논문 21-46-10-07 The Journal of Korean Institute of Communications and Information Sciences '21-10 Vol.46 No.10
https://doi.org/10.7840/kics.2021.46.10.1596

점진적 병렬 웨이브겐

김 형 용•, 우 범 준*, 김 남 수°

Progressive Parallel WaveGAN

Hyung Yong Kim•, Beom Jun Woo*, Nam Soo Kim°

요 약

본 논문에서는 점진적 병렬 웨이브넷(progressive parallel waveNet)을 제안하고 이를 음성 합성 보코더인 병렬 웨이브겐(parallel waveGAN) 생성기(generator)에 적용하여 성능을 확인 하였다. 음성 합성 보코더에서 높은 성능을 보이는 병렬 웨이브겐의 생성기인 병렬 웨이브넷(parallel waveNet)은 학습 시 많은 GPU(graphical processing unit)를 필요로 한다. 이를 해결하기 위해서, 병렬 웨이브넷의 구조를 음성의 업샘플링(up-sampling) 과정을 활용하여 점진적인 구조로 변형 하였다. 이는 병렬 웨이브넷에서 연산하는 특징 벡터들의 타임 축 크기 문제를 효과적으로 해결한 구조이다. 또한 음성을 넓은 주파수대역을 점진적으로 처리함으로써 학습의 안정성 및 성능을 높일 수 있는 구조이다. 실험 결과 기존 병렬 웨이브겐과 비교하여 학습 시 적은 GPU 사용량을 보였으며, 음성 생성 시 생성 속도 역시 빠른 것을 확인 할 수 있었다. 최종적으로 음성의 품질을 객관적인 지표로 측정 하였을 때, 소폭 향상된 결과 역시 확인 할 수 있었다.

Key Words : signal processing, neural network, generative model, speech synthesis vocoder, parallel waveGAN, progressive structure

ABSTRACT

In this paper, we propose a progressive parallel waveNet and apply it as a generator to parallel waveGAN, a speech synthesis vocoder, to confirm the performance of the proposed model. Parallel waveNet, a generator of parallel waveGAN showing high performance in speech synthesis vocoder, requires a lot of GPUs for training. To solve this problem, the parallel wavenet structure was transformed into a progressive structure using an up-sampling process of speech. This is a structure that effectively solves the large dimension problem of the parallel wavenet. In addition, it is a structure that can increase the stability and performance of learning by gradually estimating a wide frequency band of speech. As a result of the experiment, it was confirmed that less GPU usage on training and faster inference speed compared to parallel waveGAN, and inference speed was also faster. Finally, when speech quality was measured as an objective measurement, a slightly improved result was also confirmed.

※ 본 연구는 2021년도 정부(경찰청)의 재원 [과제명: 성문분석을 통한 실시간 화자검색 기술 개발 / 과제번호: PR01-02-040-17]으로 지원받아 수행되었습니다.

• First Author : Seoul National University Department of Electrical and Computer Engineering and Institute of New Media and Communications and Human Interface Laboratory, hykim@hi.snu.ac.kr, 학생회원

° Corresponding Author : Seoul National University Department of Electrical and Computer Engineering and Institute of New Media and Communications and Human Interface Laboratory,, nkim@snu.ac.kr, 종신회원

* Seoul National University Department of Electrical and Computer Engineering and Institute of New Media and Communications and Human Interface Laboratory, bjwoo@hi.snu.ac.kr, 학생회원

논문번호 : 202107-164-A-RU, Received July 12, 2021; Revised July 26, 2021; Accepted July 26, 2021

Ⅰ. 서 론

딥러닝의 발전으로 인해 음성 신호처리의 다양한 연구 분야들이 많은 발전을 보이고 있다. 그 중, 딥러닝 기반 음성 합성[1]은 기존 기술인 파라매트릭(parametric) 음성 합성[2]과 비교 했을 때 매우 높은 성능을 보였다. 특히, 음성 합성 보코더에 있어서 기존에 이루지 못했던 높은 성능을 보이며, 음성 생성 속도 역시 매우 빠른 성과를 보이게 되었다.

이러한 결과의 대표적인 모델은 웨이브넷(WaveNet)[3]으로 연구 결과의 발표 당시 사람의 음성과 거의 비슷한 수준의 성능을 보이며 많은 관심을 끌게 되었다. 웨이브넷은 자동 회귀(auto-regressive)한 모델로 기존의 기술들이 음성을 주파수 대역으로 변환하여 신호처리 기법 및 딥러닝 기술을 적용했던 것과는 다르게 음성을 타임 도메인에서 샘플 단위로 생성하는 모델이다. 기본 웨이브넷은 생성 모델(generative model)이기 때문에 임의의 이미지를 생성하는 모델처럼 데이터 분포와 유사한 임의의 소리를 생성하였다. 이후, 컨디셔널 정보(conditional information)로 멜 스펙트로그램(Mel spectrogram)을 활용함으로서 보코더의 기능을 할 수 있게 되었다.

