Smart Media Journal / Vol.15, No.3 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2026.15.3.54

Diffusion-TTS 보코더 최적화를 위한 신경망 경량화 및 고속화 기법에 관한 연구

(A Study on Neural Network Compression and Acceleration Techniques for Diffusion-TTS Vocoder Optimization)

박 수 빈*, 김 진 성*, 정 세 훈***, 심 춘 보**
(Su Bin Park, Jin Seong Kim, Se Hoon Jung, Chun Bo Sim)

요 약

TTS(Text-to-Speech) 기술은 주어진 텍스트로부터 자연스러운 음성을 합성하는 것을 목표로 하며, 인공지능 분야에서 연구가 활발히 진행되고 있다. 생성형 모델인 Diffusion 기반의 TTS는 세밀한 인간 음성 특징을 효과적으로 학습하여 매우 자연스러운 음성을 생성할 수 있다. 그러나 순방향 및 역방향 프로세스의 반복으로 상당한 연산량과 시간이 소모된다. 이를 해결하기 위해 본 논문에서는 각 입력 채널에 별도의 커널을 적용하는 기법인 Depthwise Separable Convolution과 입력에 따라 동적으로 커널을 조정하는 LVC(Location-Variable Convolution)를 결합한 Depthwise Separable LVC를 사용하여 최적화된 Diff-TTS 모델을 제안한다. 제안된 모델은 음성 품질을 유지하면서 문장 당 연산량이 줄어든 것을 확인했으며, 이는 채널별 연산과 통합을 통해 음성 합성의 속도와 품질을 효과적으로 향상하는 것을 보여준다.

■ 중심어 : 딥러닝 ; 음성합성 ; 디퓨전 ; 깊이별 분리 가능 컨볼루션 ; 위치 변수 컨볼루션

Abstract

Text-to-Speech (TTS) technology aims to synthesize natural-sounding speech from given text, and research in the field of artificial intelligence is actively underway. Diffusion-based TTS, a generative model, effectively learns detailed human speech characteristics and can produce remarkably natural-sounding speech. However, the repeated forward and backward processes consume significant computational resources and time. To address this issue, this paper proposes an optimized Diff-TTS model using Depthwise Separable LVC, which combines Depthwise Separable Convolution (DSC), a technique that applies separate kernels to each input channel, and Location-Variable Convolution (LVC), which dynamically adjusts the kernel based on the input. The proposed model demonstrates a reduced computational load per sentence while maintaining speech quality, demonstrating that channel-specific computation and integration effectively improves the speed and quality of speech synthesis.

■ keywords : Deep Learning ; Text to Speech ; Diffusion; Depthwise Separable Convolution ; Location Variable Convolution

Ⅰ. 서 론

음성합성(Text To Speech, TTS) 기술은 음성, 언어 및 기계 학습을 활용하는 연구 분야로, 텍스트를 사람이 이해하기 쉽고 자연스러운 음성으로 합성하는 연구 분야다. 초기에은 음성 합성을 위해 복잡한 과정과 많은 시간이 소요되었다. 이를 개선하기 위해 심층 신경망(Neural Network)을 음성합성 모델에 적용한 신경망 기반 TTS (Neural TTS)가 등장했다[1-9].

신경망 기반 TTS는 언어적 특징으로부터 직

* 준회원, 순천대학교 IT-Bio융합시스템전공
** 정회원, 순천대학교 IT-Bio융합시스템전공
*** 종신회원, 순천대학교 컴퓨터공학전공

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00407739) and This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2025-25434896).
접수일자 : 2026년 03월 10일 게재확정일 : 2026년 03월 20일
수정일자 : 2026년 03월 20일
교신저자 : 심춘보 e-mail : cbsim@scnu.ac.kr

접 파형을 생성하며, 자연스러운 음성합성이 가능해졌다. 그러나 음성 데이터는 보코더(Vocoder)에서 음성 추론 및 합성 과정에 많은 시간이 소요되는 문제점이 발생했다. 이를 해결하고자 GAN(Generative Adversarial Networks)[10]이나 Diffusion[11] 등과 같은 생성 모델을 적용하는 연구가 활발히 이뤄지고 있다[12].

Diff-TTS(Diffusion Text To Speech)[13]는 Diffusion 모델 기반의 TTS 기술로 음성 합성의 품질을 높였을 뿐만 아니라 학습 및 합성에 대한 안전성도 크게 향상되었다[14]. NaturalSpeech2[15]와 NaturalSpeech3[16]와 같은 최신 연구들은 Latent Diffusion 구조와 대규모 학습을 통해 음성의 자연스러움과 일반화 성능을 크게 향상시켰다.그러나 Diffusion 모델의 특성상 지속적인 노이즈(Noise) 제거로 원본 데이터 노이즈 역시 사라지는 문제가 발생했다. 또한, 합성된 음성의 품질을 향상하는 데 비해 다른 음성합성 모델과 비교하여 느린 추론 속도를 나타냈다.

이를 해결하기 위해 연구에서는 Depthwise Separable LVC(Location Variable Convolution)를 적용한 Diff-TTS를 제안한다. 기존 Diff-TTS 구조에 Depthwise Separable Convolution[17]과 LVC[18]를 결합한 모듈을 적용하여, 채널별 독립적인 특징 추출과 위치 가변적 컨볼루션 연산을 동시에 수행하도록 설계하였다. 이를 통해 각 채널에서 입력 음성 특징을 효율적으로 분리 및 처리함으로써, Diffusion 과정에서 발생하는 불필요한 정보 손실을 줄이고 원본 음성의 세밀한 특성을 보다 효과적으로 보존할 수 있다. 또한 Depthwise Separable 구조를 활용하여 연산량을 감소시키고, 결과적으로 모델의 추론 속도 개선을 기대할 수 있다.

