pISSN 2005-8063
eISSN 2586-5854
2021. 9. 30.
Vol.13 No.3
pp. 71-78

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2021.13.3.071



# Text-to-speech with linear spectrogram prediction for quality and speed improvement

Hyebin Yoon*

*Department of English Language and Literature, Korea University, Seoul, Korea*

**Abstract**

Most neural-network-based speech synthesis models utilize neural vocoders to convert mel-scaled spectrograms into high-quality, human-like voices. However, neural vocoders combined with mel-scaled spectrogram prediction models demand considerable computer memory and time during the training phase and are subject to slow inference speeds in an environment where GPU is not used. This problem does not arise in linear spectrogram prediction models, as they do not use neural vocoders, but these models suffer from low voice quality. As a solution, this paper proposes a Tacotron 2 and Transformer-based linear spectrogram prediction model that produces high-quality speech and does not use neural vocoders. Experiments suggest that this model can serve as the foundation of a high-quality text-to-speech model with fast inference speed.

**Keywords:** speech synthesis, machine learning, artificial intelligence, text-to-speech (TTS)

## 1. 서론

음성 합성(speech synthesis)은 기계를 사용해 인간의 발화를 생성하는 기술이다. 텍스트를 음성으로 변환하는 과정이라는 점에서 Text-To-Speech(TTS)라고도 한다. 다양한 종류의 음성 합성 기술 중에서도 최근에는 딥러닝(deep learning)을 사용한 음성 합성이 사용되는 추세이다.

딥러닝 기반의 음성 합성 방식은 두 가지로 분류될 수 있다. 첫 번째는 텍스트를 선형 스펙트로그램으로 변환한 후, 선형 스펙트로그램으로부터 음성을 복원하는 방식이다. 대표적으로는 Tacotron(Wang et al., 2017)과 Deep Convolutional TTS(DCTTS; Tachibana et al., 2018)가 있다. Wang et al.(2017)의 Tacotron은 Recurrent Neural Network(RNN) 기반의 음성 합성 모델로, 텍스트를 선형 스펙트로그램으로 변환한 후, Griffin-Lim Algorithm (GLA; Griffin & Lim, 1984)을 사용하여 음성을 생성한다. Encoder와 decoder 중간에는 attention mechanism을 사용함으로써 모델이 텍스트와 스펙트로그램 간의 관계를 학습하도록 한다. Tachibana et al.(2018)의 DCTTS는 Convolutional Neural Network(CNN)를 사용하여 선형 스펙트로그램을 생성하며, RNN 기반에 비해 병렬 처리 과정이 많아서 속도가 더 빠른 모델이다. 음성 생성을 위해서는 realtime GLA라고도 일컫는 RTISI-LA(Zhu et al., 2006)를 사용한다.

* hby1117@korea.ac.kr, Corresponding author

Received 1 August 2021; Revised 19 September 2021; Accepted 19 September 2021

위와 같은 방식은 선형 스펙트로그램으로부터 음성을 복원할 때 일반 GLA, fast GLA, RTISI-LA와 같은 GLA 계열의 알고리즘을 사용한다. 그러나 GLA를 통한 음성 생성은 속도가 빠르지만, 음질이 낮다는 문제가 있다(Shen et al., 2018; Wang et al., 2017).

두 번째는 텍스트를 멜 스펙트로그램(mel-scaled spectrogram)으로 변환하는 음향 모델과 멜 스펙트로그램을 음성으로 변환하는 보코더(vocoder) 모델을 사용하는 방식이다. 이 방식은 현재 딥러닝 기반 음성 합성에서 좀 더 폭넓게 사용된다(Hsu et al., 2020). 대표적인 음향 모델로는 Shen et al.(2018)의 Tacotron 2와 Ren et al.(2019)에서 제안된 FastSpeech를 들 수 있다. Tacotron 2(Shen et al., 2018)는 Tacotron(Wang et al., 2017)의 Seq2seq 구조를 단순화하면서도 성능을 개선시킨 모델로, 선형 스펙트로그램 대신 멜 스펙트로그램을 예측한다. FastSpeech(Ren et al., 2019)는 Transformer(Vaswani et al., 2017) 기반의 음향 모델로서 스펙트로그램을 순차적으로 생성하는 Tacotron 2와 다르게 스펙트로그램의 길이를 예측하고, 예측한 길이만큼의 스펙트로그램을 병렬로 생성한다.