하지만 이러한 웨이브넷의 가장 큰 단점은 많은 GPU 사용량과 모델의 크기가 매우 크기 때문에 학습이 어렵다는 단점이 존재한다. 또한, 실제 음성 생성 시에 하나의 샘플을 생성하고 다시 하나의 샘플을 생성하는 자동 회귀한 특성으로 인해 매우 느린 생성 속도를 보였다. 이는 실시간 처리를 필요로 하는 음성 합성 문제 상황에서 매우 치명적인 단점이기 때문에 사용이 불가능 하였다.

이러한 단점을 해결하기 위해, 비자동 회귀(non-autoregressive) 한 합성 보코더 모델들이[4,5]제안되었다. 그 중 가장 유명한 모델은 병렬 웨이브넷[4]과 Clarinet[5]이 있다. 병렬 웨이브넷과 Clarinet은 사전에 학습한 선생 모델(teacher model)인 자동 회귀 한 모델을 플로우 기반(Flow-based) 학생 모델(student model)에게 지식 증류법(knowledge distillation)을 통해 학습 시키는 모델이다. 이 두 모델은 비자동 회귀한 특성을 갖기 때문에 음성 생성 시에도 매우 빠른 속도를 보이며 성능 또한 웨이브넷과 비슷한 수준까지 보고가 되었기 때문에 많은 관심을 받게 되었다. 하지만 이 두 모델의 단점은 잘 학습 시킨 자동 회귀한 한 선생 모델이 있어야하며 이를 지식 증류법을 통해 학습 시키는 일도 어렵다는 단점이 존재 했다.

그 이후로 생성 모델 중 하나인 글로우(GLOW)[6] 기반의 음성 생성 모델인 웨이브글로우(WaveGLOW)[7]가 제안되었다. 이 웨이브글로우는 생성 모델인 플로우(FLOW)[8] 모델을 발전시킨 글로우 모델을 기반으로 한다. 플로우는 매우 단순한 확률 분포에서 연속적인 역변환이 가능한 함수를 통해 매우 복잡한 확률 분포로의 변환하는 과정을 하게 된다. 역변환이 가능한 연속적인 함수들 덕분에 매우 복잡한 확률 분포에서 이를 다시 단순한 확률 분포로 변환하는 것이 가능하며 이를 통해 데이터베이스로부터 알고자하는 확률 분포인 음성 데이터의 분포를 학습할 수 있게 된다. 하지만 이러한 웨이브 글로우 역시도 높은 성능을 위해선 학습 시 큰 배치사이를 필요로 하며 낮은 배치 사이즈의 경우 최적성능에 도달 하지 못하게 된다.

이러한 플로우 기반의 음성 보코더 모델과 달리 최근 생성적 적대 모델(generative adversarial network) 기반의 음성 합성 보코더들은[9,10] 빠른 학습 속도와 높은 성능을 통해 주류의 모델이 되어가고 있다. 이러한 모델 중 병렬 웨이브겐(Parallel WaveGAN)[9]은 빠른 학습과 생성속도로 많은 관심을 받았다. 병렬 웨이브겐의 구조는 지식 증류법 없이 병렬 웨이브넷의 구조만을 생성기로 사용하였다. 분류기(discriminator)로는 MelGAN[10]에서 제안한 다중 해상도(multi-scale) 분류기를 사용하였다. 특징으로는 생성기 손실함수에 있어서 타임 도메인에서 단순하게 L_1 손실함수를 사용한 것이 아닌 다중 해상도 STFT (short-time fourier transform) 손실함수를 사용하였고 뛰어난 성능을 확보하였다.

병렬 웨이브겐이 사용하는 생성기인 병렬 웨이브넷(Parallel WaveNet)[11]은 지식 증류법을 이용하지 않고 학습을 한다고 하더라도 학습 시 많은 양의 GPU를 사용 할 수밖에 없는데, 이는 병렬 웨이브넷의 구조적인 문제에서 발생하는 문제이다. 병렬 웨이브넷은 입력과 출력 그리고 중간의 특징 벡터들 모두 매우 큰 타임 차원을 갖는 구조이다. 문제는 타임 도메인에서 긴 샘플을 유지한 상태로 특징 벡터들을 컨볼루션(convolution) 연산을 하게 되면 학습 시 많은 양의 GPU를 사용할 수밖에 없는 것이다.

최근에 많은 생성적 적대 모델 기반 이미지 생성 분야에서 점진적인 구조에 대한 연구[12-14]들이 많이이루어졌다. 특히 progressive GAN[12]과 style GAN 1,2[13.14]은 높은 해상도의 이미지를 한 번에 추정하는 것이 아닌 낮은 해상도에서부터 시작하여 점차 높은 차원의 이미지를 추정하는 방식을 제안하였고 높은

성능을 보였다. 음성 향상 분야에서도 이러한 접근 방법[15]이 사용 되었다.

