SATURI: 다중 방언 지원 한국어 TTS 모델
(SATURI: Multi-dialect Korean TTS model)

지도교수 : 강유

이 논문을 공학학사 학위 논문으로 제출함.

2025년 6월 14일

서울대학교 공과대학
건설환경공학부
김성진

2025년 8월

SATURI: 다중 방언 지원 한국어 TTS 모델
(SATURI: Multi-dialect Korean TTS model)

지도교수 : 강유

이 논문을 공학학사 학위 논문으로 제출함.

2025년 6월 14일

서울대학교 공과대학
건설환경공학부
김성진

2025년 8월

국문초록

본 논문에서는 지역별 방언의 강도를 조절할 수 있는 한국어 TTS(Text-to-Speech) 시스템 SATURI를 제안한다. 기존의 다화자 한국어 TTS 시스템과 달리, SATURI는 dialect-score 벡터를 활용하여 지역 방언의 특성을 자연스럽게 반영한 음성 합성을 제공한다. 특히, 음소 기반 소프트 인코딩 라벨링 기법을 적용한 dialect-score 벡터는 각 지역 방언의 특징을 효과적으로 표현할 수 있다. SATURI는 한국어의 다양한 방언을 지원하는 최초의 음성 합성 프레임워크로서, 음성 생성 분야에서 중요한 연구적 가치를 지닌다.

주요어 : 한국어 TTS, 다중 방언 TTS, YourTTS

목 차

제 1 장 Introduction

음성 합성(Text-to-Speech, TTS) 기술은 최근 수년간 비약적인 발전을 거듭해왔다. 자연스러운 음질과 다양한 스타일을 생성할 수 있는 TTS 모델들이 다수 등장함에 따라, 음성 기반 인터페이스의 활용 범위 또한 빠르게 확대되고 있다. 그러나 한국어 TTS 연구는 여전히 '표준어' 중심에 머물러 있으며, 지역 방언을 반영한 음성 합성에 대한 연구는 매우 제한적이다. 현재까지 제안된 한국어 방언 TTS 모델들 또한 특정 지역 방언(예: 경상도 사투리)에만 국한된 단일 방언 합성에 머무르고 있다.

이러한 한계점을 극복하기 위해, 본 연구에서는 하나의 텍스트 입력으로 다양한 지역 방언 스타일의 음성을 자유롭게 생성할 수 있는 SATURI(Multi-dialect Korean TTS) 모델을 제안한다. SATURI는 기존의 한국어 TTS 모델과 달리, 다수의 방언 스타일을 단일 모델로 통합하여 생성할 수 있으며, 각 방언 스타일의 강도(정도)를 정밀하게 조절하는 기능을 제공한다.

구체적으로는, 대규모 표준어 음성 데이터셋을 이용해 TTS 모델을 사전학습(pretraining)한 뒤, 다방언 데이터를 활용해 방언 레이블을 one-hot 혹은 soft encoding 방식으로 부여하고, 이를 바탕으로 모델을 미세조정(fine-tuning)하였다. 특히 단순한 one-hot 레이블링에 그치지 않고, Conformer 기반 방언 분류 모델을 통해 방언 강도를 연속적인 벡터 형태인 dialect-score vector로 추론하는 방식을 도입하였다. 이로써 데이터셋 내에서 존재하는 방언 강도의 미묘한 차이(예: 강한 경상도 억양 vs. 약한 억양)를 효과적으로 반영할 수 있었으며, phoneme 기반의 soft encoded label을 활용함으로써 방언 스타일을 보다 정교하게 학습할 수 있었다.

본 연구의 주요 기여는 다음과 같다:

· 한국어의 다양한 지역 방언을 지원하는 최초의 음성 합성 프레임워크를 제안하였다.

· 입력 텍스트에 대해 방언의 정도를 조절할 수 있는 TTS 시스템을 구현하였다.

· Classification network의 추론 결과를 활용하여 방언의 정도를 나타내는 Dialect-score를 새롭게 정의하였다.

제 2 장 Related works

제 1 절 Multi-speaker Korean TTS

다화자 텍스트-음성 변환(Text-to-Speech, TTS) 기술은 다양한 화자의 음색을 하나의 모델로 합성할 수 있는 능력을 통해 음성 합성 분야에서 중요한 발전을 이루어왔다. 특히, Deep Voice 2와 같은 초기 연구들은 학습 가능한 저차원 화자 임베딩을 도입하여 단일 모델이 수백 명의 화자의 음성을 합성할 수 있도록 하였다.[1] 이러한 접근 방식은 화자의 음색을 효과적으로 보존하면서도 다양한 음성을 생성하는 데 기여하였다.

최근에는 FastSpeech2와 같은 비자동회귀(non-autoregressive) 모델이 도입되어 합성 속도와 품질을 동시에 향상시키는 데 중점을 두고 있다.[2] 예를 들어, GANSpeech는 적대적 훈련(adversarial training)을 통해 고품질의 다화자 음성을 생성하며[3], SANE-TTS는 다국어 환경에서도 안정적인 음성 합성을 가능하게 한다.[4] 이러한 모델들은 다양한 화자의 음색을 효과적으로 학습하고 재현하였다.

한국어 다화자 TTS 분야에서도 유사한 발전이 이루어지고 있다. 김광현 외 연구진은 d-vector 기반의 화자 임베딩을 활용하여 새로운 화자의 음색을 소량의 음성 데이터로도 효과적으로 합성할 수 있는 시스템을 제안하였다.[5] 또한, GitHub에서 공개된 FastSpeech2-Pytorch-Korean-Multi-Speaker 프로젝트는 한국어 다화자 TTS 구현에 대한 구체적인 방법을 제공하며, 실제 적용 사례로 활용되고 있다.[6]

그러나 기존의 다화자 TTS 연구들은 주로 표준어 화자에 초점을 맞추고 있어, 지역 방언을 포함한 다양한 한국어 화자의 음성을 효과적으로 합성하는 데에는 한계가 있다. 특히, 방언의 억양, 어휘, 발음 등의 특성을 반영하여 자연스러운 음성을 생성하는 것은 여전히 도전적인 과제로 남아있다. 따라서, 다양한 지역 방언을 포함한 한국어 다화자 TTS 시스템의 개발은 향후 연구에서 중요한 방향성이 될 것이다.

