고령 화자 음성데이터 기반 한국어 음성합성 모델 파인 튜닝

김영주[1], 조광문[2*], 이도현[3]
[1]국립목포대학교 컴퓨터공학과 강사, [2]국립목포대학교 컴퓨터학부 교수, [3]국립목포대학교 스마트비즈니스학과 학생

Fine-tuning of Korean Text-to-Speech Model Based on Elderly Speaker Voice Data

Yeongju Kim[1], Kwangmoon Cho[2*], Do Hyun Lee[3]
[1]Lecturer, Dept. Computer Engineering, Mokpo National University
[2]Professor, School of Computer Science and Engineering, Mokpo National University
[3]Student, Dept. of Smart Business, Mokpo National University

요 약 본 연구는 60대에서 90대에 이르는 고령 화자의 한국어 음성 데이터를 활용하여, 제한된 데이터 환경에서도 효과적으로 학습 가능한 음성합성(Text-to-Speech, TTS) 모델 구축 방안을 제안한다. 총 50명의 고령 화자로부터 약 250분 분량의 음성 데이터를 수집하였으며, 화자당 평균 5분 수준의 소량 데이터를 기반으로 XTTS(Cross-lingual Text-to-Speech) 모델의 fine-tuning 기법을 적용하였다. 데이터 전처리 단계에서는 Whisper large-v3 모델을 활용한 자동 음성 인식(ASR)과 음성 활동 감지(VAD)를 통해 발화 구간과 전사 품질을 정제하였다. Mixed Precision 학습과 CosineAnnealing 스케줄러를 적용하여 학습 효율과 안정성을 향상시켰다. 실험 결과, 최적의 하이퍼파라미터 설정 하에서 대부분의 화자에서 낮은 Word Error Rate(WER), Character Error Rate(CER)를 달성하였으며, 초기 학습 단계에서 높은 오류율을 보인 일부 화자에 대해서도 재학습을 통해 성능이 크게 개선됨을 확인하였다. 본 연구는 고령 화자의 음성 특성을 반영한 한국어 TTS 시스템 구축 가능성을 제시하며, 화자당 데이터가 극히 제한된 환경에서도 적용 가능한 효율적인 Few-shot 학습 방법론을 제공한다.

주제어 : 음성합성, Text-to-Speech, XTTS, Fine-tuning, 고령 화자, Few-shot Learning

Abstract This study proposes an effective method for building a Text-to-Speech (TTS) model in limited data environments using Korean voice data from elderly speakers aged 60 to 90. We collected approximately 250 minutes of voice data from 50 elderly speakers (25 males and 25 females), applying fine-tuning techniques with the XTTS (Cross-lingual Text-to-Speech) model based on an average of 5 minutes of data per speaker. In the data preprocessing stage, we refined speech segments and transcription quality through Automatic Speech Recognition (ASR) using the Whisper large-v3 model and Voice Activity Detection (VAD). We improved training efficiency and stability by applying Mixed Precision learning and CosineAnnealing scheduler. Experimental results demonstrate that with optimal hyperparameter settings, most speakers achieved low Word Error Rate (WER) and Character Error Rate (CER). Even for speakers who initially showed high error rates during the initial training phase, performance was significantly improved through retraining. This study presents the feasibility of building a Korean TTS system reflecting the voice characteristics of elderly speakers and provides an efficient Few-shot learning methodology applicable in environments with extremely limited data per speaker.

Key Words : Text-to-Speech, TTS, XTTS, Fine-tuning, Elderly Speaker, Few-shot Learning

본 과제(결과물)는 2025년도 교육부 및 전라남도의 재원으로 전라남도RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다.(2025-RISE-14-001)
*교신저자 : 조광문(ckmoon@mnu.ac.kr)
접수일 2025년 12월 31일 수정일 2026년 01월 15일 심사완료일 2026년 02월 20일

1. 서론

음성합성(Text-to-Speech, TTS) 기술은 텍스트를 자연스러운 음성으로 변환하는 기술로, 최근 딥러닝 기반의 End-to-End 모델의 발전으로 인간 수준의 자연스러운 음성 생성이 가능해졌다[1,2]. Tacotron[3], FastSpeech[4], VITS[5] 등의 모델이 대표적이며, 최근에는 YourTTS[6], XTTS[7] 등 다국어를 지원하고, AdaSpeech 시리즈[8-10] 등 소량의 데이터로도 새로운 화자에 적응할 수 있는 화자 적응(speaker adaptation) 기능을 갖춘 모델들이 활발히 연구되고 있다.