본 논문에서는 병렬 웨이브겐의 생성기인 병렬 웨이브넷의 학습 시 GPU 사용량 문제를 해결하기 위해 점진적 병렬 웨이브넷을 제안하였다. 점진적 병렬 웨이브넷은 음성을 생성할 때 음성의 넓은 주파수 대역을 한 번에 생성 하는 것이 아니라 음성의 업샘플링 과정을 활용한 모델 설계를 통해 점진적으로 음성을 생성한다. 이 모델을 통해 학습 시에 GPU사용량을 대폭 낮추었으며 음성 생성 시 기존 모델과 비교 하였을 때 빠른 생성 속도를 가능하게 했다.

Ⅱ. 본 론

2.1 병렬 웨이브겐

본 항에서는 비교 대상이 되는 병렬 웨이브겐에 대한 구조와 학습 방법에 대해서 설명한다.

2.1.1 병렬 웨이브겐의 구조

병렬 웨이브겐은 병렬 웨이브넷을 기반으로 한 생성기와 다중 해상도 분류기인 MelGAN 분류기로 구성이 되어 있다. 두 모델은 일반적인 생성적 적대 모델 학습 방법에 따라 학습이 된다.

병렬 웨이브겐은 생성적 적대 학습 방법을 통해 학습이 되는데, 이 학습 방법은 비지도 학습 중 하나의 학습 방법이다. 우선 분류기는 학습 시 실제 샘플과 가짜 샘플을 잘 구분 하도록 학습 하게 된다. 여기서 가짜 샘플은 생성기가 생성한 샘플인데, 생성기의 학습은 분류기가 진짜 샘플과 구분 할 수 없도록 가짜 샘플을 만드는 방향으로 학습을 하게 된다. 이러한 두 개의 모델이 서로 적대적으로 학습을 통해 생성기가 만들어내는 샘플의 확률 분포가 실제 데이터의 확률 분포와 같아지게 만든다.

이러한 생성적 적대 모델은 컨디셔널 한 정보 없이 실제 데이터들을 생성하기 때문에 본 논문의 목적인 음성에서 추출한 멜 스펙트로그램으로부터 음성을 복원하는 보코더에는 적용이 불가능하다. 따라서 컨디셔널 생성적 적대 학습을 통해서 이를 해결한다. 이는 생성기와 분류기에 컨디셔널 정보로 멜 스펙트로그램을 입력한 후 음성을 복원한 방법이다. 그림 1은 제안하는 모델의 구조도이다. 본 논문에서 제안 하는 모델은 큰 구조에서는 병렬 웨이브겐과 동일한 구조를 갖기 때문에 이를 활용하여 병렬 웨이브겐의 학습 방법을 살펴보면, 생성기의 입력으로는 랜덤 벡터를 넣어주게 된다. 이는 일반적으로 가우시안 확률 분포를 주로 사용하게 된다. 또 다른 입력인 컨디셔널 정보는 타겟(target)이 되는 음성으로부터 추출한 멜 스펙트로그램을 컨볼루션 연산을 한 후에 랜덤 백터와 차원을 맞추기 위해 업샘플링한 벡터를 넣어주게 된다. 업샘플링 블록은 연속적인 최단입점 보간(nearest upsampling) 방법과 컨볼루션 연산을 통해 차원을 키우는 구조로 되어있다. 병렬 웨이브겐에서 사용한 생성기의 구조는 병렬 웨이브넷에서 구조만을 가져와 사용하였다.

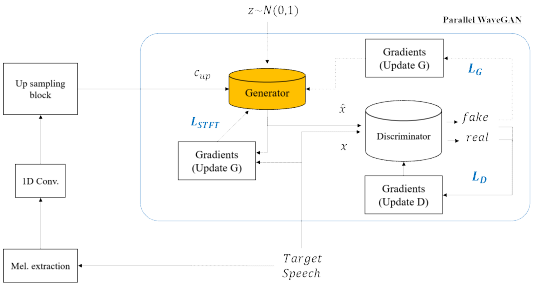

그림 1. 제안하는 모델의 전체 구조
Fig. 1. Proposed model architecture

2.1.1.1 병렬 웨이브넷

병렬 웨이브넷은 자동 회귀하게 음성의 샘플을 생성하는 웨이브넷과 같은 구조를 갖고 있다. 다만, 병렬 웨이브넷은 기본 웨이브넷과 달리 비자동 회귀 하게 컨볼루션 연산을 병렬적으로 처리하는 구조를 갖고 있다.