Ⅱ. 관 련 연 구

1. Diffusion Text To Speech

Diff-TTS는 음성합성 시스템에 최초로 DDPM(Denoising Diffusion Probabilistic Model)을 도입한 비자기회귀 모델로 정방향 과정과 역방향 과정을 거친다[13]. 정방향 과정은 가우시안 노이즈(Gaussian Noise)로 변환되는 과정으로, 각 전이 단계는 분산 스케줄을 통해 미리 정의된다. 역방향 과정은 조건부 분포를 통해 텍스트 조건과 확산 시간 단계에 해당하는 멜 스펙트로그램(Mel Spectrogram)으로 복원한다.

Diff-TTS는 텍스트 인코더(Text Encoder), 스텝 인코더(Step Encoder), 발음 예측 구간(Duration Predictor), 길이 조정 구간(Length Regulator), 디코더(Decoder)로 구성되어 있다. 텍스트 인코더는 음소 배열을 통해 텍스트의 문맥을 고려하여 각 음소의 특징을 파악하고, 잔차 블록(Residual Block)과 LSTM(Long Short-Term Memory)[19]을 통해 의미가 있는 특징을 추출한다. 발음 예측 구간에서 음성의 각 운율 단위의 지속 시간을 예측하고, 길이 조정 구간에서 음소와 멜 스펙트로그램 시퀀스(Sequence)의 길이를 일치시킨다. 스텝 인코더는 고차원 표현으로 맵핑하여 해당 단계의 시간에 따른 특징을 추출한다. 디코더는 음소와 확산 단계의 임베딩(Embedding)을 조건으로 하는 t번째 단계 잠재 변수로부터 가우시안 노이즈를 예측한다.

Diff-TTS는 기존의 Tacotron2[20] 및 Glow-TTS(Generative Flow-based Text-to-Speech)[21]보다 적은 파라미터를 사용하면서도 더 좋은 음질의 오디오 합성이 가능했다. 그러나 여전히 과도한 연산량으로 인해 속도가 중요한 음성합성 모델에서 한계를 보였다.

2. Depthwise Separable Convolution

인터넷의 발달로 빅데이터를 이용한 데이

터 분석 및 예측을 할 수 있도록 효율적으로 빠르게 처리할 수 있는 연구가 현재까지 진행되고 있다. 이를 위해 딥러닝에서 핵심적인 Convolution의 연산량을 줄여 실시간으로 사용할 수 있도록 모델의 경량화를 적용한 Depthwise Separable Convolution[15]이 제안되었다.

Depthwise Separable Convolution은 Depthwise Convolution 단계와 Pointwise Convolution 단계로 이루어져 있다. Depthwise Convolution은 채널마다 공간적인 특성을 추출하기 위한 방식으로 각 입력 채널에 대해 따로 Convolution을 수행한다. Point Convolution은 모든 채널에 1×1 Conv를 수행해 채널 수를 조절하는 역할을 한다.

일반적인 Convolution은 한 개의 필터가 채널 전체에 Convolution을 연산한다. 반면에, Depthwise Convolution은 한 개의 필터가 한 개의 채널만 연산하고, Point Convolution은 Depthwise Convolution의 출력을 입력으로 받아 임의의 개수만큼의 필터를 사용하여 연산한다. 이를 통해 일반적인 Convolution과 비교하여 연산량이 줄어들었음을 보여준다.

3. Location Variable Convolution

DNN(Deep Neural Network)이 적용된 음성합성 초기에는 자기회귀 방식으로 접근하여 고품질의 음성 합성할 수 있지만, 추론 속도가 느린 단점이 있었다[22]. 이에 흐름 기반 생성 모델이 제안됐으며, 이전보다 훨씬 빠른 추론 속도를 보였다[23-24]. 이 모델들은 Wavenet[25]을 기반으로, Wavenet과 유사한 네트워크는 시간적 특징을 추출하기 위해 많은 Convolution 층을 사용한다. 이를 개선하고자 시간 의존적 특징을 더 효과적으로 추출할 수 있는 LVC가 제안됐다.

LVC는 파형 간격마다 다른 커널의 계수를 사용할 수 있으며, 멜 스펙트로그램을 조건으로 활용하여 계수를 구하게 된다. 커널 예측 부분이 구간마다 다른 계수를 출력하고, 이를 활용해 구간마다 다른 가중치 값을 가지고 Convolution 연산을 수행한다. 여기서 지역적 조건은 그 시간 위치에 따른 음향 특징으로 불리는 멜 스펙트로그램이 활용되었다. 최종적으로 LVC는 기존 Convolution 연산보다 더 적은 레이어(Layer) 수로 장기 의존성을 더 효과적으로 추출할 수 있었다. Parallel WaveGAN[26]과 비교하여 음성 파형을 생성하는 데 실제로 걸린 시간을 기준으로 0.03초의 차이를 보이며 속도 지표에서 매우 높은 성능을 확인할 수 있었다.

Ⅲ. 본 론

3장에서는 제안하는 Depthwise Separable Convolution과 LVC를 결합한 Depthwise Separable LVC와 이를 적용한 Diff-TTS를 기술한다. 본 논문에서는 FastSpeech2[27]를 백본으로 사용하였으며, 그림 1은 제안하는 음성합성 모델의 전체 구성도이다.