그러나 기존 TTS 연구는 주로 청년 및 중장년층의 명료한 음성을 중심으로 진행되어, 고령 화자의 음성 특성을 충분히 반영하지 못하는 한계가 있다. 고령 화자의 음성은 음도 변화의 감소, 발성 떨림(tremor), 기식음(breathiness) 증가 등의 특징을 보이며[11], 이를 학습하기 위해서는 특화된 데이터셋과 방법론이 필요하다. 또한 일반적으로 고품질 TTS 모델 학습에는 화자당 수십 시간의 음성 데이터가 필요하지만[12], 실제 응용 환경에서는 충분한 데이터 확보가 어려운 경우가 많다.

본 연구는 이러한 문제를 해결하기 위해 60대에서 90대까지의 고령 화자 50명으로부터 화자당 평균 5분 이내 수집한 음성 데이터를 활용하여, 제한된 데이터 환경에서도 고품질 한국어 TTS 모델을 구축하는 방법론을 제시한다.

제안하는 방법은 다음과 같이 구성된다. 첫째, 고령 화자 한국어 TTS 데이터셋의 수집, 전처리, 관리를 위한 체계적인 방법론을 제시, 둘째, XTTS 모델의 fine-tuning과 zero-shot 기능을 활용한 Few-shot 학습 전략을 제안, 셋째, 실시간 학습 모니터링과 하이퍼파라미터 최적화를 통한 성능 개선 방법을 제시, 넷째, WER과 CER 기반의 정량적 평가, 다섯째, 학습 실패 사례의 원인 분석과 체계적인 재학습 프로세스를 제안한다.

2. 이론적 배경

2.1 딥러닝 기반 음성합성 모델

딥러닝 기반 TTS 시스템은 크게 autoregressive 모델과 non-autoregressive 모델로 분류된다. Tacotron 2[3]는 attention 기반의 sequence-to-sequence 모델로 높은 품질의 음성을 생성하지만, 추론 속도가 느린 단점이 있다. 이를 개선한 FastSpeech[4]와 FastSpeech 2[2]는 병렬 처리가 가능하여 실시간 응용에 적합하다. 최근에는 End-to-End 학습이 가능한 VITS(Variational Inference with adversarial learning for end-to-end Text-to-Speech)[5]가 제안되어, vocoder 없이도 고품질의 음성을 직접 생성할 수 있게 되었다. VITS는 Variational Autoencoder(VAE)와 GAN을 결합하여 학습 효율과 음성 품질을 모두 향상시켰다.

2.2 다국어 및 화자 적응 TTS

다국어 TTS 시스템은 여러 언어의 음성을 단일 모델로 학습하여 언어 간 지식 전이(transfer learning)를 가능하게 한다. YourTTS[6]는 zero-shot 다국어 TTS를 가능하게 하여, 새로운 화자의 소량 데이터만으로 해당 화자의 음성을 합성할 수 있다. .XTTS[7]는 대규모 다국어 zero-shot TTS 모델로, 다국어 지원과 함께 fine-tuning 및 zero-shot 학습을 모두 지원한다. 특히 5분 이상의 음성 데이터로 fine-tuning이 가능하여, 제한된 데이터 환경에서도 효과적인 화자 적응이 가능하다는 장점이 있다. IMS-Toucan[13]은 메타 학습을 통해 7000개 이상의 언어를 지원하는 대규모 다국어 TTS를 구현하였다.

2.3 Few-shot 및 Zero-shot TTS

Few-shot TTS는 소량의 데이터만으로 새로운 화자의 음성을 학습하는 기술이다. Meta-learning 기반의 접근[14]과 speaker embedding을 활용한 방법[15]이 대표적이다.AdaSpeech[8]는 conditional layer normalization을 통해 화자 적응 성능을 향상시켰으며, AdaSpeech 2[9]는 전사되지 않은 데이터로도 안정적인 학습이 가능하도록 개선되었다. AdaSpeech 4[10]는 zero-shot 시나리오에서 화자 특성을 basis vector로 분해하여 일반화 성능을 향상시켰다. USAT[16]는 zero-shot과 few-shot 적응을 통합한 범용 화자 적응 프레임워크를 제시하였다.Zero-shot TTS는 학습 단계에서 보지 못한 화자의 음성도 reference 음성만으로 합성 가능하게 하는 기술로, speaker encoder를 활용한 방법이 주로 사용된다[17].

