의 목표이다. 최근 encoder, quantizer 및 decoder 를 모두 딥러닝 모델로 대체하는 종단형 실시간 음성 코덱인 Soundstream[1], Encodec[2] 등의 모델이 개발되었다. 하지만 해당 모델을 학습해 보면 학습이 불안정한 경우가 있다. 본 논문에서는 Soundstream 모델을 baseline 으로 하여, 학습이 불안정한 이유를 분석하고 해결책을 제시한다.

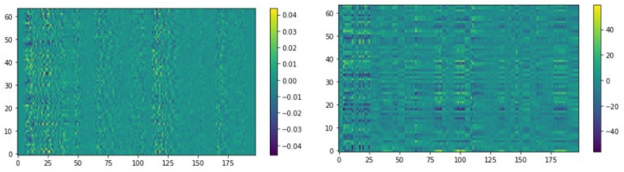


[그림 1] Soundstream 모델 구조. RVQ 는 residual vector quantizer 를 의미한다.

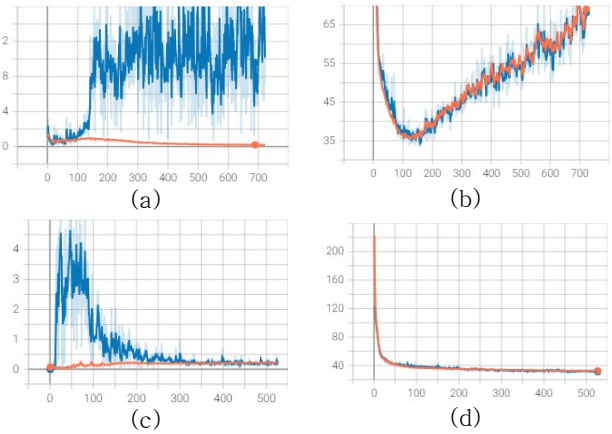
Soundstream 의 모델 구조는 [그림 1]과 같다. Generator 는 음성을 입력으로 받으며, 1D convolutional layers 로 이루어진 encoder, residual vector quantizer, 그리고 1D transposed convolutional layers 로 이루어진 decoder 를 거쳐 음성을 출력한다. Discriminator 는 음성의 complex spectrogram 을 입력으로 받아 GAN[3] 학습을 한다.

하지만 Soundstream 을 그대로 학습해보면 매우 불안정하여 학습이 되지 않는다. 학습 과정을 분석한 결과, 그 이유로 두 가지를 들 수 있다. 첫 번째로, [그림 2]에서처럼 학습이 진행됨에 따라 Encoder 의 출력이 매우 커지는 현상이 나타났다. 이로 인해 학습이 불안정한 것으로 추정할 수 있으며, 특히 mixed-precision training 을 하는 경우 encoder 출력이 float 의 maximum value 인 65536 보다 커져서 무한대가 되어 더 이상 학습을 진행할 수 없었다. 이를 해결하기 위해 encoder 출력 매 frame 마다 L2-norm 을 적용하였다.

두 번째로, 모델이 trainset 에 overfitting 되는 현상이 나타났다. [그림 3]의 (a)를 보면 Discriminator 가 train set 에 대해서는 real 과 fake 를 잘 구분하는 반면, valid



[그림 2] 입력으로 음성을 주었을 때 encoder 출력 결과. 왼쪽은 initialization 직후, 오른쪽은 15 epoch 학습 후 encoder 출력. 가로축은 time frame, 세로축은 출력의 dimension 을 의미.



[그림 3] Training loss(파란색) 및 validation loss(주황색). (a)는 변경 전 모델의 discriminator loss, (b)는 변경 전 모델의 mel spectrogram loss, (c)는 변경 후 모델의 discriminator loss, (d)는 변경 후 모델의 mel spectrogram loss.

set 에 대해서는 제대로 구분하지 못하는 것을 확인할 수 있다. 또한, (b)에서 mel spectrogram loss 는 150 epoch 정도에서 minimum 달성 후 오히려 계속 증가하는 모습을 볼 수 있다. Training loss 도 증가하는 것으로 보아 discriminator 의 overfitting 이 mel spectrogram loss 의 수렴을 방해하는 것으로 추측된다. Overfitting 을 줄이기 위해 기존의 complex spectrogram 대신 spectrogram 의 magnitude 를 discriminator 의 입력으로 주었으며, discriminator 의 구조를 2D convolutional layers 에서 1D convolutional layers 로 변경하였다. 해당 변경점들을 적용했을 때, [그림 3]의 (c)에 나와있듯이 discriminator loss 는 training loss 과 validation loss 의 차이가 학습이 진행됨에 따라 사라지며, mel spectrogram loss 역시 학습이 진행됨에 따라 감소하는 것을 확인할 수 있었다.

24kHz 영어 음성 데이터셋인 LibriTTS[4]와 16kHz 한국어 음성 데이터셋에 대해서 위 변경점을 적용하여 모델을 각각 학습 후 PESQ 점수를 측정해 보았다. 최근 많이 사용되는 신호처리 기반 음성 코덱인 Opus 의 8kbps wideband 와 비교한 결과는 [표 1]과 같다. Soundstream 은 LibriTTS 에서는 7.5kbps, 한국어 데이터셋에서는 8kbps 로 인코딩 한 후 PESQ 점수를 측정하였다. 두 데이터셋 모두 Opus 코덱 대비 PESQ 점수가 우수한 것을 확인할 수 있었다.

	LibriTTS	한국어 데이터셋
Opus (8kbps)	2.74	3.05
Soundstream	2.86	3.08

[표 1] Soundstream 과 Opus 코덱의 PESQ 점수.

Ⅲ. 결론

본 논문에서는 딥러닝 기반 종단형 음성 코덱에서 기존 모델의 학습이 불안정한 이유를 분석하고, 이를 개선할 수 있는 방법을 제안하였다. 제안한 방법을 적용하여 실험한 결과 안정적으로 음성 코덱 모델을 학습할 수 있었으며, 여러 데이터셋에서 기존 신호처리 기반 음성 코덱 대비 우수한 품질을 내는 것을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2022 년도 BK21 Four 정보기술 미래인재 교육연구단에 의하여 지원되었음.

참 고 문 헌

[1] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, “Soundstream: An End-to-End Neural Audio Codec,” IEEE/ACM Trans. Audio, Speech and Lang. Proc., 2022, pp. 495-507.

[2] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, Yossi Adi, “High Fidelity Neural Audio Compression,” arXiv:2210.13438, 2022.

[3] Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville and Yoshua Bengio, “Generative Adversarial Nets,” In Advances in Neural Information Processing Systems, 2014, pp. 2672-2680.

[4] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: a corpus derived from LibriSpeech for text-to-speech,” arXiv:1904.02882, 2019.