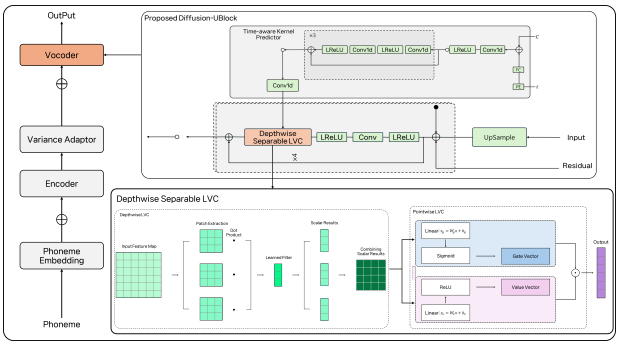


그림 1. 제안하는 모델이 적용된 보코더 프로세스
Fig. 1 Overall configuration of the proposed speech synthesis model

1. 데이터 수집 및 전처리

본 논문에서는 음성 데이터 학습에 많이 사용되는 LJSpeech 데이터 세트[28]를 활용하여 전처리를 진행했다. LJSpeech 데이터 세트는 7권의 논픽션(Non Fiction) 책에서 발췌한 구절을 읽는 한 명의 화자의 짧은 오디오 클립 13,100개로 구성된 음성 데이터이다. 음성 데이터 훈련과 추론하기 위해서 오디오 및 텍스트 파일을 10,480개의 훈련 데이터, 1,310개의 검증 데이터, 1,310개의 데이터로 구성한다.

자연어인 텍스트 데이터와 파형으로 구성된 음성 원본 데이터는 바로 입력 데이터로 사용할 수 없으며, 음성합성을 위한 정렬 및 이진화 등의 데이터 전처리 과정을 거쳐야 한다. 먼저 입력되는 텍스트와 오디오 원본 데이터의 강제 정렬을 진행한다. 이를 통해 음성 WAV(Waveform Audio File Format) 파일 구조를 표준 샘플로 리샘플링(Resampling) 및 정규화(Normalization)를 진행하여 멜스펙트로그램 및 피치 등 추가 특성을 계산하여 말뭉치 파일을 생성한다. 다음으로, 음성 및 말뭉치 파일을 기반으로 MFA(Montreal Forced Aligner)[29]의 음향 모델을 학습하여, 입력된 음성과 텍스트를 프레임 단위로 정렬하는 강제 정렬 모델을 생성한다[30]. 이를 통해 각 스크립트에 나오는 단어들을 음소 단위로 매칭시킬 수 있다. 마지막으로 정렬된 데이터를 특정 포맷으로 변환하여 저장할 수 있는 이진화를 진행하여 음성 데이터의 입출력 속도를 높일 수 있도록 한다.

2. Depthwise Separable LVC를 적용한 Diff-TTS 모델 설계

기존 Diff-TTS 보코더는 Diffusion 과정에서 동일한 Convolution 연산을 반복적으로 수행함으로써 연산량이 증가하고, 이에 따라 추론 속도가 저하되는 한계가 존재한다. 이에 본 논문에서는 보코더 구조에 Depthwise Separable LVC를 결합한 Diffusion 기반 모델을 제안한다. Diffusion 모델 내부에 Depthwise Separable LVC를 적용하여 입력 특징을 보다 효율적으로 처리할 수 있도록 한다.

그림 2는 제안하는 모델이 적용된 보코더 프로세스를 나타낸다. 먼저, 입력된 음성 특징은 Convolution 연산을 통해 시간에 따른 음향 패턴을 추출한다. 이후 단계적 정제 구조를 통한 다운샘플링(Downsampling)으로 해상도를 낮춘다. 이를 통해 음성 세기와 발음 등의 추상적인 표현을 학습하고, 연산의 효율성을 높인다.

다음으로 다운샘플링된 특징와 정보를 Depthwise Separable LVC를 적용한 Diffusion 모델로 입력한다. Iterative Refinement 과정으로 노이즈가 제거된다. Noise Predictor는 현재 단계의 노이즈를 예측하고, 이를 기반으로 음성 신호를 복원한다. 또한 다중 해상도 처리를 위해 DBlock 구조를 활용해 서로 다른 스케일의 특징을 통합적으로 반영한다.

제안하는 구조에서는 Diffusion U-Block 내부에 Depthwise Separable LVC를 적용하여, Depthwise Convolution과 Pointwise Convolution을 통해 입력 특징을 효율적으로 처리하고, 이후 위치 가변적 필터를 적용함으로써 시간적 변화에 따른 특성을 효과적으로 반영한다.

마지막으로 Convolution 연산으로 최종 음성 파형을 생성하며, 전체 구조를 통해 연산량 감소와 추론 속도 개선을 동시에 달성할 수 있다.

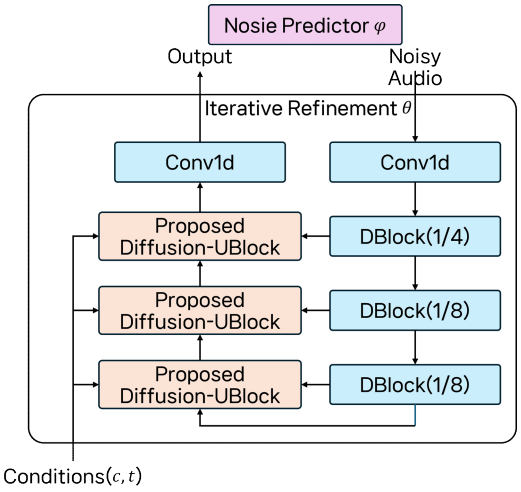


그림 2. 제안하는 모델이 적용된 보코더 프로세스
Fig. 2 Vocoder process with the proposed model applied

그림 3은 제안하는 Diffusion-UBlock 아키텍처다. 본 연구에서는 U-Block 내부에 Depthwise Separable LVC를 적용하여 연산 효율성을 고려한다. 먼저 업샘플링을 통해 해상도를 높여주고, Convolution의 입력 특징의 지역적 패턴과 Leaky ReLU(Rectified Linear Unit)의 비선형성을 통해 학습 성능을 향상한다.