음향 모델에서 생성한 멜 스펙트로그램은 뉴럴넷 기반 보코더 모델로 넘어가 음성으로 변환된다. 뉴럴넷 기반 보코더는 모델이 생성하는 음성의 질을 향상시키려는 필요성에서 도입되었다(Song et al., 2020). van den Oord et al.(2016)의 WaveNet은 이전 시간대의 음성으로부터 다음 시간대의 음성을 예측하는 causal CNN 기반의 보코더로, 양질의 음성을 생성하지만 병렬처리를 하지 않아서 속도가 느리다는 단점이 있다. Prenger et al.(2018)의 WaveGlow는 flow 기반의 보코더로 전체 시간대의 음성을 샘플 레이트 대비 그룹 사이즈로 나눠 나오는 숫자만큼의 분절 구간을 병렬화하고, 각 분절에서 각 그룹 사이즈만큼의 샘플을 각각 생성함으로써 WaveNet에 비해 속도가 향상되었다. Kumar et al.(2019)의 MelGAN은 CNN과 Generative Adversarial Network(GAN) 기반의 보코더로, WaveGlow보다도 속도가 향상되었다.

뉴럴넷 기반 보코더는 출력 음성의 품질 향상을 위한 훈련이 가능하기 때문에 GLA에 비해 양질의 음성을 생성한다는 장점이 있다(Shen et al., 2018). 그러나 뉴럴넷 보코더를 사용하는 경우에는 다음의 문제를 감수해야 한다. 첫 번째, 모델의 훈련에 소요되는 시간이 길다. 음향 모델과 보코더 모델을 각자 훈련한 후, 두 모델을 조율해야 하기 때문이다. 음향 모델과 보코더 모델이 각자 훈련이 완료되었으며 성능이 높다고 하더라도, 두 모델을 결합했을 때의 성능은 다를 수 있다. 이 경우에는 두 모델의 파라미터를 다시 설정하여 각자 재훈련해야 한다. 게다가 MelGAN(Kumar et al., 2019)과 같은 GAN 기반 보코더는 추가적으로 generator와 discriminator의 훈련의 균형을 맞춰야 하는데, GAN의 훈련은 데이터나 파라미터, 모델 구조 등에 굉장히 민감하기 때문에 어려움이 수반된다(Arjovsky et al., 2017). 두 번째, 음성 합성에 필요한 메모리의 용량이 커진다. 두 개의 모델을 사용해야 하므로 더 많은 양의 메모리를 확보해야 하는데, 이는 범용으로 사용되는 GPU 장비의 수준을 초과할 개연성이 높고, 따라서 실제 서비스에는 어려움이 있을 수 있다. 세 번째, 실제 서비스에서는 GPU를 이용하지 못하는 임베딩 환경이나 CPU 환경이 요구될 때가 있는데, 이러한 경우에 뉴럴넷 기반 보코더의 음성 합성의 속도가 GLA에 비해 현격히 느리다. 기본적으로 GLA 계열은 STFT의 다중 루프(loop) 구성이고, 뉴럴넷 기반 보코더는 매트릭스 병렬 처리 연산의 연쇄 구성이기 때문에 이러한 제한 환경에서의 속도 차이는 현격하다. 즉, GPU를 이용하면서 동시에 더 많은 vRAM을 이용할 수 있는 환경이 아니라면, GLA와 보코더 모델들의 속도는 비교 대상이 아니다.

본 논문에서는 위와 같은 실 서비스 환경에서의 뉴럴넷 기반 보코더의 문제를 해결하기 위해서 첫 번째 방식의 음성 합성 모델을 사용할 것을 제안한다. 물론 앞서 언급한 바와 같이 GLA를 사용하는 첫 번째 방식은 보코더 모델을 사용하는 두 번째 방식에 비해 음질이 떨어진다는 단점이 있다. 그러나 이 문제는 선형 스펙트로그램을 예측하는 모델의 성능 향상을 통해 개선될 수 있다. 모델로부터 예측된 선형 스펙트로그램이 아닌 실제 음성의 선형 스펙트로그램에 대한 GLA의 음질이 매우 낮다고 볼 수는 없다. Prenger et al.(2018)의 성능 실험에서 GLA, WaveNet, WaveGlow의 Mean Opinion Score(MOS)는 각각 3.823±0.1349, 3.885±0.1238, 3.961± 0.1343으로 GLA 자체의 성능은 WaveNet, WaveGlow 등의 보코더 모델과 큰 차이가 나지 않는다. 또한 뉴럴넷 기반 보코더는 한정된 데이터로 학습한 모델이기 때문에 실제 음성의 멜 스펙트로그램을 입력값으로 받더라도 훈련 시에 보지 못했던 데이터에 대해 항상 높은 성능을 보인다는 보장이 없다. 이에 반해 GLA는 convergent한 phase reconstruction이기 때문에 모델에서 선형 스펙트로그램만 준수하게 예측할 수 있다면 안정성이 높다.