그림 2.는 병렬적 웨이브넷의 구체적 블록도 이다. 병렬적 웨이브넷은 동일한 연산을 수행하는 웨이브넷 블록을 30번 쌓은 구조이다. 동일한 연산에서 세부적으로 달라지는 부분은 dilation 부분으로 매 레이어마다 2의 배수만큼 dilation이 늘어나게 된다. 이러한 dilation은 총 10개의 사이클(cycle)을 갖게 되고, 이러한 사이클을 총 3번 반복하는 구조이다.

병렬적 웨이브겐에서 생성기로 활용한 병렬적 웨이브넷의 입력은 크게 2가지이다. 하나는, 출력과 동일한 크기를 갖는 주로 가우시안에서 추출한 랜덤 노이즈를 입력으로 넣어주게 된다. 다른 하나는, 음성에서 추출한 멜을 출력과 동일한 크기로 맞춰주기 위해 업샘플링 블록을 통과한 임베딩 벡터를 컨디셔널 정보로 입력 받는다. 이후, 입력들은 다양한 컨볼루션을 통과하게 되고 게이티드 활성화 유닛에서 목적함수를 줄일 수 있는 방향으로 새로운 특징 벡터를 추출 하게 된다. 이후, 출력된 벡터는 2개의 경로로 갈리게 되는데 하나는 레지듀얼 커넥션과 더해진 벡터를 새로운

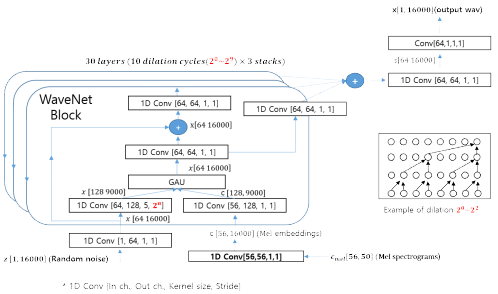

그림 2. 병렬 웨이브넷
Fig. 2. Parallel waveNet

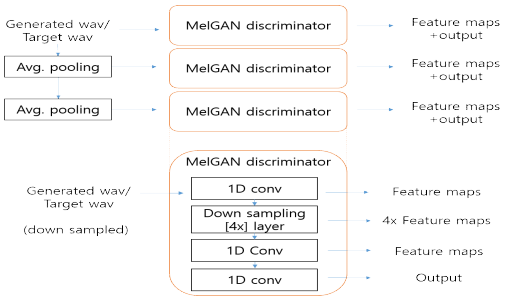

그림 3. 다중 해상도 분류기
Fig. 3. Multi-resolution discriminator

컨볼루션을 통과한 후 다음 레이어의 입력으로 되먹임 시킨다. 또 다른 하나는 컨볼루션을 통과한 후 출력부로 내보내 지게 되는데, 모든 레이어의 출력 벡터들은 모두 더해진 후에 2개의 추가적인 컨볼루션을 통과한 이후에 최종적인 음성으로 출력 된다.

병렬적 웨이브넷의 특징은 dilation이 늘어남에 따라 컨볼루션이 확보하는 receptive field가 늘어나는 것이다. 병렬적 웨이브넷은 기존의 모델들이 음성을 프레임(frame) 단위에서 주파수 정보를 처리하는 것과 다르게, 음성을 타임 도메인에서 샘플 단위로 처리하는 구조이기 때문에 receptive filed의 확보가 매우 중요하다. 웨이브넷 논문에서는 dilation을 통해서 이러한 문제를 해결하였다.

2.1.1.2 다중 해상도 분류기

병렬 웨이브겐 에서는 분류기로 MelGAN에서 제안한 다중해상도 분류기를 사용하였다. MelGAN 분류기는 다중해상도 분류기 구조를 가지고 있다. 일반적인 분류기와 다르게 다중 해상도 분류기는 여러 개의 종속된 분류기(sub discriminator)로 구성이 되어있다. 이 분류기들은 서로 다른 해상도에서 실제와 가짜 샘플을 구분 한다.

그림 3은 다중 해상도 분류기의 블록도 이다. 총 3개의 분류기로 구성이 되어있으며, 상위 레벨의 분류기는 생성된 음성을 구분하고 하위 레벨로 갈수록 평균 풀링(pooling)을 통해서 다운 샘플링 된 음성을 구분하게 된다. 이러한 구조를 통해 음성을 다중 해상도로 적대적 학습 시키게 되면, 목표가 되는 음성의 데이터 분포를 더 정확하게 학습 시킬 수 있다. 또한, MelGAN 분류기에서는 중간의 특징 벡터의 값 역시도 같아질 수 있도록 L_1 손실함수를 주는 구조를 통해 높은 성능을 확보 하였다. 본 논문에서 제안하는 모델 역시 MelGAN에서 사용한 다중 해상도 분류기를 사용한다. 본 논문에서 분류기를 위해 사용한 손실함수는 다음과 같다.

$$L_{D_k}(G, D_k) = E_{x \sim P_{data}}[(1 - D_k(x))^2] + E_{z \sim N(0,I)}[D_k(G(z))^2], \forall k = 1,2,3 \quad (1)$$

$$L_{FM}(G, D_k) = E_{x \sim P_{data}}[\sum_{i=1}^{T} \frac{1}{N} \| D_k^{(i)}(x) - D_k^{(i)}(G(\hat{x})) \|_1] \quad (2)$$

k 는 그림 3에서 3개의 분류기 들의 번호이며, 나머지는 병렬 웨이브겐[9]과 동일하다.