이후 Depthwise Convolution과 Pointwise Convolution으로 구성된 Depthwise Separable LVC를 적용하여 연산량을 줄이면서도 특징 정보를 효과적으로 유지할 수 있도록 한다.

특히 Time-Aware Kernel Predictor는 입력 특징 x_t와 시간 임베딩 t를 기반으로 시점별 동적 커널 K_t를 생성하며, 식은 (1)과 같이 표현된다.

$$K_t = (x_t, t) \tag{1}$$

생성된 커널은 Depthwise Separable LVC에 적용되어 시간에 따라 변화하는 음성 신호의 특성을 반영한다. 이를 통해 기존의 고정 커널 기반 연산과 달리 시간에 적응적인 동적 필터링이 가능하며, 불필요한 연산을 줄이면서도 음성 신호의 시간에 따른 변동성을 반영할 수 있다.

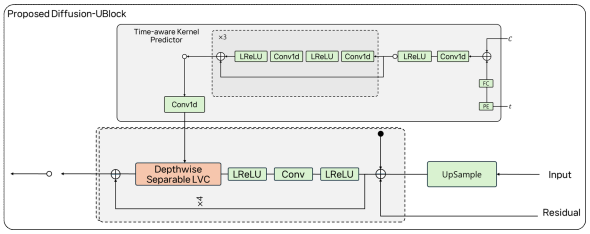


그림 3. 제안하는 Diffusion-UBlock 아키텍처
Fig. 3. Proposed Diffusion-UBlock architecture

3. Depthwise Separable LVC 설계

본 논문에서는 LVC와 Depthwise Separable Convolution을 결합한 Depthwise Separable LVC 구조를 제안한다. 해당 구조는 Depthwise LVC와 Pointwise LVC의 단계로 진행되며, 시간적인 흐름의 정보를 추가하기 위해 Time-Aware Kernel Predictor를 적용한다. Time-Aware Kernel Predictor는 시간 정보를 고려하여 커널 예측 구간에서 시점별로 적용할 동적 커널을 생성한다. 이 커널은 Depthwise Convolution을 적용해 각 채널의 시계열 상의 위치 정보를 반영한다.

그림 4는 Depthwise LVC의 과정을 나타낸다. 입력된 데이터는 패치 추출(Patch Extraction)을 통해 채널마다 작은 공간적 데이터를 추출되고, 채널 순서대로 나열되어 패치 벡터를 이룬다. 이후 내적(Dot Product)을 통해 채널마다 학습된 필터를 독립적으로 적용하여 특징을 추출하고, 이에 대한 채널별 결과를 각각 합산하여 하나의 스칼라(Scalar)값을 추출한다. 최종적으로 채널마다 하나의 스칼라 값들을 다시 하나의 벡터로 재조립하여 각 공간 위치별 채널 출력을 구성한다. 이 과

정은 기존의 공간적 Convolution보다 연산량을 크게 줄이면서도, 시간 축 상의 구조적 정보를 효과적으로 반영한다.

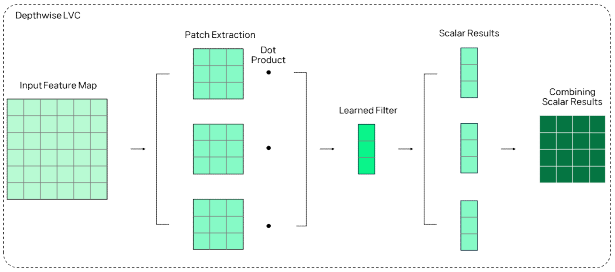


그림 4. Depthwise LVC의 과정
Fig. 4. Depthwise LVC's process

그림 5는 Pointwise LVC의 과정을 나타낸다. Depthwise LVC의 출력을 입력하며, 이때 게이트 생성 과정(Gating Process)과 값 생성 과정(Value Generation Process)으로 두 개의 서로 다른 가중치 행렬이 병렬로 곱해진다. 게이트 생성 과정은 게이트용 가중치 행렬을 이용해 채널마다 얼마나 통과할지를 결정하는 신호를 생성한다. 값 생성 과정은 값용 가중치 행렬을 이용해 입력 벡터의 선형 변환을 수행하고, ReLU 함수를 거쳐 값 벡터를 만든다. 이때, 공간적 윈도우는 1×1 크기로만 작용하므로 채널 간의 조합만 일어난다. 게이트 벡터와 값 벡터를 채널별로 곱해 출력한다. 이를 통해 값 벡터의 각 성분이 게이트의 강도만큼 조절된다.

이와 같이 Depthwise Separable LVC는 시계열 특징 추출, 조건 기반 활성화, 채널 간 통합을 모듈화함으로써, 기존 LVC 구조 대비 계산 효율성과 표현력을 동시에 확보할 수 있다.

