

WAVENET : A GENERATIVE MODEL FOR RAW AUDIO

[https://arxiv.org/pdf/1609.03499](https://arxiv.org/pdf/1609.03499.pdf)

0. Introduction

- 기존 음성 생성 방식은 스펙트로그램 변환 등 간접 표현을 사용해 원시 신호의 미세 구조를 완전히 복원하기 어려웠음.
- 원시 오디오(raw audio)는 매우 높은 샘플링 레이트를 가지므로 긴 시퀀스 의존성과 복잡한 패턴을 동시에 모델링해야 하는 어려움이 존재함.
- 이 연구는 raw audio를 직접 확률적으로 생성하는 딥러닝 모델 Wavenet을 제안하며, 자연스러운 음성·음악 합성을 목표로 함.
- 핵심 기여는 dilated causal convolution, autoregressive audio modeling, 고품질 음성 합성 성능, 조건부 생성 구조로 요약됨.

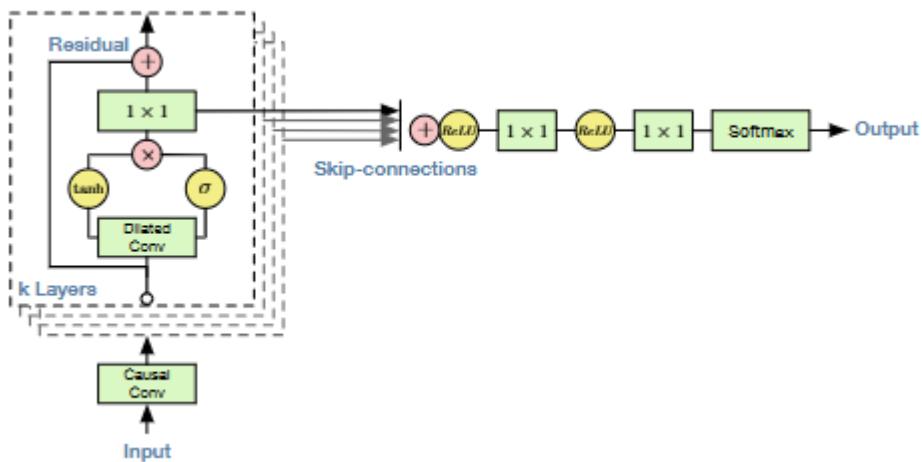
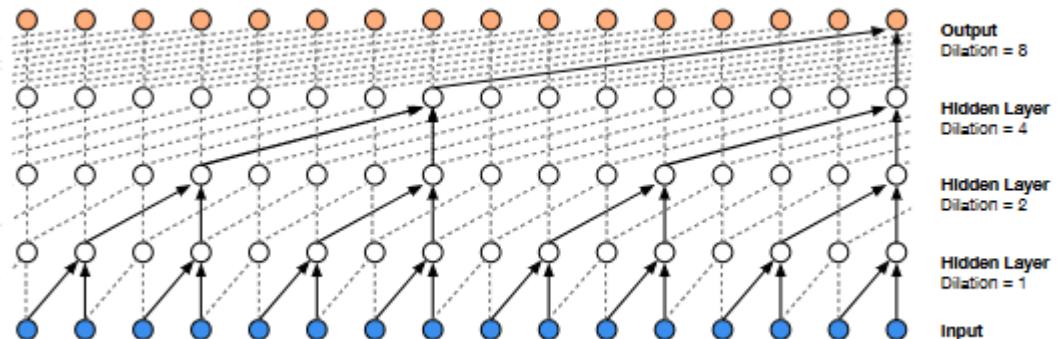
1. Overview

- Wavenet은 각 오디오 샘플을 이전 모든 샘플에 조건부로 예측하는 autoregressive 생성 모델임.
- 모델은 RNN이 아닌 causal convolution을 기반으로 하며, 미래 정보를 사용하지 않는 구조를 유지함.
- receptive field를 크게 확장하기 위해 dilated convolution을 사용해 긴 의존성을 효율적으로 학습함.
- 입력 오디오는 8-bit μ -law quantization을 통해 discrete token으로 변환됨.
- 모델은 음성 합성, 음악 생성, 스피커 변환 등 다양한 생성 태스크에 적용 가능함.

2. Challenges

- raw audio는 단일 초당 수천 개 이상의 샘플을 포함하므로 길이가 매우 길고 복잡한 시퀀스를 다뤄야 함.
- 긴 시퀀스에서 장기·단기 패턴을 동시에 모델링하기 위한 구조가 필요함.
- RNN 기반 접근법은 계산 비용이 크고 병렬화가 어려워 대규모 오디오 생성에 적합하지 않음.
- 오디오 신호는 주기적 패턴과 비주기적 잡음이 혼재되어 있어 확률적 모델링 난도가 높음.
- 조건부 생성(스피커/언어/텍스트)을 안정적으로 통합하는 구조 설계가 필요함.

3. Method



- Autoregressive formulation
 - $p(x) = \prod_t p(x_t | x_1 \dots x_{t-1})$ 형태로 전체 waveform을 모델링함.