제 2 절 Multilingual TTS Systems

다국어 텍스트-음성 합성(Text-to-Speech, TTS) 시스템은 다양한 언어 환경에서 자연스러운 음성을 생성하는 데 필수적인 기술로, 특히 저자원 언어(low-resource language)에 대한 지원이 중요한 과제로 부각되고 있다. 기존의 TTS 모델들은 주로 고자원 언어(high-resource language)에 초점을 맞추어 개발되어, 다양한 언어를 포괄하는 데 한계가 있었다.

이러한 한계를 극복하기 위해 제안된 YourTTS 모델은 VITS 구조를 기반으로 다국어 및 다화자 음성 합성을 위한 zero-shot 학습을 가능하게 하였다. YourTTS는 단일 화자의 데이터만으로도 새로운 언어에 대한 음성 합성이 가능하며, 1분 미만의 음성 데이터로도 고품질의 음성 합성을 실현할 수 있다. 이러한 특성은 특히 데이터 수집이 어려운 저자원 언어에 대해 효과적인 접근을 제공하며, 다양한 언어 환경에서의 TTS 시스템 개발에 새로운 가능성을 열

어주고 있다.[7]

그러나 YourTTS를 비롯한 기존의 다국어 TTS 모델들은 지원하는 언어의 수가 제한적이며, 특히 한국어와 같은 특정 언어에 대한 지원이 부족한 실정이다. 또한, 다국어 환경에서의 음성 합성은 언어 간의 발음, 억양, 리듬 등의 차이를 효과적으로 반영해야 하는 추가적인 도전 과제를 안고 있다.

이러한 배경에서, 본 연구는 다국어 TTS 시스템의 한계를 극복하고, 특히 한국어를 포함한 다양한 언어에 대한 자연스러운 음성 합성을 목표로 한다. 이를 위해, 사투리와 표준어 간의 미묘한 발화 차이를 반영할 수 있는 새로운 접근 방식을 제안하며, 다국어 환경에서의 TTS 시스템의 성능을 향상시키고자 한다.

제 3 장 Proposed Method

제 1 절 Korean Standard and Dialect Dataset

본 연구에서는 표준어를 포함한 한국어 다중 방언 TTS 모델 학습을 위해 AI Hub에서 제공하는 한국어 표준어 및 방언 발화 데이터셋을 사용하였다. 표준어 발화 데이터셋으로는 '다화자 음성합성 데이터'[8]를 활용하였으며, 해당 데이터는 3,400명 이상의 한국인 화자가 약 2,000개 문장을 녹음한 것으로, 총 10,000시간 이상의 음성 데이터를 포함한다. 또한, 화자의 지역 정보(예: 서울, 경기, 충청 등)가 함께 제공되어 지역별 특성을 일부 반영할 수 있다. 표준어 데이터셋은 방언 데이터셋에 비해 상대적으로 음질이 양호한 편이다.

방언 발화 데이터셋으로는 강원도, 경상도, 충청도, 제주도, 전라도 지역의 '한국어 방언 발화' 데이터를 사용하였다[9,10,11,12,13]. 해당 데이터는 2,000명 이상의 화자가 조용한 환경에서 녹음한 3,000시간 이상의 음성으로 구성되어 있으며, 표준어 데이터에 비해 음질은 다소 낮은 편이다.

모델 학습을 위해 데이터 전처리 과정을 수행하였다. 각 데이터셋에는 음성에 대응하는 문장 형태의 전사 정보가 제공되며, 이를 모델이 효과적으로 학습할 수 있도록 음소 단위로 변환하였다. 이 과정에는 Phonemizer 중 g2pkk를 사용하였다[14]. 그림 1은 5개 지역 방언 데이터셋의 오디오 길이 분포를 나타내며, 대부분 1초에서 6초 사이의 길이를 갖는 오디오가 다수를 차지함을 확인할 수 있었다. 너무 짧은 오디오는 잡음 영향을 크게 받을 수 있고, 너무 긴 오디오는 음소와 발화 간 관계 모델링에 어려움을 줄 수 있으므로, 본 연구에서는 1~6초 길이의 오디오만을 학습에 사용하였다.

또한, 텍스트 레이블 분석 결과, '?!,.' 이외의 특수문자가 포함된 데이터가 일부 존재하였다. 이들 특수문자 외 토큰은 발화에 큰 영향을 미치지 않는다고 판단하여 학습 과정에서 제외하였다. 아울러 영어와 숫자가 포함된 데이터는 전사 오류로 간주하여 제외하였으며, 실제 영어 및 숫자를 발화한 경우는 모두 한국어 표기법에 따라 음성처럼 표기된 경우만 학습에 포함시켰다.

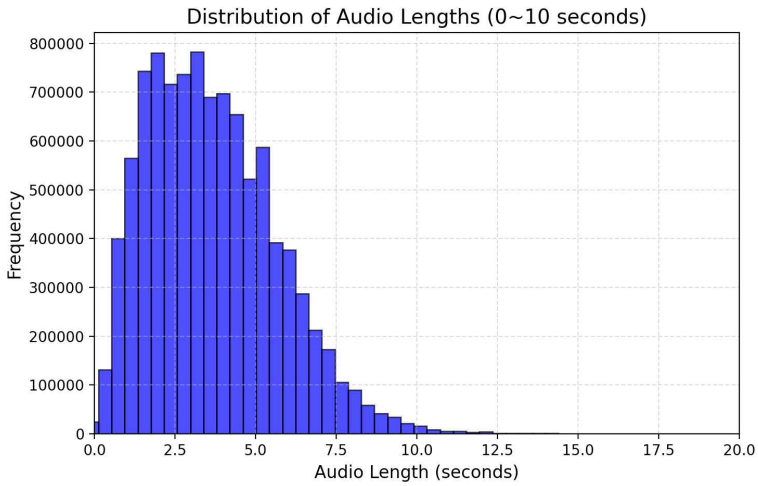

그림 1. 5개 지역 방언 데이터셋에 대한 오디오 길이 분석

제 2 절 Model Architecture

한국어 다중 방언 TTS 모델이 5개 지역 방언의 특징을 효과적으로 학습하도록 하기 위해, 본 연구에서는 사전 학습된 표준어 TTS 모델을 기반으로 하기로 하였다. 모델 아키텍처로는 다국어 및 다중 화자를 지원하며 TTS 분야에서 벤치마크로 널리 활용되는 YourTTS 모델을 채택하였다. 모델 구조는 그림 2에 상세히 나와있다.