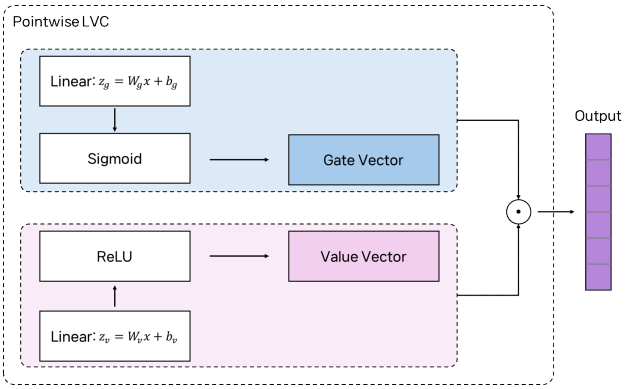


그림 5. Pointwise LVC의 과정
Fig. 5. Pointwise LVC's Process

Ⅳ. 실험 결과 및 고찰

1. 평가 지표

본 논문에서는 오디오의 품질, 음성합성의 속도를 기준으로 성능을 평가한다. 합성된 오디오 품질을 평가하기 위해서 STOI(Short-Time Objective Intelligibility)[31], PESQ(PerceptualEvaluation of Speech Quality)[32]를 활용한다. STOI는 음성 신호의 명료도를 평가하는 객관적인 지표로, 식은 (2)와 같다. 음성 신호를 짧은 시간 구간으로 나누어 분석하며, 원본 신호와 처리된 신호 간의 상관관계를 계산하여 명료도를 평가한다. PESQ는 음성 신호의 품질을 객관적으로 평가하는 지표로, 식은 (3)과 같다. 원본 신호와 처리된 신호를 비교하여 음질 왜곡과 간섭을 분석하고, 이를 기반으로 음질 점수를 산출한다. 두 성능 지표는 통계적인 비교를 위해 평균값을 사용하며, 기존의 모델과의 음질의 차이를 비교한다.

$$STOI = \frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\widetilde{d_{j,m}} \quad (2)$$

$$PESQ_{MOS} = 4.5 - 0.1D_{sym} - 0.0309D_{asym} \quad (3)$$

음성합성의 속도를 평가하기 위해서 RTF (Real Time Factor)[33]와 T_{utt}(Time per Utterance)[34]을 활용한다. RTF는 음성 인식 및 합성 시스템의 처리 속도를 평가하는 지표로, 식은 (4)와 같다. 처리 시간과 실제 시간의 비율을 나타낸다. RTF는 1보다 작을 때 시스템이 실시간보다 빠르게 데이터를 처리하며, 1보다 클 때는 느리게 처리한다는 것을 의미한다. T_{utt}은 한 문장을 합성하는 데 걸리는 시간 평균을 나타내는 지표로, 식은 (5)와 같다. 전체 문장을 합성하는 데 걸리는 총 시간을 총 문장의 개수로 나누어 한 문장의 평균 합성 시간을 구한다. 두 성능 지표는 통계적인 비교를 위해 평균값을 사용하며, 기존 모델들과의 음성합성 처리 속도를 비교하고 평가한다.

$$PTF = \frac{T_{processing}}{T_{audio}} \tag{4}$$

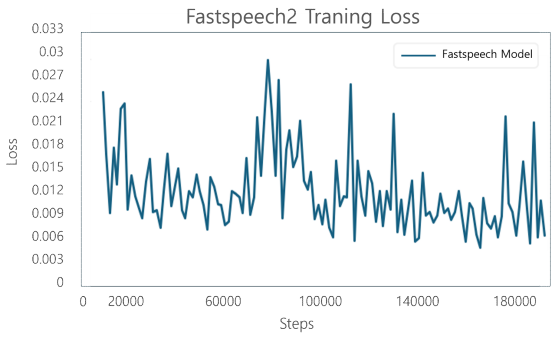
$$T_{utt} = \frac{1}{N}\sum_{i=1}^{N} t_i \tag{5}$$

2. 실험 결과

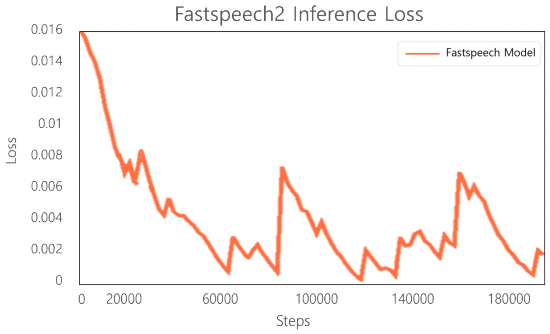
본 논문에서 제안하는 음성합성 모델은 LJ Speech 데이터 세트를 통해 평가한다. 먼저 제안하는 모델의 학습에 따른 성능을 비교 및 평가하고, 이후 다른 모델들과 성능을 비교 및 평가한다.

음성 파형 데이터를 시간 및 주파수 특징 벡터로 변환하기 위한 하이퍼파라미터(Hyperparameter)는 다음과 같다. 샘플링 레이트(Sampling Rate)는 24,000 Hz, 멜 스펙트로그램 빈(Bins) 수는 80으로 가장 표준화된 지표로 설정했다. Diffusion 모델의 노이즈 스케줄의 시간 인덱스 하이퍼파라미터는 다음과 같다. 노이즈 주입 및 제거의 단계의 총수인 타임스텝은 기본적인 파라미터를 기반으로 1,000으로 설정한다. 타임스텝에서 스케줄의 시작과 끝의 값을 나타낸 β_0와 β_T는 1.0e-6와 0.01으로 설정했다. 스케줄러는 RSqrt(Reciprocal Square-root-schedule), 학습률은 2e-4, 옵티마이저는 Adam으로 설정했다. 제안하는 모델의 성능 확인을 위해 Steps를 100,000, 200,000, Epoch를 50, 1,000으로 설정했다. 또한, FastSpeech2, Tacotron2, FastDiff-TTS[35], ProDiff-TTS[36]와 제안하는 모델의 성능을 비교 및 평가를 진행한다.

그림 6은 FastSpeech2의 Loss를 나타낸 그래프이며, 그림 7은 제안하는 모델의 Loss를 나타낸 그래프이다. (a)의 Training Loss 그래프를 비교했을 때, 두 모델 모두 Steps 100k에서 점차 진폭이 감소했다. 그러나 Steps 100k 단계 이후부터 FastSpeech2의 변화폭이 크게 변하는 것을 볼 수 있었고, 제안하는 모델은 안정적으로 감소하는 것을 볼 수 있었다. (b)의 Inference Loss를 비교했을 때, Fastspeech2는 초반에 0으로 안정적으로 수렴하고 있었으나 Steps 100k 단계 이후부터 진폭이 크게 변하는 모습을 보였다. 반면에 제안하는 모델은 초반에 진폭의 변화와 높게 나왔으나, Steps 100k 단계에서 평균 Loss가 저하되면서, Steps 200k 단계에서 0에 가깝게 수렴함을 보여준다. 이를 통해 FastSpeech2보다 제안하는 모델이 Loss 변동 폭 및 수렴 안정성이 높은 것을 확인할 수 있었다.