2.4 고령 화자 음성 연구

고령 화자의 음성은 생리학적 노화로 인한 특수한 특성을 보인다. 성대의 경직, 호흡 조절 능력 저하, 조음 기관의 정밀도 감소 등으로 인해 음성의 떨림, 기식음 증가, 발화 속도 변화 등이 나타난다[11].ASR 분야에서는

고령 화자 음성 인식 성능 향상을 위한 연구가 진행되었으나[18], TTS 분야에서는 상대적으로 연구가 부족한 실정이다. 특히 한국어 고령 화자 음성합성 연구는 매우 제한적이어서, 본 연구가 이 분야의 기초 연구로서 의의를 갖는다.

3. 연구방법

3.1 데이터셋 구축

본 연구에서는 60대에서 90대까지의 고령 화자 50명(남성 25명, 여성 25명)으로부터 음성 데이터를 수집하였다. 화자당 약 30문장, 5분 분량의 음성을 녹음하였으며, 전체 데이터셋은 1,500문장, 총 250분, 약 1.7GB 규모이다.

데이터는 연령대(60대, 70대, 80대, 90대)와 성별에 따라 디렉토리를 분류하여 체계적으로 관리하였다. 각 화자별로 메타데이터 파일을 작성하였으며, 모든 음성은 조용한 환경에서 녹음되었고 샘플링 레이트 22.05kHz, 16-bit PCM WAV 형식으로 저장하였다.

```
sentence01.wav | 안녕하세요.
sentence02.wav | 반갑습니다.
sentence03.wav | 오늘 날씨가 좋습니다.
```

[Fig. 1] Voice data collection and metadata format

3.2 모델 선정

본 연구에서는 화자당 5분이라는 제한된 데이터 환경을 고려하여 fine-tuning과 zero-shot 기능을 모두 지원하는 XTTS 모델을 선정하였다. XTTS는 5분 이상의 소량 데이터만으로도 효과적인 fine-tuning이 가능한 Few-shot 학습을 지원하며, 한국어를 포함한 다국어 환경에서 안정적으로 작동한다. 또한 fine-tuning이 충분하지 않을 경우 zero-shot 모드로 전환하여 사용할 수 있어 학습 실패에 대한 대체 방안을 제공한다.

또한, Alltalk TTS v2 프레임워크를 활용하여 XTTS-v2 모델의 fine-tuning을 수행하였으며, 웹 기반 사용자 인터페이스를 통해 학습 과정을 효율적으로 관리하였다.

3.3 학습 환경 구성

하드웨어는 NVIDIA Tesla V100-PCIE-32GB×2 GPU, Intel Xeon Gold 6136×2 CPU(24코어/48스레드), RAM 251GB를 사용하였다. 소프트웨어는 Linux 운영체제, Python 3.11, Miniconda 가상환경, Alltalk TTS v2 프레임워크를 사용하였으며, SSH를 통한 원격 접속과 웹 UI를 통해 학습을 진행하였다.

3.4 데이터 전처리

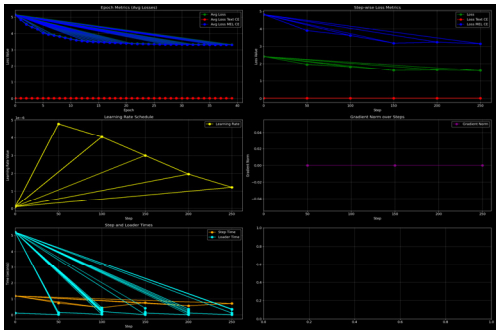
Whisper large-v3 모델을 사용하여 음성 파일로부터 텍스트를 자동 추출하였다. Mixed Precision(FP32와 FP16 혼용)을 적용하여 학습 효율을 향상시켰다[19]. Voice Activity Detection(VAD)을 활성화하여 음성이 포함된 구간만 자동으로 추출하였으며, Min Audio Length 2초, Max Audio Length 12초로 설정하였다. 전체 데이터를 학습 데이터와 검증 데이터로 분할하였으며, 데이터 양이 제한적임을 고려하여 검증 데이터 비율을 5%로 최소화하였다. XTTS-v2는 한국어를 포함한 다국어가 사전 학습된 토크나이저를 내장하고 있으므로, 별도의 BPE 토크나이저 학습은 수행하지 않았다.