YourTTS의 기본 구조는 대부분 유지하였으나, 본 연구에서는 다국어가 아닌 다중 방언 TTS 모델 구현을 목표로 하였기에, 기존 Lang ID 대신에 방언 클래스 정보를 담은 dialect-score vector를 도입하였다.

본 연구에서 새롭게 정의한 dialect-score vector는 [표준어 score, 충청도 score, 강원도 score, 경상도 score, 제주도 score, 전라도 score]의 6차원 벡터 형태로 구성된다. 각 지역별 score는 0에서 1 사이의 값을 가지며, 값이 0에 가까울수록 해당 지역 방언의 영향이 적음을, 1에 가까울수록 해당 지역 방언의 특성이 강함을 의미한다.

또한, YourTTS 모델에서 화자 임베딩을 추출하는 Speaker Encoder 부분에는 사전학습된 ECAPA-TDNN[15]을 사용하였다. ECAPA-TDNN은 입력된 화자의 음성 데이터로부터 192차원의 고유 발화 특성 벡터를 출력하며, 두 화자의 발화 특징이 유사할수록 두 벡터 간 코사인 유사도가 높아지는 특성을 지닌다.

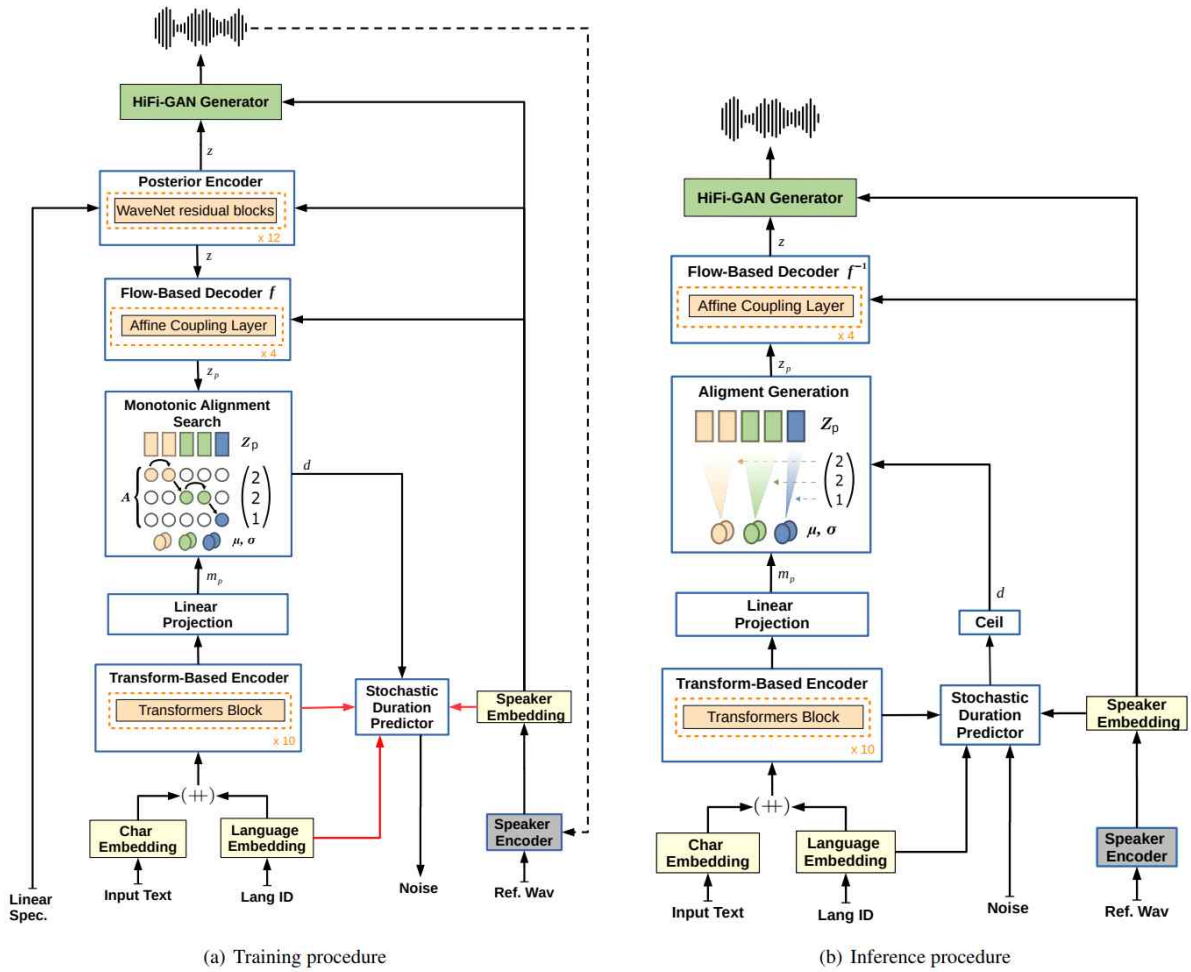

그림 2. YourTTS diagram depicting (a) training procedure and (b) inference procedure

제 3 절 Pretraining a Korean Standard TTS Model

본 연구에서는 YourTTS 기반의 한국어 TTS 사전 학습 모델이 기존에 존재하지 않고, Lang ID 대신 벡터 형태인 dialect-score vector를 사용해야 했기 때문에 모델을 초기 상태에서부터 학습하였다. 표준어 데이터셋은 발화자의 다양한 지역 정보를 포함하고 있으나, 사전 학습에는 지역 정보가 '서울'인 데이터만을 선별하여 사용하였다.

AI Hub에서 제공하는 '다화자 음성합성 데이터'는 전체 용량이 약 2.1TB에 달하는 대규모 데이터이므로, 사전 학습에 적합한 하나의 음성 ZIP 파일과 라벨 ZIP 파일을 선정하였다. 그 중 화자의 다양성과 남녀 비율이 적절히 분포된 TL1.zip과 TL23.zip 파일을 활용하였다. 이들 파일 내에서 화자의 다양성과 남녀 비율 1:1 조건을 유지하며, 표준어 음성 100시간 분량을 랜덤하게 추출한 뒤, 시간 기준으로 16:1:1 비율로 Trainset, Validset, Testset으로 분할하였다. 표준어 TTS 학습을 위해 dialect-score vector는 표준어 점수만 1.0으로 설정하고, 나머지 지역 점수는 모두 0으로 설정한 [1.0, 0.0, 0.0, 0.0, 0.0, 0.0] 벡터를 사용하였다.