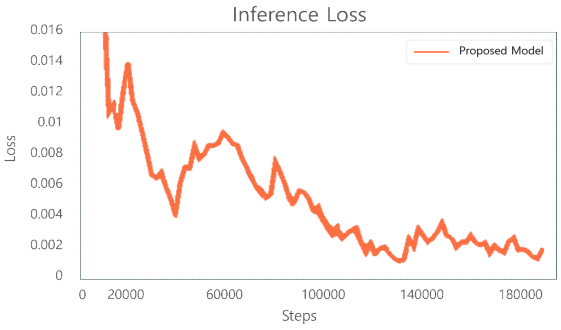
(a) FastSpeech2 Training Loss



(b) FastSpeech2 Inference Loss

그림 6. FastSpeech2의 Loss를 나타낸 그래프
Fig. 6. Loss of FastSpeech2

(a)Training Loss



(b)Inference Loss

그림 7. 제안하는 모델의 Loss를 나타낸 그래프
Fig. 7. Loss of the proposed model

표 1은 LJSpeech 데이터 세트에서 Steps 100,000, Epoch 1,000일 때 성능평가표이다. 품질 측면에서 PESQ는 제안하는 모델이 Tacotron2와 0.096의 차이로 낮은 성능을 보였으나, FastDiff-TTS와 ProDiff-TTS와 크게 차이는 나지 않았다. 또한, STOI에서 ProDiff-TTS와 0.001의 차이로 두 번째로 좋은 성능을 나타냈다. 이를 통해 노이즈 예측 기반 Diffusion 구조로 음성의 전반적인 자연스러움과 명료도를 효과적으로 보존하여 품질 면에서 크게 차이 나지 않은 것을 알 수 있었다

속도측면에서 RTF의 평균은 제안하는 모델이 가장 좋았던 FastSpeech2와 0.001의 차이로 두 번째 좋은 성능을 나타냈다. 그러나 T_{utt}에서는 FastSpeech2와 0.05의 차이로 가장 좋은 성능을 나타냈다. 이를 통해 Depthwise Separable LVC 구조를 통한 연산량을 줄이고, 채널 단위 연산 후 통합하는 방식으로 각 문장에 처리 속도가 높아진 것으로 보인다.

표1. LJSpeech 데이터 세트에서 Steps 100,000, Epoch 1,000일 때 성능평가표
Table 1. Performance evaluation with 100,000 steps and 1,000 epochs

Model	AVG STOI(↑)	AVG PESQ(↑)	AVG RTF(↓)	T_{utt}(↓)
Tacotron2	0.183	**1.141**	0.060	2.08
FastSpeech2	0.232	0.992	**0.010**	1.68
FastDiff-TTS	0.362	1.068	0.012	1.76
ProDiff-TTS	**0.364**	1.063	0.040	1.99
Proposed	0.363	1.045	0.011	**1.63**

표 2는 LJSpeech 데이터 세트에서 Steps 2

00,000, Epoch 1,000일 때 성능평가표이다. Steps의 추가를 통해 모든 모델의 성능이 향상되었다. 품질을 비교했을 때, PESQ는 가장 성능이 좋은 Tactron2와 차이의 폭이 크게 줄어든 것을 확인할 수 있었다. 또한, STOI는 여전히 크게 차이가 나지 않으면서, 두 번째로 좋게 추출된 것을 볼 수 있었다.

속도를 비교했을 때, RTF의 평균은 가장 좋았던 FastSpeech2와 0.001의 차이로 두 번째로 좋은 성능을 볼 수 있었다. 그러나 T_{utt}에서는 FastSpeech2와 0.11의 차이로 더 좋은 성능을 보였다. 이는 Steps의 수가 늘어남에 따라 Diffusion 모델의 특성에 따라 학습이 진행될수록 노이즈 제거 과정이 안정화되고, 음성 품질이 점진적으로 개선되기 때문으로 분석된다.

표 2. LJSpeech 데이터 세트에서 Steps 200,000, Epoch 1,000일 때 성능평가표
Table 2. Performance evaluation with 200,000 steps and 1,000 epochs

Model	AVG STOI(↑)	AVG PESQ(↑)	AVG RTF(↓)	T_{utt}(↓)
Tacotron 2	0.184	**1.143**	0.062	2.04
FastSpeech2	0.234	0.994	**0.010**	1.53
FastDiff-TTS	**0.364**	1.067	0.012	1.61
ProDiff-TTS	**0.364**	1.064	0.050	1.89
Proposed	0.363	1.053	0.011	**1.50**

품질 면에서 Steps에 따른 성능을 비교했을 때, 제안하는 모델은 기존의 모델들 성능 향상 폭보다 더 큰 폭으로 향상되었다. 특히 음성의 품질을 나타내는 PESQ에서는 다른 기존의 모델들보다 0.008이라는 성능 향상의 폭을 보여주었다. 한편, PESQ에서 여전히 기존 모델 대비 미세한 성능 차이가 존재한다. 이는 제안하는 모델이 연산 효율성을 고려한 구조를 적용함에 따라 일부 고주파 성분이나 세밀한 스펙트럼 정보 복원에서 제한이 발생할 수 있기 때문으로 분석된다.

T_{utt}에서는 0.13이 더 줄어들어 한 문장당 처리에 사용된 시간이 줄어들었다. 이는 Steps를 추가하는 것으로 문장의 단어와 지속 시간, 음의 높이 등의 위치적 정보와 시간적 정보가 잘 나타나는 것을 알 수 있었다.