따라서 본 논문에서는 GLA를 사용하면서도 보코더 모델을 사용하는 음성 합성 모델과 성능이 비슷한 선형 스펙트로그램 예측 모델을 제시하고자 한다.

본 논문은 다음과 같이 구성된다. 2장에서는 본 모델의 구조와 훈련 방식을 설명한다. 3장에서는 본 모델의 성능과 속도 비교 실험 결과를 서술한다. 4장에서는 3장에 대한 해석과 본 논문의 한계 및 의의를 다룬다.

## 2. 제안 모델

### 2.1. 구조

본 논문에서는 Tacotron 2(Shen et al., 2018)의 인코더(encoder)-디코더(decoder)와 Transformer(Vaswani et al., 2017) 기반 선형 스펙트로그램 디코더를 결합한 선형 스펙트로그램 예측 모델을 제안한다. 본 모델의 전체 구조는 그림 1과 같이 텍스트 인코더(text encoder), 멜 스펙트로그램 디코더(mel-scaled spectrogram decoder), 선형 스펙트로그램 디코더(linear spectrogram decoder)로 구성된다. 텍스트 인코더(text encoder)에서 텍스트를 입력값으로 받아 처리한 후, 멜 스펙트로그램 디코더(mel-scaled spectrogram decoder)에서 텍스트 정보를 기반으로 멜 스펙트로그램을 생성한다. 이 인코더와 디코더는 Tacotron 2(Shen et al., 2018)의 구조

를 따른다. 선형 스펙트로그램 디코더(linear spectrogram decoder)에서는 멜 스펙트로그램을 입력값으로 받아 선형 스펙트로그램을 생성한다. 선형 스펙트로그램 디코더는 Transformer(Vaswani et al., 2017)의 구조를 기반으로 한다. 마지막으로 Fast GLA(Perraudin et al., 2013)를 사용해 선형 스펙트로그램에서 음성을 복원한다.

Tacotron(Wang et al., 2017), DCTTS(Tachibana et al., 2018) 등의 선형 스펙트로그램 예측 모델은 텍스트를 멜 스펙트로그램으로 변환한 후, 멜 스펙트로그램을 선형 스펙트로그램으로 변환하는 과정을 거친다. 선형 스펙트로그램의 차원은 멜 스펙트로그램의 차원에 비해 항상 높기 때문에, 저차원의 멜 스펙트로그램을 일단 예측하고 그 이후에 선형 스펙트로그램의 차원 공백을 메꾸는 시도를 하는 것이 유리하기 때문이다. 따라서 본 논문에서도 우선 멜 스펙트로그램을 예측하고 추후 이를 선형 스펙트로그램으로 확장시킨다.

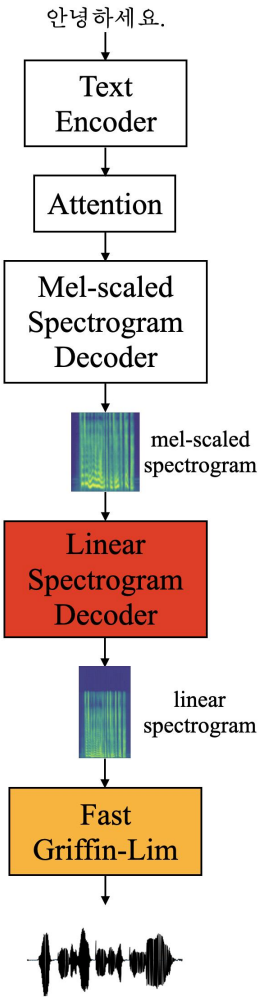


**그림 1.** 선형 스펙트로그램 예측 모델
**Figure 1.** Linear spectrogram prediction model

### 2.1.1. 텍스트 인코더

텍스트 인코더(text encoder)는 텍스트를 입력값으로 받은 후, 텍스트에 관련된 중요한 특징을 추출하여 디코더(decoder)로 전달하는 부분으로, Tacotron 2(Shen et al., 2018)의 인코더(encoder)와 동일한 구조를 지닌다. 즉, 1-dimensional CNN과 bidirectional RNN으로 구성된다. 일반적인 RNN은 이전의 시간대에서 이후의 시간대로의 한 방향의 맥락(context)만 고려하지만, bidirectional RNN은 이후의 시간대에서 이전의 시간대로의 방향도 고려하기 때문에 더 많은 양의 정보를 추출할 수 있다는 장점이 있다. 본 모델의 텍스트 인코더와 Tacotron 2의 인코더와의 차이점은 사용하는 활성화 함수(activation function)에 있다. Tacotron 2에서는 ReLU를 사용하지만, 텍스트 인코더에서는 leaky-ReLU를 활용한다. 음수인 입력값을 모두 0으로 변환하는 ReLU와 달리, leaky-ReLU는 음수인 입력값에 대해서도 0이 아닌 값을 출력함으로써 vanishing gradient 문제의 발생을 감소시킨다.