학습은 NVIDIA GeForce RTX 2080 Ti GPU를 활용하였으며, 배치 크기는 28로 설정하고 총 63 epoch까지 진행하였다. 학습 과정에서 Mel-spectrogram reconstruction loss, KL divergence loss, Feature matching loss, Duration predictor loss의 변화는 각각 그림 3, 그림 4, 그림 5, 그림 6에 나타내었다. Mel-spectrogram reconstruction loss와 Duration predictor loss가 적절히 감소하였으며, 테스트용 음성 합성 결과에서도 우수한 성능이 확인되어 표준어 TTS 사전 학습은 63 epoch에서 종료하였다.

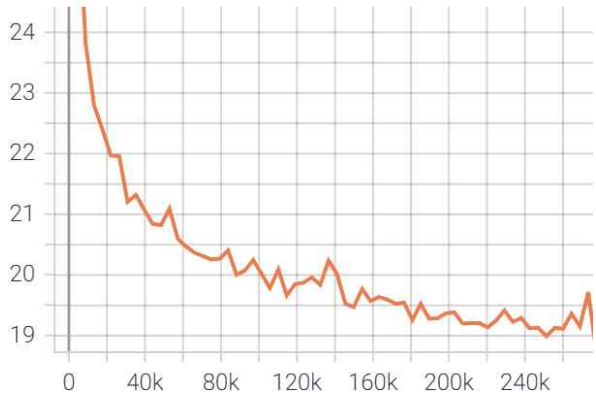

그림 3. Mel-spectrogram reconstruction loss

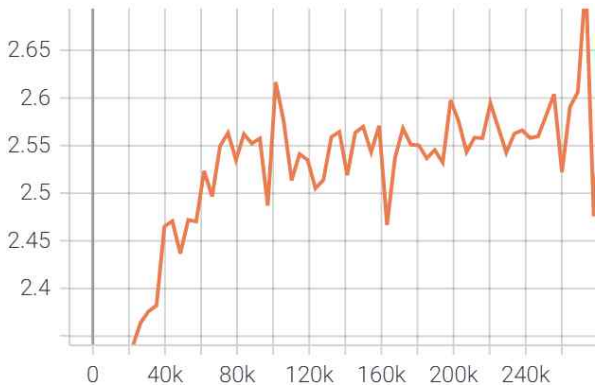

그림 4. KL divergence loss

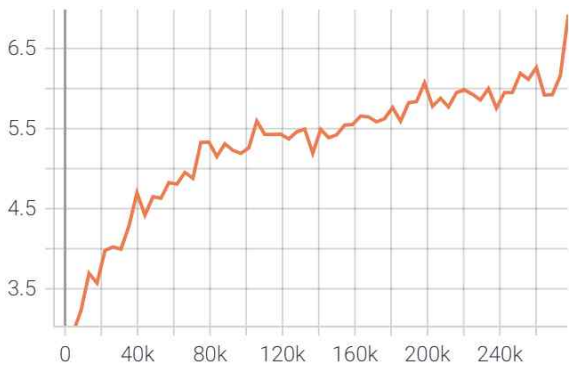

그림 5. Feature matching loss

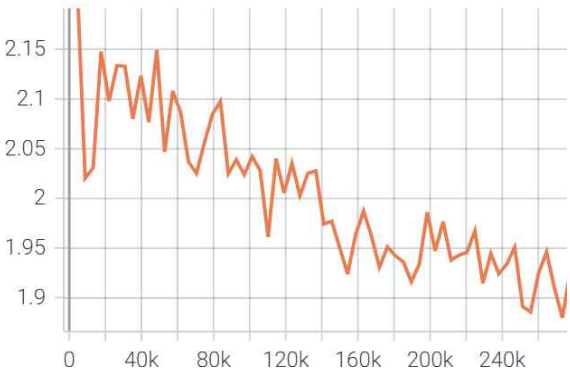

그림 6. Duration predictor loss

제 4 절 Fine-tuning Korean Multi-dialect TTS with One-Hot Labeling Method

사전학습된 한국어 표준어 TTS 모델을 기반으로 방언 데이터셋에 대해 미세조정(fine-tuning)을 수행하였다. 방언 데이터셋의 텍스트 레이블은 방언 발음을 그대로 반영한 dialect form과 표준어로 변환한 standard form 두 가지 형태가 존재하지만, 본 연구에서는 dialect form을 활용하여 학습을 진행하였다.

Dialect-score vector는 one-hot 인코딩 방식을 적용하여, 해당 지역 방언에만 1점을 부여하고 나머지 지역에는 0점을 할당하였다. 예를 들어, 충청도 방언 데이터의 경우 [0.0, 1.0, 0.0, 0.0, 0.0, 0.0] 벡터를 사용하였다.

각 지역별 데이터(표준어, 경상도, 충청도, 강원도, 제주도, 전라도)는 약 20시간 분량으로 확보하였으며, 이를 시간 기준으로 Trainset, Validset, Testset에 각각 8 : 1 : 1 비율로 분할하였다. 학습은 NVIDIA GeForce RTX 2080 Ti GPU를 사용하였으며, 배치 크기(batch size)는 28, 총 67 epoch 동안 진행되었다.

제 5 절 Fine-tuning Korean Multi-dialect TTS with Soft Labeling Method

4절에서 수행한 것과 유사하게, 사전학습된 한국어 표준어 TTS 모델을 기반으로 방언 데이터셋에 대해 미세조정(fine-tuning)을 진행하였다. 다만, 이번에는 Dialect-score vector에 Phoneme 기반의 soft-encoded label 방식을 적용하였다.

이 방식은 특정 지역 방언의 강도가 강할수록 1에 가까운 값을, 약할수록 0에 가까운 값을 부여하는 형태이며, 나머지 지역 점수는 0으로 처리한다. 다만, 총 점수의 합이 1이 되도록 표준어 점수를 소량 포함시키는 방식을 사용하였다. Soft-encoded label에 대한 자세한 설명은 4장에서 다룬다.

표준어, 경상도, 충청도, 강원도, 제주도, 전라도 각 지역별 데이터는 약 20시간 분량이며, 이를 시간 기준으로 Trainset, Validset, Testset에 8 : 1 : 1 비율로 분할하였다. 학습은 NVIDIA GeForce RTX 2080 Ti GPU에서 배치 크기 28로 총 60 epoch 동안 수행되었다.