2.2 점진적 병렬 웨이브겐

본 항에서는 본 논문에서 제안한 점진적 병렬 웨이브겐에 대해서 설명한다. 점진적 병렬 웨이브겐은 기존 병렬 웨이브겐의 생성기인 병렬 웨이브넷을 개선한 점진적 웨이브넷의 구조를 제안하였다. 분류기는 앞에서 설명한바와 같이 MelGAN에서 제안한 다중 해상도 분류기를 사용하였다.

2.2.1 점진적 병렬 웨이브넷

그림 2에서 설명한 병렬적 웨이브넷은 음성을 샘플(sample) 단위로 처리하며 구조적인 특징들로 인해 타임 축으로 매우 큰 크기의 피쳐 벡터들을 처리하게 된다. 예를 들면, 연속적인 웨이브넷 블록들은 입력과 출력 모두 사전에 정의한 음성 길이의 타임 샘플의 크기를 유지해야 한다. 이러한 문제로 인해서 학습 시 백프로파게이션(back propagation)을 위해 매우 많은 GPU 메모리가 필요로 하다. 때문에, 다른 음성 분야에서 사용한 배치 사이즈와 달리 굉장히 작게 배치사이즈를 구성 할 수밖에 없으며 이러한 문제로 학습의 불안정성과 최적 성능 미달 및 속도 저하가 발생 하게 된다.

본 논문에서 제안하는 점진적 병렬 웨이브넷은 음성의 업샘플링의 처리과정을 모델을 구성하는데 활용하였다. 음성은 샘플링 레이트 마다 포함 하고 있는 주파수 대역이 다르며 높은 샘플링 레이트의 음성일수록 넓은 주파수 대역의 광대한 정보를 포함 하고 있다. 따라서 점진적 음성 생성 방식은 높은 샘플링 레이트의 음성을 생성하는데 효과적인 방법이다.

그림 4는 점진적 병렬 웨이브넷에 대한 블록도 이다.본 논문에서는 16kHz 음성을 기준으로 설명을 하겠지만, 원하는 샘플링 레이트의 음성에서 적용이 가능하다. 점진적 병렬 웨이브넷의 구조를 보면 낮은 레이어에서는 4kHz 음성을 먼저 생성하고, 2배의 크기를 갖는 8kHz 그리고 최종적으로 16kHz 음성을 생성하는 구조이다. 이를 통해서 높은 샘플링 레이트 음성의 넓은 주파수 대역을 점진적으로 생성 할 수 있는 모델이다.

입력으로는 최종 출력 음성에 비해서 타임 축의 크기가 4배 줄어든 랜덤 잡음 샘플이 입력되게 된다. 컨디셔널 정보로는 음성으로부터 멜을 추출하고 이 멜을 업샘플링 블록을 통해서 해당 레이어의 랜덤 잡음 입력 샘플과 같은 크기를 갖도록 만들어준다. 이렇게 들어간 입력은 기존의 병렬 웨이브넷에서 사용한 동일한 레이어들을 통과하게 된다.

기존 병렬 웨이브넷과 마찬가지로 2개의 경로 중 레지듀얼 커넥션과 더해진 후 되먹임이 되는 출력 벡터는 전치 컨볼루션(transposed convoluton)을 통해 크기를 2배로 늘리고 상위 레이어의 입력으로 들어가게 된다. 이렇게 입력된 벡터는 하위 레이어 에서와 마찬가지로 동일한 웨이브넷 블록을 통과하게 되어 최종적으로 목표로 하는 음성과 동일한 크기를 가질 때 까지 반복적으로 출력 벡터를 생성하게 된다.