### 2.1.2. 멜 스펙트로그램 디코더

멜 스펙트로그램 디코더(mel-scaled spectrogram decoder)는 텍스트 인코더로부터 받은 텍스트 특징을 바탕으로 멜 스펙트로그램을 예측하는 부분으로, Tacotron 2(Shen et al., 2018)의 디코더와 동일한 구조를 지닌다. 즉, prenet, unidirectional RNN, linear projection과 postnet으로 구성된다. Tacotron 2와 마찬가지로, linear projection 중 한 개는 inference 환경에서 멜 스펙트로그램 생성이 중단되어야 하는 지점을 나타내는 stop token을 예측한다.

### 2.1.3. Attention

Attention은 인코더의 입력값 텍스트와 디코더의 출력값 멜 스펙트로그램의 구성 요소를 매핑(mapping)하는 데 사용된다. 다시 말해, 텍스트를 구성하는 각 글자와 멜 스펙트로그램을 구성하는 각 프레임(frame)을 연결(map)한다. Attention은 인코더에서 추출한 정보를 디코더로 전달하는 통로이기도 하다. 본 논문에서는 Tacotron 2(Shen et al., 2018)와 동일하게 location-sensitive attention을 사용한다. Location-sensitive attention은 텍스트와 멜 스펙트로그램과의 관계를 계산할 때, 특정 글자와 멜 스펙트로그램 프레임 간의 의미의 유사성뿐만 아니라, 위치의 근접성까지 고려한다.

### 2.1.4. 선형 스펙트로그램 디코더

선형 스펙트로그램 디코더(linear spectrogram decoder)는 멜 스펙트로그램을 선형 스펙트로그램으로 변환하는 작업을 하며, Transformer(Vaswani et al., 2017) 기반의 구조이다. Vaswani et al.(2017)의 Transformer는 원래 자연어 처리를 위해 제안된 encoder-decoder 구조의 모델이다. 그러나 Transformer는 이 외에도 음성 합성(Ren et al., 2019), 가창 합성(Chen et al., 2020) 등 다양한 분야에서 활용된다. 본 논문에서는 Transformer의 encoder와 decoder의 근간이 되는 multi-head attention, point-wise feed-forward network와 positional encoding 기법을 사용해 멜 스펙트로그램을 선형 스펙트로그램으로 변환했다.

선형 스펙트로그램 디코더의 기본적인 구조는 그림 2와 같다.

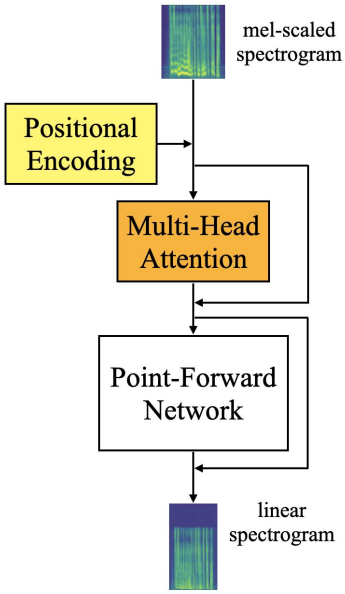


**그림 2.** 선형 스펙트로그램 디코더
**Figure 2.** Linear spectrogram decoder

모델이 입력값의 시간 혹은 위치 정보를 파악할 수 있도록 positional encoding이 입력값에 더해지면 입력값 임베딩은 시간 혹은 위치 정보의 상대적 값을 포함하게 된다. 시간 혹은 위치 정보가 포함된 입력값은 Multi-head attention을 통과한다. 선형 스펙트로그램 디코더에서 사용되는 Multi-Head attention에는 두 가지 특징이 있다. 첫 번째, 이름에서 명시하듯이 하나의 입력값을 세분화하여 여러 개의 attention을 계산한다. 두 번째, 입력값과 입력값 자기 자신과의 관계를 매핑하는 self-attention을 사용한다. 텍스트와 멜 스펙트로그램과의 관계를 계산하던 location-sensitive attention과 다르게, 멜 스펙트로그램과 멜 스펙트로그램 자신과의 관계를 계산한다.