3.5 모델 Fine-tuning

GPU 서버의 사양과 데이터의 특성을 고려하여 위와 같이 하이퍼파라미터를 설정하였다. 주요 설정은 〈Table 1〉과 같다. Learning Rate는 데이터가 제한적이므로 작은 값(5e-6)을 사용하여 과적합을 방지하고 안정적인 학습을 유도하였다. CosineAnnealing 스케줄러는 학습률을 코사인 함수 형태로 감소시켜, 적은 데이터에서 과적합을 방지하는 데 효과적이다[20]. AdamW는 weight decay regularization을 적용하여 일반화 성능을 향상시키며, 현재 대부분의 음성/언어 모델에서 표준으로 사용된다[21].

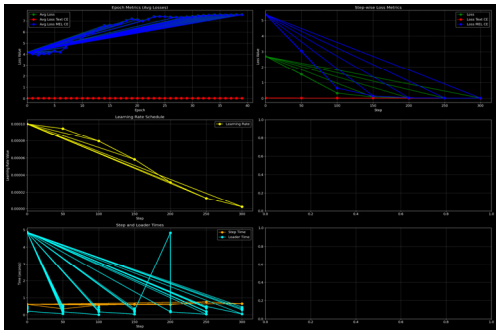
〈Table 1〉 Hyperparameter configuration

Parameter	Value	Description
Model	XTTS-v2	Fine-tuning target model
Learning Rate	5e-6	Stable learning with small rate
LR Scheduler	Cosine Annealing	Overfitting prevention
Optimizer	AdamW	Improved generalization
Epochs	40	Sufficient training iterations
Batch Size	8	GPU VRAM consideration
Gradient Accumulation	2	Effective batch size 16
Workers	8	Parallel data loading
Max Audio Size	15 sec	Include most sentences



[Fig. 2] Normal pattern

학습 과정에서 TensorBoard를 통해 Average Loss MEL CE, Training Loss, Validation Loss, Learning Rate Schedule을 실시간으로 모니터링 하였다. Fig. 2는 정상 패턴으로 Avg Loss MEL CE 지수가 점점 하강한다. 반면, Fig. 3은 비정상 패턴으로 Avg Loss MEL CE 지수가 점점 상승한다. 비정상적인 패턴(Loss 상승, 발산, 과적합)이 관찰되는 경우 학습을 중단하고 Learning Rate를 조정하여 재학습을 진행하였다.



[Fig. 3] Abnormal pattern

3.6 음성 합성 및 평가

Fine-tuning이 완료된 모델로 음성을 합성할 때, Language는 Korean, Temperature 0.75, Repetition Penalty 10, Speed 1.0x로 설정하였다.

모델의 성능을 객관적으로 평가하기 위해 WER(Word Error Rate)과 CER(Character Error Rate)을 사용하였다. WER은 음성 인식 결과와 원본 텍스트 간의 단어 수준 오류율을, CER은 문자 수준의 오류율을 측정한다.

$$WER = (S + D + I) / N \times 100\%$$
$$CER = (S + D + I) / N \times 100\%$$

여기서 S는 Substitution(치환), D는 Deletion(삭제), I는 Insertion(삽입), N은 원본 단어 또는 문자 총 개수이다.

평가 프로세스는 다음과 같은 순서로 진행된다. 첫째, Fine-tuning된 모델로 각 화자의 테스트 문장을 합성한다. 둘째, 합성된 음성을 Whisper large-v3 모델로 텍스트로 변환한다. 셋째, 원본 텍스트와 인식된 텍스트를 전처리한다. 넷째, WER과 CER을 계산한다. 다섯째, 기준치 이상의 오류율을 보이는 화자에 대해 재학습을 진행한다. 본 평가 방식은 Whisper 모델을 통한 재인식 기반으로 수행되어, TTS 출력이 Whisper에 최적화된 패턴을 보일 경우 실제 음질보다 오류율이 과소 추정될 수 있다는 한계점이 있다.