제안하는 모델은 기존 Diffusion기반 모델의 단점인 속도 문제를 개선하는데 집중했다. 품질 면에서 가장 우수한 성능을 달성하지는 않았지만, 가장 좋은 성능을 보여준 FastDiff-TTS 및 ProDiff-TTS와 유사한 성능을 달성했다. 이는 제안 모델의 채널별 연산 후 통합 구조가 반복 학습을 통해 각 채널의 정보를 효율적으로 결합하면서 점진적으로 음성 품질을 개선할 수 있기 때문이다. 또한, 속도면에서 가장 우수한 성능 보여준 FastSpeech2와 비교했을 때 더 좋은 성능을 보였다. 특히, T_{utt}에서 가장 좋은 성능을 달성해 한 문장당 처리에 사용되는 연산량이 줄어든 것을 확인할 수 있었다. 이는 각 채널 연산 후 통합 방식을 통해 한 문장당 연산량이 줄어들면서, 학습 단계가 진행될수록 성능 향상 폭이 크게 나타났음을 확인할 수 있었다. 이를 통해 Steps는 음성 합성의 품질과 각 문장에 대한 합성 속도 개선에 기여함을 확인할 수 있으며, 결과적으로 전체 음성 합성 속도 향상에 긍정적인 영향을 미치는 것으로 판단된다.

V. 결 론

인공지능의 발달로 다양한 음성 모델의 연구들이 발표되고 있으며, 생성형 모델을 통한 음성합성 모델의 연구도 다양하게 진행되고 있다. Diffusion 모델을 적용한 음성합성 모델은 자연스러운 음성을 합성할 수 있고, 최근에는 모델 구조를 수정하여 속도의 단점을 보완하는 연구가 진행되고 있다.

본 논문에서는 속도 측면을 보완하기 위해 Depthwise Separable LVC를 적용한 Diff-TTS를 제안했다. 시간 의존적 특징을 더 효과적으로 추출할 수 있는 LVC와 Convolution의 연산량을 줄여 모델의 경량화를 위한 Depthwise Separable Convolution을 결합하여 사용했다. LJSpeech 데이터 세트를 사용하여 성능평가 및 다양한 모델과의 성능 비교를 분석했다.

제안하는 모델에서 문장 길이에 따른 음성 파형을 비교했을 때, 모두 원본과 유사하게 나타난 것을 확인할 수 있었다. Loss 변화 폭과 추세도 안정적으로 감소함을 확인했다. 또한, Steps가 늘어남에 따라 T_{utt}에서는 성능이 향상되었으며, 음성의 품질에서 STOI와 PESQ 성능이 향상되었다. 이를 통해, Steps가 각 문장에 대한 합성 속도와 음성 합성의 품질개선에 기여함을 확인할 수 있었다.

기존 모델과 제안하는 모델을 비교했을 때, STOI와 RTF는 가장 좋았던 모델과 비교하여 0.001의 미미한 차이로 나타났다. 그러나 PESQ는 다른 모델에 비해 0.008 향상되었고, T_{utt}에서는 기존 모델에 비교해 0.03 차이로 가장 좋은 성능을 나타냈다. 이를 통해 제안하는 모델이 음성의 품질을 유지하면서 하나의 문장 처리에 사용되는 연산량이 줄어든 것을 확인할 수 있었다. 이는 채널별 연산 후 통합 방식이 음성합성의 속도와 품질 면에서 모두 개선한 것을 입증한다.

본 논문에서는 Depthwise Separable LVC를 적용한 Diff-TTS를 구현하고 실험을 통해 음성합성 처리 속도를 개선 가능성을 확인했다. 향후 제안하는 모델의 추가적인 조정을 통해 입력되는 대용량의 문장들을 처리 속도를 높일 수 있을 것으로 생각된다. 또한, 추가적인 음성 데이터 세트를 활용하면 합성된 음성의 품질 성능 향상도 얻을 수 있을 것으로 기대된다.

REFERENCES

[1] P. Taylor, "Text-to-speech synthesis," *Cambridge university press*, 2009.

[2] H. Zen, K. Tokuda, & A. W. Black, "Statistical parametric speech synthesis," *speech communication*, pp. 1039 - 1064, 2009.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, & T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," *In Sixth European Conference on Speech Communication and Technology*, pp. 2347－2350, 1999.

[4] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, & K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, pp. 1234 - 1252, 2013.

[5] C. Manning, & H, Schutze, "Foundations of statistical natural language processing," *MIT press*, 1999.

[6] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, & T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," *In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Proceedings, pp. 1315 - 1318, 2000.

[7] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, pp. 349 - 353, 2006.

[8] X. Wang, J. L. Trueba, S. Takaki, L. Juvela, & J. Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-networkbased speech synthesis," *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.