Multi-head attention을 통과한 입력값은 attention을 통과하기 이전의 입력값과 더해진 후, Deep Neural Network(DNN)로 구성된point-wise feed-forward network를 통과한다. 출력된 특징은 DNN을 통과하기 이전의 특징과 더해진다. 위와 같은 과정을 여러 번 반복한 후에 선형 스펙트로그램이 출력된다.

#### 2.1.5. Fast Griffin-Lim algorithm(Perraudin et al., 2013)

음성에 Short-Time Fourier Transform(STFT)을 적용했을 때 출력되는 선형 스펙트로그램에는 각 주파수의 amplitude와 phase 정보가 포함된다. 대부분의 상황에서는 amplitude 정보만 필요하기 때문에 선형 스펙트로그램의 절댓값을 음성 특징으로 취한다. 그러나 선형 스펙트로그램을 원래의 음성으로 복구시키기 위해서는 amplitude 뿐만 아니라 phase 정보가 필요하다. 손실된 phase 정보를 유추하여 선형 스펙트로그램을 음성으로 복원시키는 기술이 GLA(Griffin & Lim, 1984)이다.

Perraudin et al.(2013)은 alpha 파라미터를 사용하여 적은 iteration으로도 최적의 음성으로 복원이 가능한 fast GLA를 개발하였다. Sharma et al.(2020)의 실험에 따르면 fast GLA의 MOS는 GLA뿐만 아니라 GAN 기반 보코더보다도 높았다. 이 실험은 fast GLA의 성능이 뉴럴넷 기반 보코더와 비슷하거나 더 높을 가능성을 시사한다. 본 모델에서는 이 실험을 토대로 하여 선형 스펙트로그램 디코더에서 생성된 선형 스펙트로그램에 fast GLA를 적용한 음성을 최종 결과물로 생성한다.

### 2.2. Cost Function

본 모델의 훈련에는 5가지 종류의 cost function이 사용된다. 모델은 아래의 cost function을 모두 최소화하는 방향으로 학습된다.

첫 번째는 모델이 예측한 postnet 이전의 멜 스펙트로그램 $M_p$와 타겟(target) 멜 스펙트로그램 $M_t$와의 binary cross entropy loss이다. 본 모델에서 사용되는 멜 스펙트로그램의 값의 범위는 0과 1 사이로 한정되어 있으며, 이 값을 확률로 고려하여 식 (1)과 같은 binary cross entropy loss를 채택하였다. 두 번째는 모델이 예측한 postnet 이후의 멜 스펙트로그램 $M_{p'}$와 타겟 멜 스펙트로그램 $M_t$와의 binary cross entropy loss로 식 (2)와 같다. 세 번째는 모델이 예측한 선형 스펙트로그램 $L_p$와 타겟 선형 스펙트로그램 $L_t$와의 binary cross entropy loss로 식 (3)과 같다. 멜 스펙트로그램과 마찬가지로 선형 스펙트로그램의 범위는 0과 1 사이로 한정되었다. 네 번째는 모델이 예측한 stop token $S_p$와 타겟 stop token $S_t$와의 binary cross entropy loss로 식 (4)와 같다. Stop token은 모델이 음성 생성을 중단해야 할 때 1, 지속해야 할 때 0을 출력하도록 설계되었기 때문에 식 (1)–(3)과 동일한 cost function을 사용한다. 마지막은 guided attention loss로, location-sensitive attention의 alignment의 양상을 특정 스텝까지는 대각선으로 강제시킨다. 식 (5)의 $A$는 attention matrix에 해당하며, $W$는 attention weight을 대각선으로 강제시키는 장치이다. 모델의 총 cost function은 식 (6)과 같다.

$$C_1 = -\sum M_t \cdot \log M_p + (1-M_t)\cdot \log(1-M_p) \quad (1)$$

$$C_2 = -\sum M_t \cdot \log M_{p'} + (1-M_t)\cdot \log(1-M_{p'}) \quad (2)$$

$$C_3 = -\sum L_t \cdot \log L_p + (1-L_t)\cdot \log(1-L_p) \quad (3)$$

$$C_4 = -\sum S_t \cdot \log S_p + (1-S_t)\cdot \log(1-S_p) \quad (4)$$

$$C_5 = \frac{1}{n}\sum |A \cdot W| \quad (5)$$

$$C = C_1 + C_2 + C_3 + C_4 + C_5 \quad (6)$$

## 3. 실험

본 모델의 성능과 속도를 평가하기 위한 실험을 다음과 같이 진행하였다. 본 모델과 비교할 모델로는 Tacotron 2에 WaveGlow를 보코더로 결합한 모델을 사용했다. 먼저 두 모델을 각자 훈련한 후, 각 모델에서 20개의 음성을 생성하였다. 두 모델의 음성에 대한 Mean Opinion Score(MOS) 점수를 기반으로 성능을 비교하였다. 또한, 각 모델에서 100개의 음성을 생성한 후, 음성

생성 속도를 비교하였다.