제 4 장 Experiments

제 1 절 Experiment Overview

본 장에서 소개하는 실험들은 지역 간 방언의 특징을 가장 효과적으로 표현할 수 있는 TTS 모델 학습 방법을 규명하는 것을 목적으로 한다. 특히, 제절과 7절의 실험 결과를 비교함으로써 One-hot 방식과 Soft-label 방식 중 어느 방식이 방언 표현에 더 효과적인지를 평가하고자 한다. 또한 3절부터 6절까지의 실험을 통해, Dialect-score를 산출하는 데 사용되는 classification 모델이 어떤 입력 정보를 기반으로 할 때 강한 방언과 약한 방언의 정도를 가장 잘 반영할 수 있는지를 분석하고자 한다.

각 절에서 수행하는 실험은 다음과 같다. 2절에서는 사전 학습된 표준어 TTS 모델을 One-hot 방식으로 fine-tuning하는 실험을 다루며, 7절에서는 동일한 표준어 모델을 Soft-label 방식으로 fine-tuning하는 실험을 다룬다. 두 방식의 결과는 추후에 상호 비교를 통해 방언 표현력의 차이를 분석할 것이다. 3절에서는 Mel-spectrogram만을 입력으로 사용하는 dialect classification 모델의 학습 결과를, 4절에서는 Energy, HuBERT, F0 정보를 결합하여 학습한 결과를 다룬다. 5절에서는 F0 단독 입력으로 학습한 실험을, 6절에서는 Phoneme 정보를 기반으로 학습한 실험을 다룬다. 이들 실험을 비교 분석하여 가장 우수한 성능을 보이는 classification 모델의 입력 구성을 최종적으로 채택할 예정이다.

제 2 절 Korean Dialect TTS Evaluation with One-Hot Encoded Labels

학습 과정에서 Mel-spectrogram reconstruction loss, KL divergence loss, Feature matching loss, Duration predictor loss의 변화는 그림 7, 8, 9, 10에 나타나 있다.

방언 데이터셋으로 fine-tuning한 TTS 모델을 이용하여 직접 추론을 수행하였다. 평가 문장으로는 “언제가 놀러가기가 가장 좋은 계절이라고 생각해?”를 사용하였으며, 화자 정보는 크게 영향을 미치지 않아 데이터셋 내 임의의 화자를 선택하여 실험하였다. 각 지역별 score가 1점이고 나머지 지역이 0점인 one-hot encoded dialect-score vector를 입력하여 추론한 결과, 표준어와 경상도에 1점을 부여한 경우에는 해당 지역의 방언 특성이 명확히 반영되었다. 특히 표준어 및 경상도 방언의 억양을 잘 모사하는 것으로 확인되었다. 반면, 다른 지역의 경우 억양 차이가 뚜렷하게 느껴지지 않았다.

동일 문장에 대해 표준어와 경상도 score에 각각 0.5점을 부여한 dialect-score vector로 추론을 수행하였다. 이 경우 표준어에 1점을 준 경우보다 억양 변화가 더 크고, 경상도에 1점을 준 경우보다는 억양 변화가 적어 중간 정도의 억양 변화를 관찰할 수 있었다.

그러나 one-hot encoded label 방식은 지역별 방언 특징을 충분히 학습하지 못할 가능성이 있다. 이는 방언 데이터셋이 표준어 데이터셋에 비해 잡음이 많고, 방언 데이터 내에서도 표

준어에 가까운 발화가 다수 존재하기 때문이다. 또한 같은 지역 내에서도 사투리 강도가 다양한데, one-hot label 방식은 사투리 정도와 관계없이 동일한 라벨을 부여하여 지역 간 방언의 미묘한 차이를 모델이 학습하는 데 어려움을 초래할 수 있다.

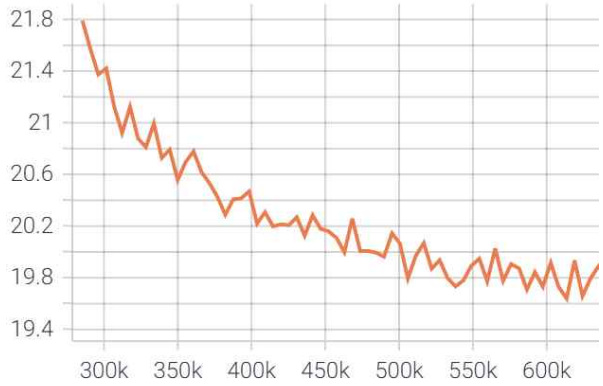

그림 7. Mel-spectrogram reconstruction loss

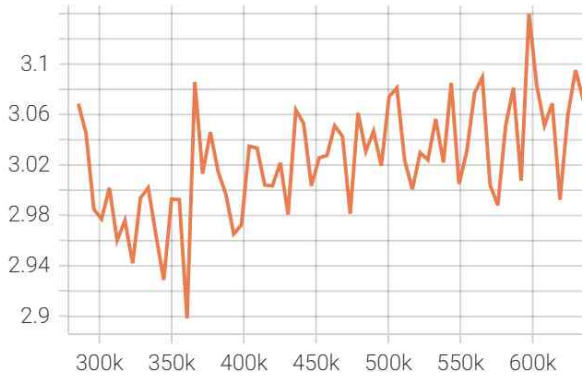

그림 8. KL divergence loss

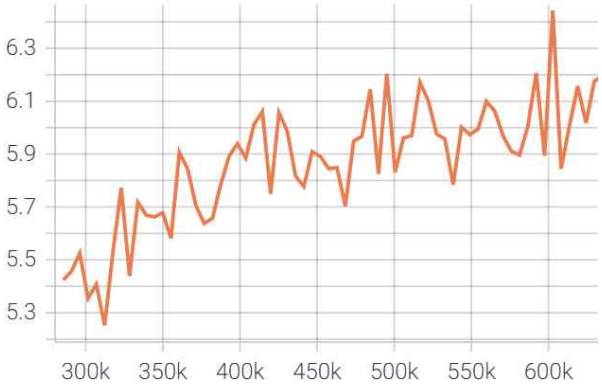

그림 9. Feature matching loss

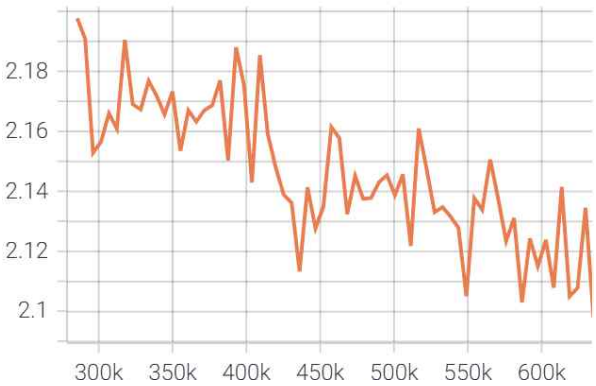

그림 10. Duration predictor loss

제 3 절 Dialect Classification Using Mel-Spectrogram

발화별로 사투리 정도를 반영한 Dialect-score vector를 부여하기 위해 분류(classification) 모델을 사용하였다. 본 연구에서는 Conformer 모델을 분류기로 채택하였으며,[16] vector 형태의 출력을 얻기 위해 Conformer의 마지막 softmax 레이어를 제거한 구조를 사용하였다.