두 번째 경로는 출력 쪽으로 나오게 되는데 10개의

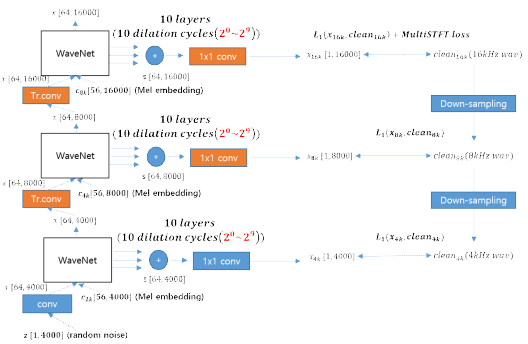

그림 4. 점진적 병렬 웨이브넷
Fig. 4. Progressive parallel waveNet

레이어 출력을 더한 이후에 해당 레이어에서 다루는 샘플링 레이트에 맞는 크기의 음성 출력을 내보내게 된다. 이렇게 출력된 음성의 크기는 최종 목표로 하는 음성의 다운 샘플링한 음성과 같게 된다. 점진적으로 음성의 주파수 대역을 추정하기 위해 각 레이어에서 다운 샘플링(down sampling)된 목표 음성과 L_1 손실함수를 목적함수로 사용하게 된다. 이렇게 추가된 손실함수와 기존의 병렬적 웨이브겐에서 생성기에 적용된 손실함수는 다음과 같다.

$$L_G(G,D) = L_{progress}(G) + L_{aux}(G) + \lambda_{adv} L_{adv}(G,D) \tag{3}$$

$$L_{progress} = \sum_{n \in N_p} \| x_n - \hat{x_n} \|_1, N_p \in 4k, 8k, 16k \tag{4}$$

$$L_s(G) = E_{z \sim N(0,I), x \sim p_{data}}[L_{sc}(x,\hat{x}) + L_{mag}(x,\hat{x})] \tag{5}$$

$$L_{sc}(x,\hat{x}) = \frac{\| |STFT(x)| - |STFT(\hat{x})| \|_F}{\| |STFT(x)| \|_F} \tag{6}$$

$$L_{mag}(x,\hat{x}) = \frac{1}{N} \| \log|STFT(x)| - \log|STFT(\hat{x})| \|_1 \tag{7}$$

$$L_{aux}(G) = \frac{1}{M} \sum_{m=1}^{M} L_s^{(m)}(G) \tag{8}$$

$$L_{adv}(G,D) = E_{z \sim N(0,I)}[(1 - D(G(z)))^2] \tag{9}$$

λ_{adv}는 적대적 손실함수의 가중치이며, N_p는 점진적 구조에서 사용한 샘플링레이트 들이다. $N(0,I)$는 노말 확률분포(normal distribution)를 의미하고, p_{data}는 학습 데이터베이스에 음성의 확률분포이다. $\| \|_F$와 $\| \|_1$은 프로베니우스 놈(Frobenius norm)과 L_1 놈(norm)이다. $STFT$는 짧은 시간 푸리에 변환을 의미하고, m은 실험 항에 있는 표 1에 정의되어있는 변수들에 맞게 변환하게 된다.

Ⅲ. 실 험

3.1 실험 환경

본 논문에서는 TIMIT 16kHz 데이터베이스를 사용하였다. 학습은 TIMIT 훈련 세트를 사용하였고 결과

측정은 TIMIT 시험 세트를 활용하였다. TIMIT 데이터베이스는 462명의 화자가 2개의 동일한 문장과 8개의 서로 다른 문장을 발화한 훈련 세트와 훈련 세트에 포함 되지 않은 162명의 화자가 2개의 동일한 문장과 8개의 서로 다른 문장을 발화한 시험 세트로 구성된 영어 데이터베이스이다. 본 논문에서는 다양한 화자가 포함되어있는 실험 및 시험 세트를 구성하기위해 TIMIT 데이터 베이스를 사용하였다.

학습에 사용한 문장의 길이는 16kHz 기준 1초에 해당하는 16000 샘플을 사용하였다. 멜 스펙트로그램은 80-7600Hz 제한된 대역에서 56차의 크기를 갖도록 추출하였고, 윈도우 크기와 FFT(fast fourier transform)사이즈는 400을 사용하였고 프레임 이동 크기는 320을 사용하였다.

3.2 모델 구조

모델의 구조는 병렬 웨이브겐을 기본으로 한다. 먼저 비교 대상이 되는 병렬 웨이브겐의 생성기 대해서 설명하면 dilation을 제외한 나머지의 동일한 연산을 하는 웨이브넷 블록을 총 30개 쌓은 구조이다. 설명을 위해서 [입력채널, 출력채널, 커널크기, stride, dilation]의 형태로 컨볼루션 연산의 파라매터(parameter)를 설명한다. 또한 stride와 dilation의 경우 별다른 설명이 없는 경우 1로 가정한다.

첫 번째로 그림 1에서 멜을 입력으로 받는 컨볼루션은 1D 컨볼루션으로 [56,56,5,1,1]을 갖는다. 업샘플링 블록은 총 4번에 나눠서 멜 스펙트로그램의 차원을 학습 시 정의한 16000 샘플과 같은 차원으로 맞춰주게 되는데 10, 8, 2, 2배로 순으로 업샘플링 하였다. 업샘플링 방법은 최단입점 보간법으로 차원을 늘린 이후에 1x1 컨볼루션을 통과 시켰다. 이때 커널의 사이즈는 21, 17, 5, 5의 순으로 감소 시켰다.