### 3.1. 모델

#### 3.1.1. 데이터

실험에 사용된 모델은 성인 여자 성우 1명의 한국어 발화를 녹음한 자체 수집 데이터로 학습되었다. 총 약 20시간에 해당하는 15,000개의 음성 데이터와 전사 데이터가 훈련에 사용되었으며, 모델이 학습하지 않은 20개의 문장을 평가용 음성 생성의 입력값으로 사용했다. 텍스트 데이터는 한글 및 스페이스, 콤마(,), 마침표(.), 느낌표(!), 물음표(?)로 구성되며, 숫자, 영어 및 기타 부호는 한글로 변환하거나 삭제하였다. 음성 데이터는 sample rate 22.05 kHz의 16-bit mono pcm으로, 음성의 전후에 공백(silence) 구간이 존재한다. 멜 스펙트로그램 및 선형 스펙트로그램을 추출할 때의 frame 크기는 1,024 samples, frame 이동 간격은 256 samples, mel bin의 크기는 80으로 설정하였다. 멜 스펙트로그램과 선형 스펙트로그램의 범위는 0과 1 사이로 제한하였다.

#### 3.1.2. 훈련 및 예측

선형 스펙트로그램 예측 모델은 Shen et al.(2018)의 훈련 파라미터와 동일한 파라미터로 훈련하였다. Adam optimizer의 learning rate은 0.001, $\beta_1$은 0.9, $\beta_2$는 0.999, $\varepsilon$은 $10^{-6}$으로 설정하였다. 성능 및 속도 비교에 사용되는 Tacotron 2 또한 Shen et al.(2018)과 동일한 파라미터를 사용하였다. 두 모델은 모두 배치(batch) 32개로 총 20만 스텝을 돌았다. Tacotron 2의 보코더로 사용되는 WaveGlow는 배치 16개로 총 20만 스텝을 돌았다. Adam optimizer의 learning rate은 $10^{-4}$, $\beta_1$은 0.9, $\beta_2$는 0.999, $\varepsilon$은 $10^{-8}$으로 설정하였다.

### 3.2. 결과

MOS는 음성 합성 모델의 성능을 평가하는 데 활용되는 보편적인 방법이다. 평가자는 음성의 자연스러움을 주로 1–5 사이의 주관적인 점수로 평가하며, 평가자들의 점수의 평균이 모델의 성능을 나타내는 지표가 된다.

본 실험에서는 20–30대의 남녀 14명을 대상으로 선형 스펙트로그램 예측 모델과 Tacotron 2+WaveGlow가 생성한 음성을 평가하도록 하였다. 각 평가자는 각 모델이 생성한 20개의 음성을 평가하였으며, 총 40개의 음성을 평가하였다. 두 모델의 음성에 대한 MOS은 95% 신뢰 구간에서 표 1과 같다. 선형 스펙트로그램 예측 모델과 Tacotron 2+WaveGlow의 MOS는 근소한 차이를 보인다.

**표 1.** Mean opinion score (MOS)
**Table 1.** Mean opinion scores (MOS)

| 모델 | MOS |
|---|---|
| Linear spectrogram prediction model | 3.65±0.085 |
| Tacotron 2 +WaveGlow | 3.55±0.086 |

본 실험에서는 MOS 실험 외에도 음성 생성 속도 실험을 진행하였다. NVIDIA Tesla A100 GPU 환경에서 본 모델과 Tacotron 2+WaveGlow의 문장 100개에 대한 평균 음성 생성 속도를 측정하였다. 문장은 최소 5자부터 최대 60자까지의 다양한 길이의 문장을 사용했다. 표 2에 따르면, 본 모델의 속도는 Tacotron 2+WaveGlow에 비해 약 6배 빠르다는 것을 알 수 있다.