4. 실험 및 결과

4.1 Fine-tuning 효과 검증

fine-tuning의 효과를 검증하기 위해, 동일 화자 및 동일 테스트 문장에 대해 XTTS zero-shot (fine-tuning 없음)과 fine-tuned 모델의 성능을 비교하였다. 평가는 무작위 추출한 화자 6명을 대상으로 수행하였으며, 일부는 높은 오류율을 보인 재학습 대상 화자도 포함되어 있다. 6명 화자의 평균 WER과 CER을 측정하였다.

〈Table 2〉 Comparison of XTTS zero-shot and fine-tuned models

Condition	Avg WER	Avg CER
XTTS zero-shot (Before fine-tuning)	41.9%	38.3%
XTTS fine-tuned (After fine-tuning)	19.7%	3.3%

실험 결과, fine-tuning을 통해 WER과 CER이 개선되어, 화자당 평균 5분 정도의 소량 데이터로도 효과적인 화자 적응이 가능함을 확인하였다.

4.2 전체 화자 성능 분석

총 50명의 화자에 대해 fine-tuning을 수행하였으며, 각 화자별로 WER과 CER을 측정하여 성능을 평가하였다. 초기 학습에서 높은 오류율을 보인 화자들은 하이퍼파라미터를 조정하여 재학습을 진행하였다.

〈Table 3〉 WER and CER measurement results for 50 speakers

Speaker Group	WER Range	CER Range	Number of speakers
Excellent	0%	0%	47
Retraining Required	36.36-45.45%	0-11.43%	3

대부분의 화자(94%)에서 WER 0%, CER 0%를 달성하여 매우 우수한 성능을 보였다. YourTTS[6] 및 AdaSpeech[8] 등 기존 Few-shot TTS 연구에서 보고된 WER은 일반 화자 기준 5~15% 수준이며, 고령 화자 대상 연구는 상대적으로 희소하다. 본 연구의 WER 0% 달성은 fine-tuning 기반의 강한 화자 적응 효과에 기인하며, 이는 동시에 과적합 가능성도 내포한다.

그러나 3명의 화자(speaker24, speaker25, speaker27)는 초기 학습에서 높은 WER을 보여 재학습이 필요하였다. 화자에 대한 자세한 정보는 〈Table 4〉와 같다.

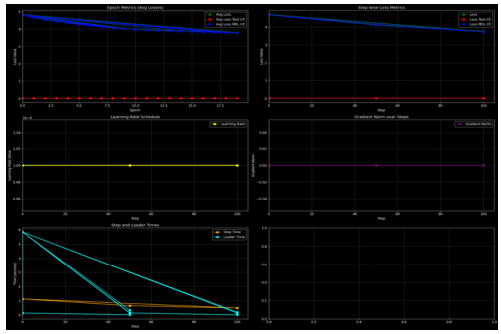
〈Table 4〉 Sample speaker information

Speaker Group	Speaker number	Number of sentences	Audio length	Age range	Gender
Retraining Required	24	30	6 min 13 s	60s	male
	25	30	5min 33s	60s	male
	27	30	5min 33s	60s	male
Excellent	44	30	5min 34s	70s	male
	49	30	5min 42s	80s	female
	50	30	5min 3s	90s	female

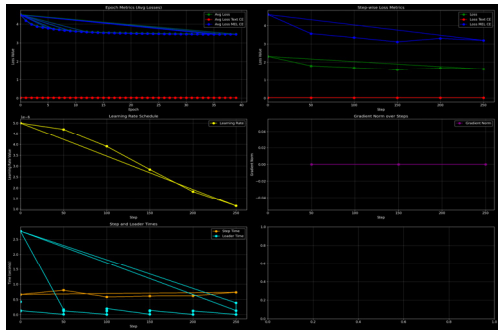
4.3 재학습 사례 분석

높은 WER을 보인 화자들의 공통적인 문제점은 다음과 같다. 첫째, Epochs 부족으로 20 epochs로 학습하여 충분한 수렴이 이루어지지 않았다. 둘째, 과도한 Evaluation Data Split으로 15%로 설정하여 학습 데이터가 과도하게 감소하였다.