그 다음으로 생성기인 그림 2의 모델 파라메터는 그림에 표기된 바와 같이 진행을 하였다. 그림 3의 경우는 그림 2의 파라메터를 동일하게 따라갔으며 전치 컨볼루션의 파라메터는 [64,64,31,2,1]로 입력 특징벡터의 차원을 늘려주었다. 마지막으로 분류기는 병렬 웨이브겐에서 사용한 MelGAN 분류기와 동일한 파라메터를 사용하여 실험을 진행 하였다. 생성기학습 시 다중해상도 STFT 손실함수의 설정 값은 병렬 웨이브겐에서 사용한 값을 사용하였으며 이는 표 1 에 나타나 있다.

표 1. 학습 시 다중 해상도 STFT 손실함수의 설정 값
Table 1. Parameters of STFT loss

STFT lossl	FFT size	Window size	Frame shift
$L_s^{(1)}$	1024	600	120
$L_s^{(2)}$	2048	1200	240
$L_s^{(3)}$	512	240	50

3.3 실험 결과 및 평가

본 논문에서 측정 하고자 하는 지표는 크게 3가지이다. 첫 번째로는 학습 과정에서 필요한 메모리와 학습 시간을 측정하여 대상 모델과 비교한다. 두 번째로는 학습 모델의 파라메터와 음성 생성 속도를 비교한다. 마지막으로 대상 모델과 객관적 음성 품질 지표를 측정한다.

3.3.1 학습 과정에서 필요한 메모리와 학습 시간

학습 과정에서 필요한 메모리를 측정하고자 동일한 배치와 문장 길이 상황에서 사용 메모리를 측정하였고, 사용한 그래픽 카드는 TITAN RTX 24GB GPU로 실험을 진행 하였다.

표 2에 볼 수 있듯이 본 논문에서 제안한 모델이 GPU 사용량에 있어서 기존 병렬 웨이브넷 대비 36% 감소시키는 것을 확인 할 수 있었다. 이는 점진적 구조로 인해 특징 벡터의 타임 축 크기를 줄여서 학습을 할 수 있기 때문이다. 또한, 멀티 GPU 및 더 긴 음성을 사용하게 될 경우 이 감소량은 더 커질 수 있다.

3.3.2 학습 모델의 파라메터와 음성 생성 속도

학습이 아닌 실제 음성을 생성 할 때, 본 논문에서 제안하는 모델이 기존 모델 대비 얼마나 빠르게 생성이 가능한지를 체크하기 위해 RTF(real-time factor)를 측정 하였다. RTF의 계산 수식은 다음과 같다.

$$RTF = \frac{\text{문장 생성 시간}(s)}{\text{문장의 길이}(s)} \tag{10}$$

RTF 측정은 TIMIT 시험 세트의 1620문장에 대해

표 2. 학습 시 모델별 메모리 사용량 비교
Tabel 2. Comparison of memory usuage

모델	배치 크기	Speech length (sample)	mel length	memory cached (GB)
병렬 웨이브겐	20	16000	54	21.6
점진적 병렬 웨이브겐				13.9

표 3. 음성 생성 시 모델 별 속도 비교
Table 3. Comparison of inference speed

모델	모델 크기 (MB)	CPU specification		RTF
		Name	Core	
병렬 웨이브겐	1.73	Intel(R) Silver 4214 2.2GHz	8	0.937
점진적 병렬 웨이브겐	1.99			0.591

표 4. 모델 별 객관적 음성 품질 평가
Table 4. Comparison of objctive speech quality

모델	PESQ	STOI
병렬 웨이브겐	3.42	0.95
점진적 병렬 웨이브겐	3.45	0.95

서 평균 RTF를 계산하였고, 실험은 8코어 CPU에서 측정을 하였으며 자세한 CPU 설정은 표 3과 같다. 표 3에서 볼 수 있듯이 음성 생성 속도가 약 37% 정도로 빨라진 것을 확인 할 수 있었다. 이는 기본 모델이 16000 샘플을 모두 컨볼루션 연산을 해야 하지만, 본 논문에서 제안한 모델은 8000과 4000으로 줄어든 샘플을 연산 하면 되기 때문에 상대적으로 빠르게 음성 생성이 가능하게 된 것이다.

또한, 그림 4에서 볼 수 있듯이 기존 생성기에서 추가된 파라매터는 주황색에서 모델에서 사용된 파라매터들이다. 표 3에서 볼 수 있듯이 기존 생성기의 파라매터는 1.72MB 이며, 제안한 생성기는 1.99MB로 매우 적은 파라매터 증가가 있었다. 분류기는 동일한 파라매터를 갖는다.