데이터는 표준어, 경상도, 충청도, 강원도, 제주도, 전라도 각 지역별로 30시간 분량을 확보하여 시간 기준으로 Trainset, Validset, Testset을 4 : 1 : 1의 비율로 분할하였다. 모델 입력으로는 오디오 파일의 Log mel-spectrogram을 사용하였으며, 파라미터는 sample_rate 16kHz, n_fft 1024, hop_length 256, win_length 1024, n_mels 80으로 설정하였다.

학습 결과 Accuracy와 Loss 변화는 그림 11과 12에 나타나 있다. 그림에서 확인할 수 있듯이 모델이 데이터셋에 과적합(overfitting)되는 경향을 보였다. 특히 방언 데이터셋은 지역별로 서로 다른 환경에서 녹음되었기 때문에, 모델이 실제 방언 특성보다는 지역별 데이터셋 간의 잡음 차이를 학습하는 현상이 발생할 수 있다고 판단하였다. 따라서 방언의 특징을 효과적으

로 학습할 수 있도록 적절한 입력 특성 설계가 필요함을 느꼈다.

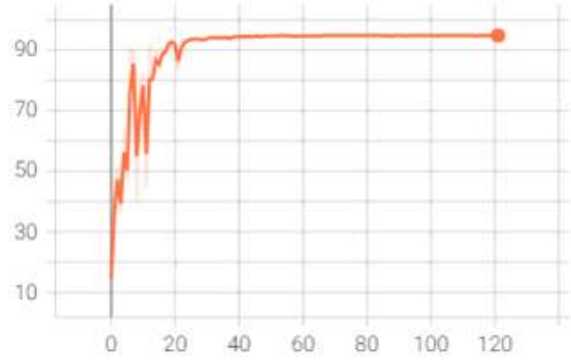

그림 11. Validation Accuracy

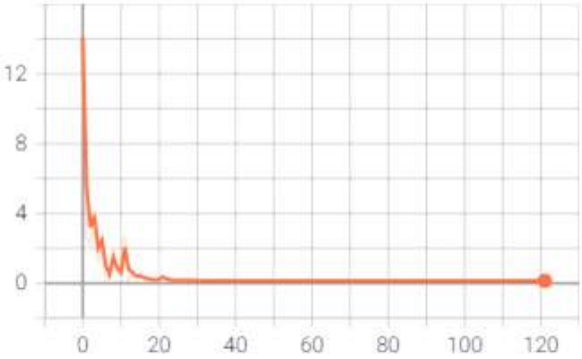

그림 12. Validation Loss

제 4 절 Dialect Classification Using Energy + HuBERT + F0

모델 구조는 3절과 동일하게 유지하되, 입력(input) 데이터만 다르게 하였다. 지역별로 강하게 발음되는 주파수 특성을 학습하기 위해 Log mel-spectrogram의 energy를 사용하였으며, 음소 및 언어적 패턴에 집중하기 위해 사전학습된 HuBERT를 활용하였다.[17] HuBERT는 음성 신호에서 의미 있는 음향 특징을 자가 지도학습(self-supervised learning) 방식으로 추출하는 모델로, 음소 인식 및 언어 표현 학습에 효과적이다. 아울러, 지역별 억양 차이가 dialect-score에 반영되도록 Log scale로 변환한 F0 정보를 함께 사용하였다. Energy와 F0는 HuBERT 출력 차원과 일치시키기 위해 모두 HuBERT 차원 크기로 선형 보간(linear interpolation)하여 맞추었다.

학습 결과 Accuracy와 Loss는 그림 13과 그림 14에 나타나 있으며, 앞 절과 마찬가지로 데이터셋에 과적합(overfitting)되는 경향을 보였다.

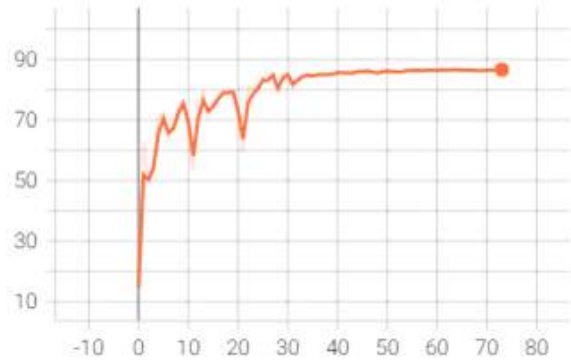

그림 13. Validation Accuracy

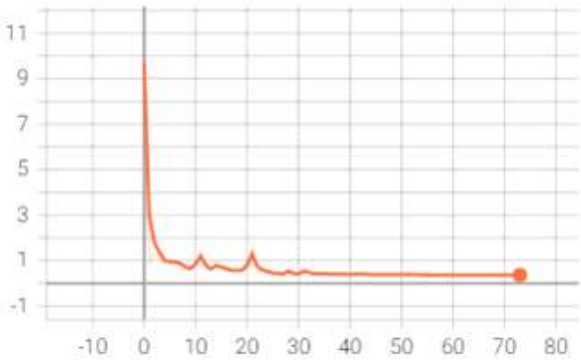

그림 14. Validation Loss

제 5 절 Dialect Classification Using F0 Only

동일한 모델 구조를 사용하되, 모델의 입력으로 F0 정보만을 사용하여 실험을 진행하였다. 이는 억양만을 바탕으로 지역 방언을 구분할 수 있는지를 평가하기 위한 시도이다. 학습 결과 Accuracy와 Loss는 그림 15와 그림 16에 나타나 있다. 결과를 통해 모델이 지역을 제대로 분류하지 못하는 것을 확인할 수 있었다. 즉, F0 정보만으로는 지역 방언을 구분하기에 어려움이 있음을 알 수 있다.