재학습 시 Epochs를 40으로, Evaluation Data Split을 5%로 조정하였다. speaker24의 경우, 초기 학습에서 WER 45.45%, CER 11.43%를 보였으나, 재학습 후 WER 0%, CER 0%를 달성하였다. Fig 4는 Speaker24의 재학습 전 패턴과 Fig 5는 Speaker24의 재학습 후 패턴이다.



[Fig. 4] Speaker 24's pattern before relearning



[Fig. 5] Speaker 24's pattern after relearning

4.4 Learning Rate 영향 분석

동일한 화자, 동일한 파라미터에서 Learning Rate만 변경하여 실험한 결과는 〈Table 5〉와 같다.

〈Table 5〉 Impact of learning rate on training stability

Learning Rate	Avg Loss MEL CE Trend	Training Stability
5e-6	Continuous decrease	Stable
1e-4	Increase or divergence	Unstable

Learning Rate가 너무 높을 경우(1e-4) Loss가 발산하여 학습이 실패하였으며, 적절한 Learning Rate(5e-6)를 사용할 경우 안정적인 학습이 가능하였다.

4.5 연령대별 및 성별 성능 분석

〈Table 6〉과 같이 재학습 후 전체 50명 화자의 평균은 WER은 6.53%, 표준편차 2.03%를 기록하였다. 이는 기존 Few-shot TTS 연구의 5~15% 범위 내에 위치하며 안정적인 성능을 보였다. 연령대별로는 60대 6.47%, 70대 6.40%, 80대 6.80%, 90대 9.25%로 60~70대 구

간에서는 차이가 미미하였다. 성별로는 남성 6.61%, 여성 6.27%로 성별에서도 차이가 미미하게 관찰되었다.

〈Table 6〉 Performance analysis by age group and gender

Category	Group	Speakers	Avg WER	WER range
Age	60s	37	6.47%	2.86%~9.86%
	70s	10	6.40%	3.49%~9.61%
	80s	2	6.80%	5.71%~7.89%
	90s	1	9.25%	9.25%
Gender	Male	37	6.61%	2.86%~9.86%
	Female	13	6.27%	3.27%~9.79%

4.6 학습 시간 분석

화자당 평균 학습 시간은 데이터셋 구축 약 5-10분, Fine-tuning(40 epochs) 약 2-3시간, 음성 합성 및 평가 약 10-15분으로, 전체 소요 시간은 화자당 약 3-4시간이며, GPU 2개를 사용하여 병렬 학습이 가능하므로, 전체 50명의 화자를 약 75-100시간 내에 학습 완료하였다.

5. 결론 및 향후 연구

본 연구는 60대에서 90대까지의 고령 화자 50명의 음성 데이터를 활용하여, 제한된 데이터 환경(화자당 5분)에서도 고품질 한국어 TTS 모델을 구축할 수 있음을 실증하였다. XTTS 모델의 fine-tuning 기법과 체계적인 하이퍼파라미터 최적화를 통해 대부분의 화자에서 낮은 WER과, CER을 달성하였으며, 초기 학습 실패 사례에 대해서도 재학습을 통해 성능을 완전히 회복할 수 있었다.

본 연구의 주요 기여는 다음과 같다. 첫째, 고령 화자 한국어 TTS 데이터셋 구축을 위한 체계적인 방법론을 제시하였다. 둘째, Few-shot 학습 환경에서의 효과적인 fine-tuning 전략을 개발하였다. 셋째, 실시간 학습 모니터링과 체계적인 재학습 프로세스를 수립하였다. 넷째, WER과 CER 기반의 정량적 평가를 통한 객관적 품질 관리 방법을 제안하였다.

본 연구의 한계점으로 첫째, 평가 방식에서 Whisper 모델을 통한 재인식 기반으로 수행되어, TTS 출력이 Whisper에 최적화된 패턴을 보일 경우 실제 음질보다 오류율이 과소 추정될 수 있다는 한계점이 있다. 둘째, WER 및 CER을 통한 정량적 평가를 수행하였으나, 실제 음성 품질의 자연스러움, 화자 유사성을 반영하는 MOS(Mean Opinion Score) 평가는 수행되지 않았다. 셋째, 화자당 30문장·5분 수준의 낭독체 데이터만을 수집하여 발화 다양성이 제한되었으며, 고령 화자 특유의 발성 떨림(tremor)·기식음(breathiness) 등의 음성 특성이 fine-tuning 과정에서 노이즈로 처리될 가능성이 있다. 넷째, 단조로운 낭독체 문장으로만 구성되어 감정 표현의 다양성이 부족하다는 한계가 있다.