3.3.3 객관적 음성 품질 지표

보코더의 음성 품질을 측정하기 위하여 객관적 음질 평가를 시도하였다. 음성 보코더의 성능은 보통 주관적 평가를 실시 하지만, 해당 논문의 목표는 계산량 및 메모리 사용량이 목적이기 때문에 주관적 평가를 실시하지 않았다. 객관적 지표는 PESQ (perceptual evaluation speech quality)[16]와 STOI(short-time objective intelligibility)[17]를 사용하였다.

표 3에서 볼 수 있듯이 본 논문에서 제안한 모델이 기존 모델과 비교 했을 때 PESQ 수치로 약 0.03 향상 된 것을 확인 할 수 있었다. 또한 STOI는 같은 수치를 나타냈다. 실제로 청음을 했을 때, 두 모델의 결과는 많은 차이가 느껴지지 않았다. 이는, 약간의 모델 크기 증가와 비해 성능 저하가 거의 없었으며 오히려 객관적 지표에서는 더 높은 성능을 보인 것을 확인 하였다.

Ⅳ. 결 론

본 논문에서는 음성 합성 보코더의 모델 중 병렬 웨이브겐의 생성기인 병렬 웨이브넷을 개선한 점진적 웨이브넷을 제안하였다. 음성의 업샘플링 처리과정을 적용하여 웨이브넷을 점진적 웨이브넷으로 개선 시켰다. 이 모델을 통해 기존의 병렬 웨이브넷이 가지고 있는 학습 GPU 사용량 문제를 해결 하였으며, 실제 음성 생성에 있어서도 빠른 생성속도를 확인 하였다. 이러한 모델의 변화에도 불구하고 음성 품질의 객관적 지표에 있어서 약간의 개선이 있었다. 이러한 점진적 웨이브넷은 병렬 웨이브겐 뿐만 아니라 병렬 웨이브넷이 사용 되는 다양한 모델 구조에서 사용이 가능하다.

References

[1] J. Y. Lee, S. J. Cheon, B. J. Choi, N. S. Kim, and D. H. Hong, "Speech style modeling method using mutual information for end-to-end speech synthesis," *J. KICS*, vol. 44, no. 9, pp. 1641-1647, 2019.

[2] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proc. ICASSP*, pp. 1229-1232, 2007.

[3] A. V. D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

[4] S. Kim, S. Lee, J. Song, J. Kim, and S. Yoon, "FloWaveNet : A generative flow for raw audio," in *Proc. ICML*, pp. 3370-3378, 2019.

[5] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICLR*, 2019.

[6] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," arXiv preprint arXiv:1807.03039, 2018.

[7] R. Prenger, R. Valle, and B. Catanzaro,

"Waveglow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, pp. 3617-3621, 2019.

[8] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," arXiv preprint arXiv:1505.05770, 2015.

[9] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, pp. 6199-6203, 2020.

[10] K. Kumar, R. Kumar, T. Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. Brebisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. NeurIPS*, pp. 14881-14892, 2019.

[11] A. V. D. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. V. D. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. ICML*, pp. 3915-3923, 2018.

[12] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. ICLR*, pp. 1-12, 2018.

[13] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. CVPR*, pp. 4401-4410, 2019.

[14] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," arXiv preprint arXiv:1912.04958, 2019.

[15] H. Y. Kim, J. W Yoon, S. J. Cheon, W. H. Kang, and N. S. Kim, "A multi-resolution approach to GAN-Based speech enhancement," *Appl. Sci.*, vol. 11, no. 2, 2021.

[16] ITU-T, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs,* Rec. ITU-T P. 862; 2000. Available online: https://www.itu.int/rec/T-REC-P.862.

[17] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 11, pp. 2009-2022, 2016.

김 형 용 (Hyung Yong Kim)

2013년 8월 : 광운대학교 전기공학과 학사 졸업

2014년 3월~현재 : 서울대학교 전기정보공학부 석박사통합과정 박사과정

<관심분야> 음성 신호 처리, 음성 향상, 뉴럴 네트워크

[ORCID:0000-0001-6009-9530]

우 범 준 (Beom Jun Woo)

2019년 2월 : 카이스트 전기전자 공학부 졸업

2019년 3월~현재 : 서울대학교 전기정보공학부 석박사통합과정 박사과정

<관심분야> 음성 신호 처리, 음성 분리, 뉴럴 네트워크

[ORCID:0000-0002-2940-0134]

김 남 수 (Nam Soo Kim)

1988년 : 서울대학교 전자공학과 학사 졸업

1990년 : 한국과학기술원 전기 및 전자공학과 석사 졸업

1994년 : 한국과학기술원 전기 및 전자공학과 박사 졸업

1993년~1998년 : 삼성종합기술원 전문연구원

1998년~현재 : 서울대학교 전기정보공학부 교수

<관심분야> 음성 신호 처리, 음성 인식, 통계적 신호처리, 패턴 인식, 휴먼 인터페이스

[ORCID:0000-0002-0568-4902]