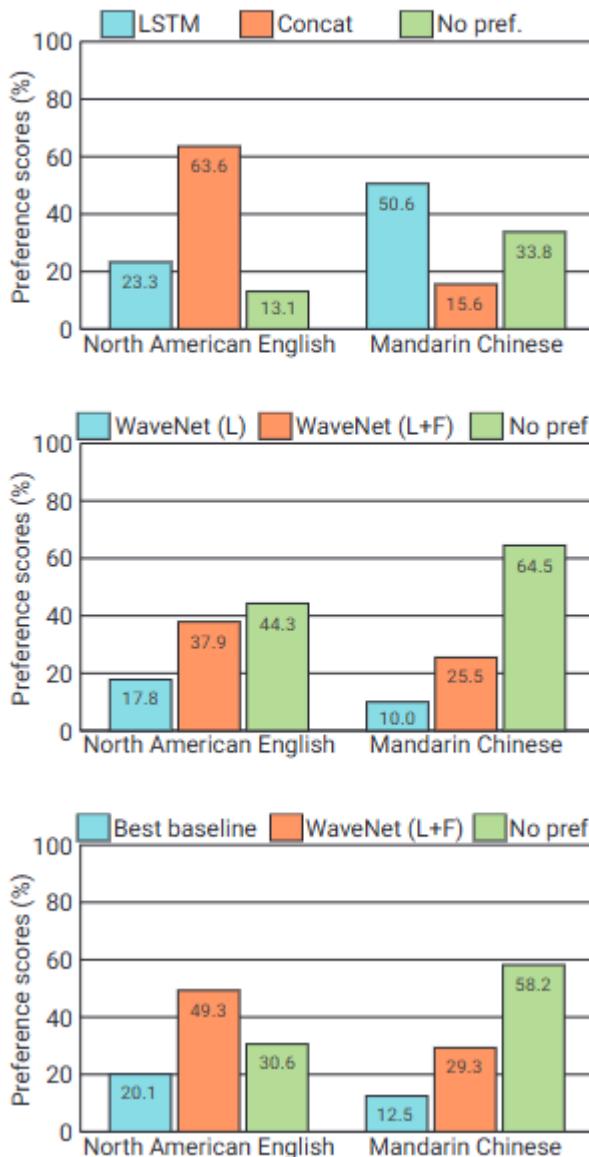
- μ -law quantization(8-bit, 256-level)으로 discrete modeling을 수행함.
- Causal convolution
 - 합성 시 미래 timestamp 정보 접근을 차단하여 causal structure를 보장함.
 - RNN 없이도 시간 순서를 유지하는 구조 제공.
- Dilated convolution
 - dilation factor를 1, 2, 4 ... 형태로 증가시키며 receptive field를 지수적으로 확장함.
 - 깊은 네트워크 없이도 넓은 temporal context 학습 가능.
- Gated activation unit
 - tanh과 sigmoid를 조합한 게이트 구조를 적용하여 richer nonlinear modeling을 수행함.
- Residual & skip connections
 - 학습 안정성 향상과 deeper stack 구성 지원.
 - skip connection은 최종 softmax layer로 직접 연결됨.
- Conditional modeling
 - global conditioning: 스피커 ID 등 전체 시퀀스에 동일한 조건을 투입
 - local conditioning: 텍스트 음성 변환(TTS)에서 phoneme sequence 등을 time-aligned로 투입함.

4. Experiments

- 데이터는 다수의 음성 corpus와 music dataset을 사용해 학습함.
- baseline 대비 raw waveform 직접 생성의 성능을 평가함.
- TTS 실험에서는 local conditioning을 적용해 텍스트 기반 음성 생성 품질을 측정함.
- 확률 모형의 정합성 확인을 위해 likelihood 기반 평가를 수행함.
- 오디오 샘플은 청취 기반 평가(mean opinion score 형태)로 품질을 비교함.

5. Results

Speech samples	Subjective 5-scale MOS in naturalness	
	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	4.21 ± 0.081	4.08 ± 0.085
Natural (8-bit μ -law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071



- Wavenet은 기존 합성 시스템(예: parametric TTS, concatenative TTS) 대비 훨씬 자연스러운 음성 품질을 보임.

- 실제 사람 음성에 가까운 타이밍·억양·잡음 패턴을 재현함.
- 음악 생성에서도 반복 패턴과 비주기적 변화가 공존하는 구조를 비교적 잘 학습함.
- 조건부 모델은 스피커 ID를 정확히 반영하며 다중 화자 음성 합성에서 높은 품질을 달성 함.
- dilated convolution 구조는 긴 시퀀스에서도 안정적인 modeling 성능을 보여줌.
- 합성 속도는 autoregressive 특성 때문에 느린 편이라는 한계가 존재함.

6. Insight

- 원시 오디오의 확률적 모델링이 가능해졌다는 점에서 음성 합성 패러다임의 전환을 이 끈 연구임.
- dilated convolution은 긴 시퀀스를 효율적으로 처리하는 대안 구조로 이후 다양한 분야(CV, NLP 등)에 확장됨.
- 조건부 구조는 TTS·음성 변환·악기 스타일 변경 등 실무적 활용 범위를 크게 확장함.
- autoregressive 구조로 인해 실시간 합성이 어렵다는 문제는 후속 연구(Parallel WaveNet 등)의 출발점이 됨.