향후 연구는 본 연구의 한계를 극복하는 방향으로 확장될 필요가 있다. 첫째, 데이터 측면에서 발화 문장 수를 100문장 이상으로 확대하고 자유 발화 및 대화체 데이터를 추가 수집하여 데이터셋의 다양성을 강화할 필요가 있다. 또한 100세 이상 초고령자와 지역 방언 화자를 포함하여 연령대 및 언어 변이의 범위를 넓혀야 한다. 둘째, 고령 화자 특유의 발음 불명확성과 발성 떨림·기식음 등의 음성 특성을 명시적으로 모델링하는 고령 화자 특화 전처리 및 학습 전략 개발이 요구된다. 셋째, 기쁨·슬픔·중립 등 다양한 감정 레이블을 포함한 감정 TTS(Emotional TTS)로 확장하여 더욱 자연스럽고 풍부한 음성 합성이 가능하도록 해야 한다. 넷째, 실시간 서비스 적용을 위한 추론 속도 최적화와 함께, MOS 등 주관적 청취 평가를 도입하여 정량적 지표와 주관적 음질 평가를 결합한 종합적인 평가 체계를 구축해야 한다. 마지막으로, 학습 모니터링과 재학습 프로세스를 자동화하는 파이프라인을 구축함으로써 대규모 화자에 대한 확장성을 갖춘 지능형 TTS 시스템으로 발전시킬 수 있을 것이다.

본 연구는 고령화 사회에서 고령자 친화적 음성 기술 개발의 필요성에 부응하며, 제한된 데이터 환경에서도 효과적인 TTS 모델 학습이 가능함을 보여주었다. 특히 화자당 5분이라는 최소한의 데이터로 완벽한 성능을 달성한 것은, 데이터 수집이 어려운 소수 집단에 대해서도 고품질 음성 기술을 제공할 수 있는 가능성을 제시한다.

REFERENCES

[1] A.van den 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," in Proceedings of the 9th ISCA Speech Synthesis Workshop, 2016, pp.125-125.

[2] Y.Ren, C.Hu, X.Tan, T.Qin, S.Zhao, Z.Zhao, and T.Y.Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[3] J.Shen, R.Pang, R.J.Weiss, M.Schuster, N.Jaitly, Z.Yang, Z.Chen, Y.Zhang, Y.Wang, R.Skerrv-Ryan, R.A.Saurous, Y.Agiomvrgiannakis, and Y.Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018, pp.4779-4783.

[4] Y.Ren, Y.Ruan, X.Tan, T.Qin, S.Zhao, Z.Zhao, and T.Y.Liu, "FastSpeech: Fast, robust and controllable text to speech," in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2019, pp.3165-3174.

[5] J.Kim, J.Kong, and J.Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in Proceedings of the International Conference on Machine Learning (ICML), 2021, pp.5530-5540.

[6] E.Casanova, J.Weber, C.D.Shulby, A.C.Junior, E.Gölge, and M.A.Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in Proceedings of the International Conference on Machine Learning (ICML), 2022, pp.2709-2720.

[7] E.Casanova, K.Davis, E.Gölge, G.Göknar, I.Gulea, L.Hart, A.Aljafari, J.Meyer, R.Morais, S.Olayemi, and J.Weber, "XTTS: A massively multilingual zero-shot text-to-speech model," in Proceedings of the Interspeech, 2024, pp.4978-4982.

[8] M.Chen, X.Tan, B.Li, Y.Liu, T.Qin, S.Zhao, and T.Y.Liu, "AdaSpeech: Adaptive text to speech for custom voice," in Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[9] Y.Yan, X.Tan, B.Li, G.Qin, T.Qin, S.Zhao, Y.Shen, W.C.Cheng, and T.Y.Liu, "AdaSpeech 2: Adaptive text to speech with untranscribed data," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021, pp.6613-6617.

