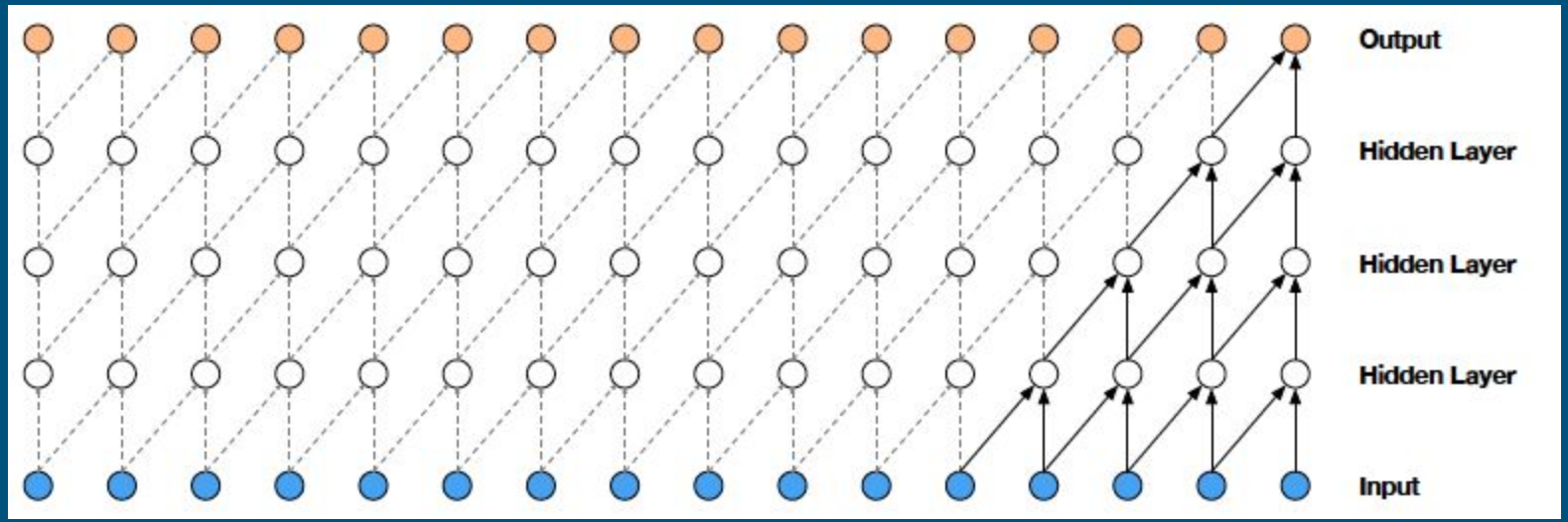
WaveNet 논문 리뷰

이희종

Introduction

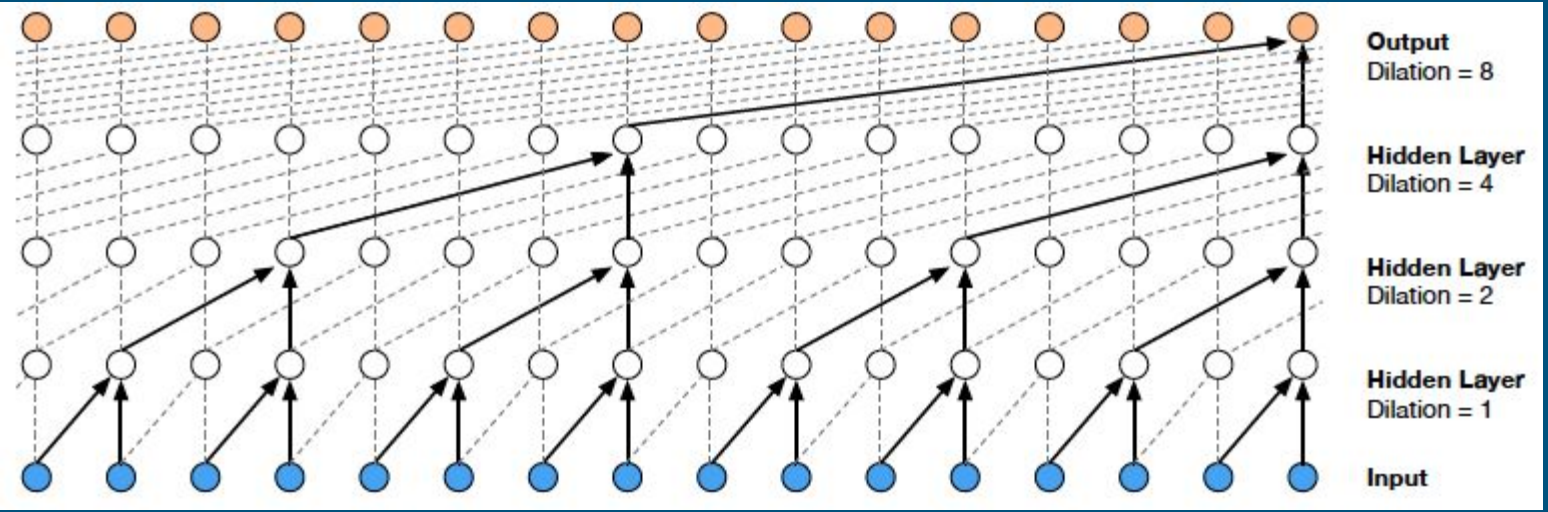
- 모델의 부제를 살펴보면 A Generative Model for Raw Audio 즉 정제되지 않은 오디오 자료를 생성하는 모델 이라는 뜻이다.
- 이후 WaveNet의 단점과 부족한점을 보완하는 과정에서 나오는 모델들에 의해 Neural Vocoder(음성 생성 기술)에 많은 발전이 있었다.
- 음성인식 및 음악을 포함 많은 음성 관련 분야에서 좋은 성능을 보이면서 이용되고 있다.