**표 2.** 음성 생성 속도
**Table 2.** Inference speed

| 모델 | 평균 속도 (초) |
|---|---|
| Linear spectrogram prediction model | 0.4381 |
| Tacotron 2+ WaveGlow | 2.7146 |

## 4. 논의 및 결론

본 논문에서는 뉴럴넷 기반 보코더 없이도 준수한 음질의 음성을 빠른 속도로 생성하는 선형 스펙트로그램 예측 모델을 제시하고, 보코더 기반 모델과의 비교를 통해 성능 및 속도를 평가하였다. 실험 결과, 선형 스펙트로그램 예측 모델이 보코더 기반 모델보다 GPU에서 속도가 더 빠르며, 따라서 CPU나 임베딩 환경에서의 서비스 응용 가능성 또한 높다. 아울러 음질 측면에서 성능이 미세하게 좋다는 점을 발견하였다.

위의 결과에 대해서는 다음과 같은 해석이 가능하다. 먼저, 본 모델은 보코더 기반 모델보다 규모가 작기 때문에 연산량이 줄어들어 속도 면에서 우세하다.

다음은 성능 실험에 대한 해석이다. MOS 실험의 결과는 선형 스펙트로그램 예측 모델이 보코더 모델을 사용한 Tacotron 2보다 근소한 차이로 성능이 좋다는 점을 시사한다. 1장에서 언급했던 바와 같이, 고품질의 선형 스펙트로그램을 입력값으로 받는 GLA는 뉴럴넷 기반 보코더와 미세한 성능의 차이를 보인다. 본 모델이 GLA를 사용했음에도 WaveGlow 보코더를 사용한 모델보다 성능이 미세하게 높게 나온 것은 본 모델이 실제 선형 스펙트로그램과 유사한 선형 스펙트로그램을 예측한다는 점을 암시한다.

또한, GLA의 안정성과 fast GLA의 향상된 성능이 결과에 영향을 미쳤을 것으로 해석된다. 1장에서 언급한 바와 같이, GLA는 학습하는 모델이 아니기 때문에 보코더보다 더 안정성이 있으며, 2장에서 언급했던 바와 같이, fast GLA는 기존의 GLA와 일부 뉴럴넷 기반 보코더에 비해 성능이 높기 때문이다. 그러나 이 MOS 결과는 본 실험에서 사용한 특정한 여성 성우 화자 한 명의 데이터에 대해 나온 것이기 때문에, 다양한 화자에 대한

MOS 실험이 진행되어야 할 필요가 있다.

본 논문에서는 14명의 평가자와 20개의 음성으로 선형 스펙트로그램 예측 모델의 성능을 평가했다. 그러나 모델의 성능이 포괄적으로 측정되기 위해서는 다양한 그룹의 평가자들이 더 많은 음성을 듣고 평가해야 할 필요성이 있다. Wang et al.(2017)은 100개의 음성에 대해 한 개의 음성 당 8개의 평가를 받도록 했다. Li et al.(2019)은 38개의 음성에 대해 한 개의 음성 당 최소 20개의 평가를 받도록 했다. 또한, 다양한 환경에서의 CPU와 GPU를 사용한 음성 생성 속도 측정이 필요하다.

그럼에도 불구하고 본 논문은 보코더 기반 모델과 비슷한 성능으로 정해진 시간 내에 더 많은 음성을 생성할 수 있는 모델의 가능성을 제시했다는 점에서 의의가 있다. 또한, 선형 스펙트로그램 예측 모델에서 선형 스펙트로그램 디코더는 멜 스펙트로그램을 생성하는 다양한 모델의 뒷부분에 결합되어 GLA와 같이 사용된다면 보코더 모델 대신 빠른 속도로 음성을 생성할 수 있다. 즉, 텍스트 인코더와 멜 스펙트로그램 디코더를 다른 모델로 교체할 수 있다는 유연성이 존재한다.

선형 스펙트로그램 예측 모델의 가능성을 확장시키기 위해서는 우선 선형 스펙트로그램 예측 모델과 여러 보코더 기반 모델의 성능을 비교 분석할 필요가 있다. 또한, 다양한 데이터셋에 대한 선형 스펙트로그램 예측 모델의 성능 비교 실험이 필요하다. 그 외에도 선형 스펙트로그램 디코더와 다른 모델과의 결합에 관한 연구도 진행될 수 있다. 그동안 음성 합성에서 보코더의 역할이 강조되었으며, 성능과 속도가 모두 높은 보코더 기반 모델에 대한 연구가 진행되었다. 본 논문과 같이 뉴럴넷 기반 보코더를 사용하지 않고도 높은 성능과 속도를 보이는 1개의 모델을 개발하는 방안을 집중적으로 모색해야 할 것이다.