[10] Y.Wu, X.Tan, B.Li, L.He, S.Zhao, R.Song, T.Qin, and T.Y.Liu, "AdaSpeech 4: Adaptive text to speech in zero-shot scenarios," in Proceedings of the Interspeech, 2022, pp.2568-2572.

[11] K.Takeda, G.Kawahara, and M.Morise, "Acoustic characteristics of speech in elderly individuals: A comprehensive review," Journal of Voice, Vol.34, No.6, pp.812-823, 2020.

[12] A.Gibiansky, S.Arik, G.Diamos, J.Miller, K.Peng, W.Ping, J.Raiman, and Y.Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017, pp.2962-2970.

[13] F.Lux, S.Meyer, L.Behringer, F.Zalkow, P.Do, M.Coler, E.A.P.Schmaltz, J.Neugebauer, T.Vu, and M.Seelbach Benkner, "Meta learning text-to-speech synthesis in over 7000 languages," in Proceedings of the Interspeech, 2024, pp.1155-1159.

[14] Y.Chen, Y.Assael, B.Shillingford, D.Budden, S.Reed, H.Zen, Q.Wang, L.C.Cobo, A.Trask, B.Laurie, C.Gulcehre, A.van den Oord, O.Vinyals, and N.de Freitas, "Meta-TTS: Meta-learning for few-shot speaker adaptive text-to-speech," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol.28, pp.2742-2752, 2020.

[15] S.O.Arik, J.Chen, K.Peng, W.Ping, and Y.Zhou, "Neural voice cloning with a few samples," in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2018, pp.10019-10029.

[16] Y.A.Li, C.Han, V.S.Raghavan, G.Mischler, and N.Mesgarani, "USAT: A universal speaker-adaptive text-to-speech approach," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol.32, pp.2892-2906, 2024.

[17] Y.Jia, Y.Zhang, R.Weiss, Q.Wang, J.Shen, F.Ren, P.Nguyen, R.Pang, I.L.Moreno, Y.Wu, Z.Ma, X.Chen, R.Skerry-Ryan, and Y.Xiao, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2018, pp.4480-4490.

[18] R.Vipperla, M.Wolters, K.Georgila, and S.Renals, "Ageing voices: The effect of changes in voice parameters on ASR performance," EURASIP Journal on Audio, Speech, and Music Processing, Vol.2010, article ID 525783, 2010.

[19] P.Micikevicius, S.Narang, J.Alben, G.Diamos, E.Elsen, D.Garcia, B.Ginsburg, M.Houston, O.Kuchaiev, G.Venkatesh, and H.Wu, "Mixed precision training," in Proceedings of the International Conference on Learning Representations (ICLR), 2018.

[20] I.Loshchilov and F.Hutter, "SGDR: Stochastic gradient descent with warm restarts," in Proceedings of the International Conference on Learning Representations (ICLR), 2017.

[21] I.Loshchilov and F.Hutter, "Decoupled weight decay regularization," in Proceedings of the International Conference on Learning Representations (ICLR), 2019.

김 영 주(Yeongju Kim) [정회원]

■ 2017년 2월 : 국립목포대학교 컴퓨터공학과(공학박사)
■ 2017년 9월 ~ 현재 : 국립목포대학교 컴퓨터공학과 강사

〈관심분야〉
AI Security, AI in Education(AIED), Text-to-Speech, Generative AI, Vibe Coding, Transfer Learning, Computer Vision

조 광 문(Kwangmoon Cho) [종신회원]

■ 1995년 8월 : 고려대학교 전산과학과(이학박사)
■ 1995년 9월 ~ 2005년 2월 : 삼성전자 통신연구소 선임연구원
■ 2000년 3월 ~ 2005년 2월 : 백석대학교 정보통신학부 교수
■ 2005년 3월 ~ 현재 : 국립목포대학교 컴퓨터학부 교수

〈관심분야〉
사물인터넷, 통신 소프트웨어, 인공지능 교육, 웹 서비스

이 도 현(Do Hyun Lee) [준회원]

■ 2024년 3월 : 국립 목포대학교 스마트비즈니스학과 (재학 중)

〈관심분야〉
인공지능, 소프트웨어, 정보통신