Causal convolution layers



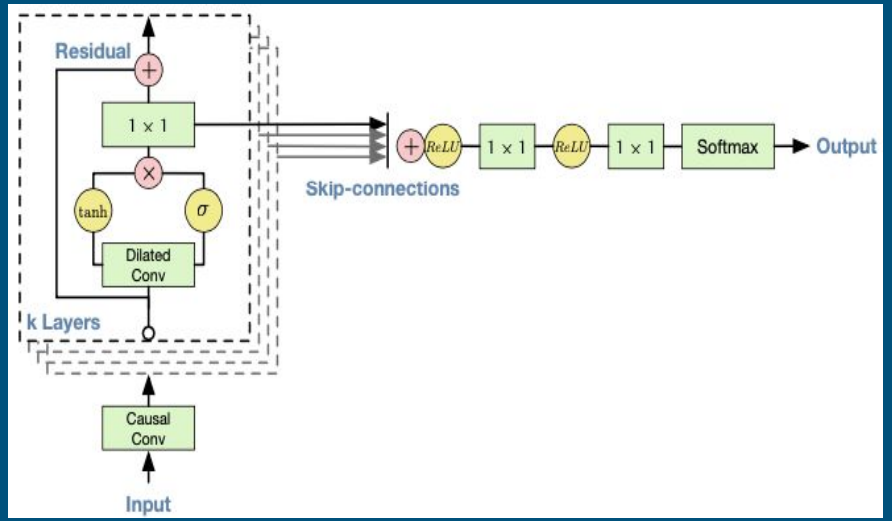
새로운 음성을 생성하기 위해 위의 와 같은 convolution layer를 거치지만 보다시피 output에 관여하는 time stamp(input)은 5개 밖에 되지 않는다. 이를 해결하기 위에 다음 슬라이드의 dilated convolution 을 이용하게 된다.

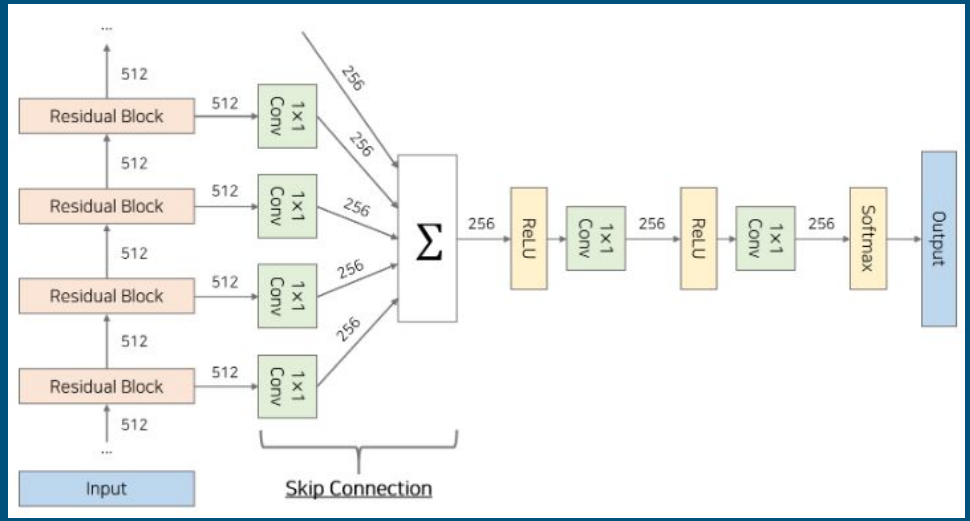
Dilated convolution layers



Dilated(확장된) convolution layers는 많은 오디오 샘플수로 인해 layer를 많이 쌓을수가 없는데 이를 해결하기 위해 사이의 연결을 스킵하는 방식을 사용했다. 이로인해 output에 수렴도 더 잘되고 모델의 깊이를 늘려 정확한 모델을 만들수 있다.

Overview of the model architecture





모델의 전체적인 구성을 보면 결과를 도출하기 위해서 softmax를 사용했고 skip-connections 부분은 이전 슬라이드에서 설명했던 dilated convolution layer다

Conclusion

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)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
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

위의 표에서 볼 수 있듯이 이전의 다른 Vocoder 보다 더 좋은 결과를 보여주고 있다.