## References


Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. Retrieved from https://arxiv.org/abs/1701.07875

Chen, J., Tan, X., Luan, J., Qin, T., & Liu, T. Y. (2020). HiFiSinger: Towards high-fidelity neural singing voice synthesis. Retrieved from https://arxiv.org/abs/2009.01776

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(2), 236-243.

Hsu, P., Wang, C., Liu, A. T., & Lee, H. (2020). Towards robust neural vocoding for speech generation: A survey. Retrieved from https://arxiv.org/abs/1912.02461

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brebisson, A., ... Courville, A. (2019). MelGAN: Generative adversarial networks for conditional waveform synthesis. Retrieved from https://arxiv.org/abs/1910.06711

Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M., & Zhou, M. (2019). Neural speech synthesis with transformer network. Retrieved from https://arxiv.org/abs/1809.08895

Perraudin, N., Balazs, P., & Søndergaard, P. L. (2013, October). A fast Griffin-Lim algorithm. *Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 1-4). New Paltz, NY.

Prenger, R., Valle, R., & Catanzaro, B. (2018). WaveGlow: A flow-based generative network for speech synthesis. Retrieved from https://arxiv.org/abs/1811.00002

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2019, December). FastSpeech: Fast, robust and controllable text to speech. *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems* (pp. 3156-3164). Vancouver, BC.

Sharma, A., Kumar, P., Maddukuri, V., Madamshetti, N., Kishore, K. G., Kavuru, S. S. S., Raman, B., ... Roy, P. P. (2020). Fast Griffin Lim based waveform generation strategy for text-to-speech synthesis. *Multimedia Tools and Applications*, *79*(41), 30205-30233.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., ... Wu, Y. (2018, April). Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4779-4783). Calgary, AB.

Song, W., Xu, G., Zhang, Z., Zhang, C., He, X., & Zhou, B. (2020, October). Efficient WaveGlow: An improved WaveGlow vocoder with enhanced speed. *Proceedings of the 21st Annual Conference of the International Speech Communication Association* (pp. 225-229). Shanghai, China.

Tachibana, H., Uenoyama, K., & Aihara, S. (2018, April). Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4784-4788). Calgary, AB.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., ... Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. Retrieved from https://arxiv.org/abs/1609.03499

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., ... Polosukhin, I. (2017). Attention is all you need. Retrieved from https://arxiv.org/abs/1706.03762

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., ... Saurous, R. A. (2017, August). Tacotron: Towards end-to-end speech synthesis. *Proceedings of the 18th Annual Conference of the International Speech Communication Association* (pp. 4006-4010). Stockholm, Sweden.

Zhu, X., Beauregard, G. T., & Wyse, L. (2006, July). Real-time iterative spectrum inversion with look-ahead. *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo* (pp. 229-232). Toronto, ON.

• **윤혜빈 (Hyebin Yoon)** 교신저자
고려대학교 문과대학 영어영문학과 박사과정
서울시 성북구 안암로 145
Tel: 02-3290-1980
Email: hby1117@korea.ac.kr
관심분야: 음성학, 언어공학

# 음질 및 속도 향상을 위한 선형 스펙트로그램 활용 Text-to-speech

윤 혜 빈

고려대학교 영어영문학과

**국문초록**

인공신경망에 기반한 대부분의 음성 합성 모델은 고음질의 자연스러운 발화를 생성하기 위해 보코더 모델을 사용한다. 보코더 모델은 멜 스펙트로그램 예측 모델과 결합하여 멜 스펙트로그램을 음성으로 변환한다. 그러나 보코더 모델을 사용할 경우에는 많은 양의 컴퓨터 메모리와 훈련 시간이 필요하며, GPU가 제공되지 않는 실제 서비스 환경에서 음성 합성이 오래 걸린다는 단점이 있다. 기존의 선형 스펙트로그램 예측 모델에서는 보코더 모델을 사용하지 않으므로 이 문제가 발생하지 않지만, 대신에 고품질의 음성을 생성하지 못한다. 본 논문은 뉴럴넷 기반 보코더를 사용하지 않으면서도 양질의 음성을 생성하는 Tacotron 2 & Transformer 기반의 선형 스펙트로그램 예측 모델을 제시한다. 본 모델의 성능과 속도 측정 실험을 진행한 결과, 보코더 기반 모델에 비해 성능과 속도 면에서 조금 더 우세한 점을 보였으며, 따라서 고품질의 음성을 빠른 속도로 생성하는 음성 합성 모델 연구의 발판 역할을 할 것으로 기대한다.

**핵심어:** 음성 합성, 기계 학습, 인공지능, Text-to-speech (TTS)