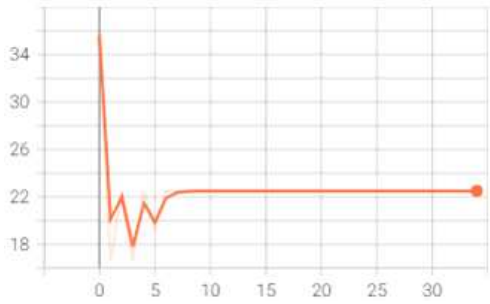

그림 15. Validation Accuracy

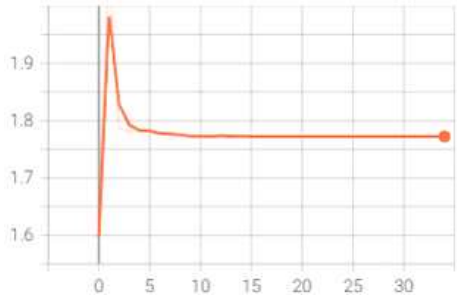

그림 16. Validation Loss

제 6 절 Dialect Classification Using Phonemes

지역별로 사용하는 특이한 사투리 단어가 존재한다는 점에 착안하여, classification 모델의 입력으로 phoneme sequence를 사용한 실험을 진행하였다. 각 지역별 특이한 사투리 단어의 사용 빈도와 dialect-score가 비례할 것이라는 가정을 세웠다. Phonemizer로는 표준어 TTS pretraining 시 사용했던 g2pkk를 활용하였다.

모델 구조는 앞서 3, 4, 5절에서 사용한 Conformer를 기본으로 하였으나, Conformer의 encoder에 입력하기 전에 embedding 레이어를 추가하여 차원 확장을 수행하는 부분을 더하였다.

학습 결과는 그림 17과 그림 18에 나타나 있으며, phoneme sequence를 사용했을 때 지역별 구분이 어느 정도 잘 이루어지는 것을 확인할 수 있었다. 이에 18 epoch 시점의 checkpoint를 사용하여 dialect-score를 추론해보았다. 그림 19는 사투리 정도가 강하다고 판단된 샘플 5개에 대한 dialect-score 추론 결과이며, 그림 20은 사투리 정도가 약하다고 판단된 샘플 5개에 대한 결과이다. 제주도 방언에 해당하는 dialect-score는 벡터의 다섯 번째 값에 해당한다. 사투리가 강한 샘플들은 1에 가까운 높은 dialect-score 값을 보였으며, 사투리가 약한 샘플들은 대부분 0에 가까운 낮은 값을 나타냈다.

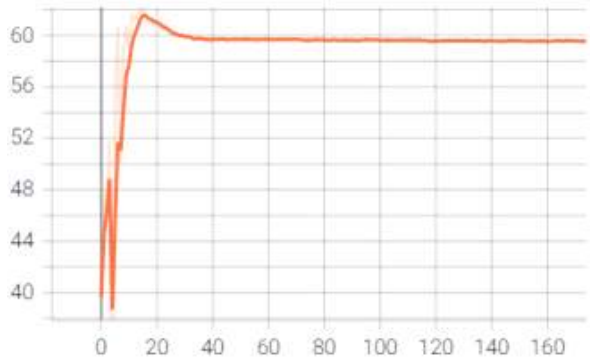

그림 17. Validation Accuracy

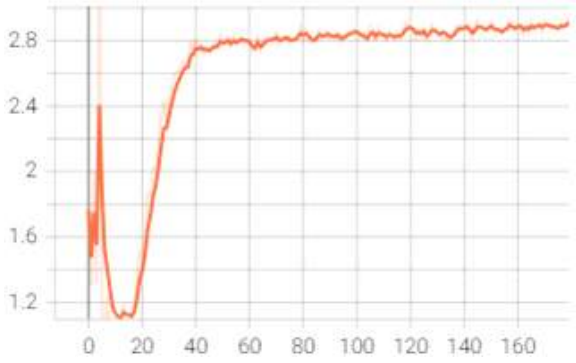

그림 18. Validation Loss

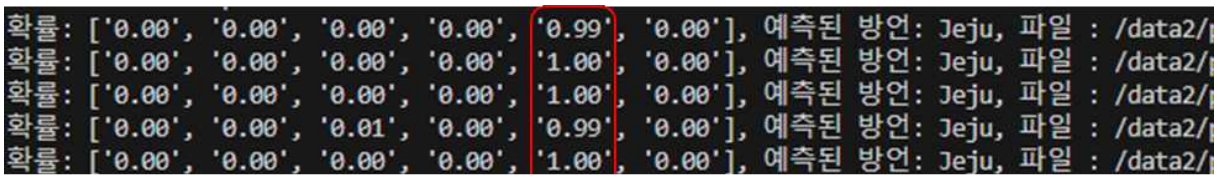

그림 19. 사투리의 정도가 강한 제주도 발화에 대한 dialect-score vector

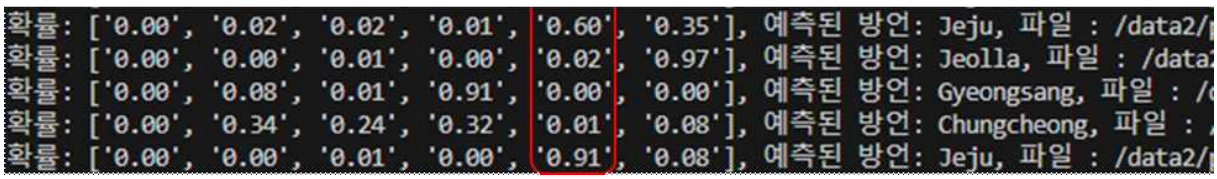

그림 20. 사투리의 정도가 약한 제주도 발화에 대한 dialect-score vector

제 7 절 Korean Dialect TTS Evaluation with Phoneme-Based Soft Encoded Labels

Phoneme sequence를 입력으로 사용한 classification 모델이 사투리 정도를 비교적 정확하게 판단하여 dialect-score를 계산할 수 있음을 앞서 확인하였다. 따라서 1절에서 finetuning에 사용한 동일한 trainset 데이터에 대해 classification 모델로부터 dialect-score vector를 추론하여 soft-encoded label을 계산하였다.

soft-encoded label은 다음과 같은 절차로 산출되었다.