4804 - 4808, 2018.
[9] H. Zen, A. Senior, & M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” *In 2013 ieee international conference on acoustics, speech and signal processing*, pp. 7962 - 7966, 2013.
[10] A. Wali, Z. Alamgir, S. Karim, A. Fawaz, M. B. Ali, M. Adan, & M. Mujtaba, “Generative adversarial networks for speech processing : A review,” *Computer Speech and Language, vol. 72(article 101308)*, 2022
[11] H. Jonathan, J. Ajay, & A. Pieter, “Denoising Diffusion Probabilistic Models,” *arXiv preprint,* arXiv:2006.11239, 2020
[12] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, & L. Xie. “Multi-band melgan: Faster waveform generation for high-quality text-to-speech,” arXiv preprint, arXiv:2005.05106, 2020.
[13] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, & N. S. Kim, “Diff-tts: A denoising diffusion model for text-to-speech,” arXiv preprint, arXiv:2104.01409, 2021.
[14] C. Zhang, C. Zhang, & S. Zheng, “A Survey on Audio Diffusion Models: Text To Speech Synthesis & Enhancement in Generative AI,” arXiv preprint, arXiv:2303.13336, 2023
[15] K. Shen, Z. Ju, X. Tan, et al., “NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers,” arXiv preprint, arXiv:2304.09116, 2023
[16] Z. Ju, Y. Wang, K. Shen, et al., “NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models,” arXiv preprint, arXiv:2403.03100v3, 2024
[17] C. François, “Xception: Deep Learning with Depthwise Separable Convolutions,” arXiv preprint, arXiv:1610.02357, 2017.
[18] Z. Zhen, W. Jianzong, C. Ning, & X. Jing, “LVCNet: Efficient Condition-Dependent Modeling Network for Waveform Generation,” arXiv preprint, arXiv:2102.10815, 2021.
[19] J. Schmidhuber, F. f. Informatik, T. U. München, “LONG SHORT-TERM MEMORY,” *IEEE Customer Center*, pp. 1735 - 1780, 1997.
[20] S. Jonathan, P. Ruoming, J. Ron, et al., “TACOTRON2: Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” arXiv preprint, arXiv:1712.05884, 2018.
[21] J. Kim, S. Kim, J. Kong, S. Yoon, “Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search,” arXiv preprint, arXiv:2005.11129, 2020.
[22] F. Yu, V. Koltun, “Multi-Scale Context Aggregation by Dilated Convolutions,” arXiv preprint, arXiv:1511.07122, 2016.
[23] S. Bai, J. Z. Kolter, V. Koltun, “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling,” arXiv preprint, arXiv:1803.01271, 2018.
[24] P. Hsu, C. Wang, A. T. Liu, & H. Lee, “Towards Robust Neural Vocoding for Speech Generation: A Survey,” arXiv preprint, arXiv:1912.02461, 2020.
[25] A. van-den-Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, & K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” arXiv preprint, arXiv:1609.03499, 2016.
[26] Y. Ryuichi, S. Eunwoo, J.M. Kim, “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” arXiv preprint, arXiv:1910.11480, 2020.
[27] R. Yi, H. Chenxu, T. Xu, Q. Tao, Z. Sheng, Z. Zhou, & L. Tie-Yan, “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech,” arXiv preprint, arXiv:2006.04558, 2020.
[28] K. Ito, L. Johnson, “The Linda Johnson Speech Dataset,” *Keith Ito, https://paperswithcode.com/dataset/ljspeech*, 2025 (Accessed Jun., 23).
[29] T. Alessio, B. Alessio “Multilingual MFA: Forced Alignment on Low-Resource Related Languages,” arXiv preprint, arXiv:2504.07315, 2025.
[30] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, et al., “The Kaldi Speech Recognition Toolkit,” *In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 1 - 4, 2011.
[31] C. H. Taal, R. C. Hendriks, R. Heusdens, & J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214-4217, 2010.
[32] J. G. Beerends, A. P. Hekstra, A. W. Rix, M. P. Hollier, “Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-band,” T*elephone Networks and Speech Codecs Journal of the Audio Engineering Society*, pp. 749-752, 2002.
[33] OpenVoice Tech, Real-time factor, *OpenVoice Tech Wiki*,

https://openvoice-tech.net/index.php/Real-time-factor, (Accessed Jun., 23, 2025).
[34] W. Nigel, V. Alejandro, "Studies in the Use of Time Into Utterance as a Predictive Feature for Language Modeling," *Departmental Technical Reports (CS)*, pp. 1-7, 2010.
[35] R. Huang, M. W. Y. Lam, J. Wang, D. Su, D. Yu, Y. Ren, & Z. Zhao, "FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis," arXiv preprint, arXiv:2204.09934, 2022.
[36] H. Rongjie, Z. Rongjie, L. Huadai, L. Jinglin, C. Chenye, R. Yi, "ProDiff:Progressive Fast Diffusion Model For High-Quality Text-to-Speech," arXiv preprint, arXiv:2207.06389, 2022.

저 자 소 개

박수빈(준회원)

2024년 순천대학교 대학원 멀티미디어공학과 졸업(공학사)
2026년 순천대학교 대학원 멀티미디어공학과 졸업(공학석사)
<주관심분야 : 음성합성, 딥러닝, 디퓨전, 머신러닝>

김진성(준회원)

2021년 국립순천대학교 컴퓨터공학과 졸업(공학사).
2023년 국립순천대학교 멀티미디어공학과 졸업(공학석사).
2023년 ~현재 국립순천대학교 대학원 IT융합바이오융합시스템전공 박사과정.
<주관심분야 : 컴퓨터비전, 딥러닝, 시맨틱 세그멘테이션, 빅데이터 분석, 머신러닝>

정세훈(종신회원)

2012년 순천대학교 대학원 멀티미디어공학과 졸업(공학석사).
2017년 순천대학교 대학원 멀티미디어공학과 졸업(공학박사).
2018년 영산대학교 빅데이터융합전공 조교수
2020년 안동대학교 창의융합학부 조교수
2022년 ~현재 순천대학교 컴퓨터공학과 부교수
<주관심분야 : 블록체인, 딥러닝, 빅데이터 분석>

심춘보(정회원)

1996년 전북대학교 컴퓨터공학과 졸업(공학사)
1998년 전북대학교 대학원 컴퓨터공학과 졸업(공학석사).
2003년 전북대학교 대학원 컴퓨터공학과 졸업(공학박사).
2005년 ~현재 순천대학교 인공지능공학부 교수
<주관심분야 : 빅데이터, 블록체인, 딥러닝, 생성모델, 자연어 처리, 강화학습>