① 발화된 지역에 해당하는 dialect-score는 그대로 유지한다.

② 발화 지역 이외의 지역에 대한 dialect-score는 모두 0으로 변경한다.

③ 6개 dialect-score의 합이 1이 되도록 표준어 dialect-score(벡터의 첫 번째 값)를 조정한다.

예를 들어, 제주도 발화에 대해 classification 모델이 [0.00, 0.02, 0.02, 0.01, 0.60, 0.35]의 결과를 출력한 경우, soft-encoded label은 [0.40, 0.00, 0.00, 0.00, 0.60, 0.00]로 계산된다. 이렇게 산출한 soft-encoded label을 활용하여 기존에 사전학습된 Korean Standard TTS 모델을 finetuning하였다. 학습 과정 중 Mel-spectrogram reconstruction loss, KL divergence loss, Feature matching loss, Duration predictor loss의 변화는 각각 그림 21, 그림 22, 그림 23, 그

림 24에 나타나 있다.

Finetuning된 TTS 모델로 추론을 수행하였으며, 2절과 동일한 문장인 “언제가 놀러가기가 가장 좋은 계절이라고 생각해?”를 사용하였다. 화자 정보도 동일한 화자를 선택하여 설정하였다. 한 지역에만 1점을 부여하고 나머지 지역에는 0점을 준 dialect-score vector를 입력으로 하여 추론을 진행하였다. 결과는 2절과 유사하였다. 표준어 score를 1점으로 준 경우와 경상도 score를 1점으로 준 경우에서 확연한 발화 차이를 확인할 수 있었으며, 특히 경상도 score에 1점을 준 경우 억양 차이가 뚜렷하게 관찰되었다. 경상도 이외 지역에 1점을 준 케이스에서는 뚜렷한 억양 차이는 나타나지 않았으나, 충청도 score 1점 케이스에서는 발화 속도가 느렸고, 제주도 score 1점 케이스에서는 비교적 빠른 발화 속도가 관찰되었다. 또한, 경상도 score에 0.5점을 부여한 테스트도 실시하였으며, 1절과 마찬가지로 0점과 1점일 때의 중간 정도 억양 변화를 확인할 수 있었다.

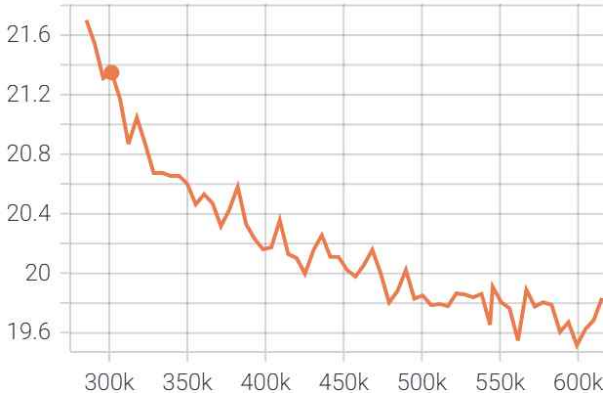

그림 21. Mel-spectrogram reconstruction loss

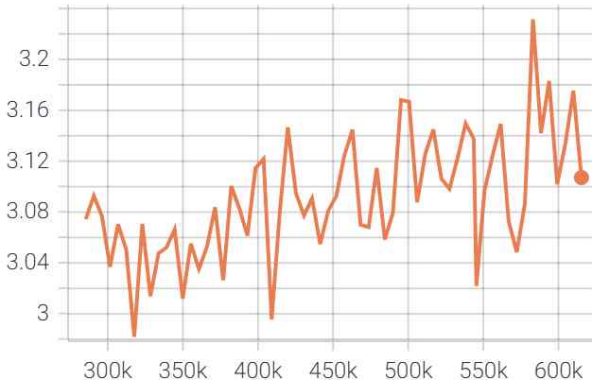

그림 22. KL divergence loss

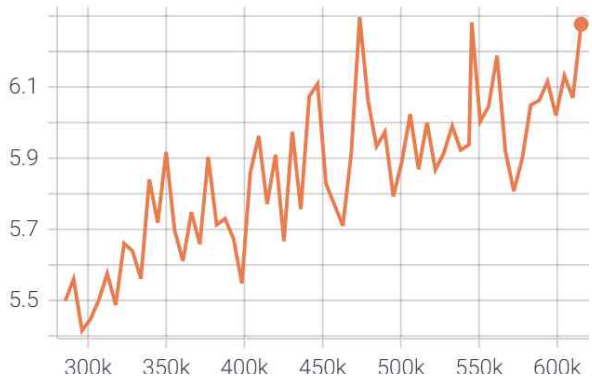

그림 23. Feature matching loss

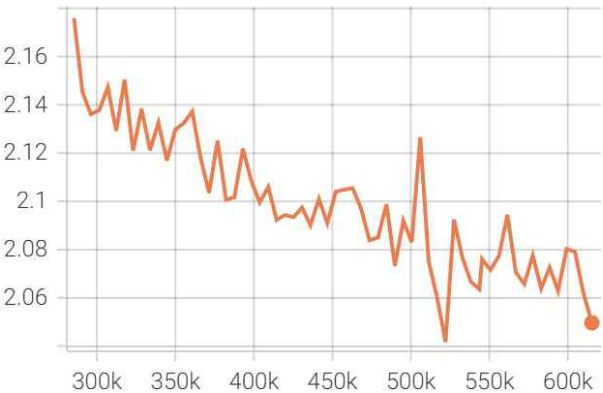

그림 24. Duration predictor loss

제 5 장 Discussion

본 연구는 한국어의 다양한 지역 방언을 지원하는 최초의 음성 합성 프레임워크를 제안하고, 입력 텍스트에 대해 방언의 정도를 조절할 수 있는 TTS 시스템을 구현하였다. 특히, classification network의 추론 결과를 활용하여 방언의 정도를 나타내는 Dialect-score vector를 정의하고 이를 기반으로 한 soft-encoded labeling 방식을 통해 기존의 one-hot 방식보다 더 유연하게 방언 특성을 표현하고자 하였다.

실험 결과, phoneme 기반 classification 모델을 통해 지역별로 사용되는 사투리 단어의 차이를 포착하고, 방언 강도를 연속적으로 표현하는 Dialect-score vector를 생성할 수 있었으며, 이를 TTS 모델의 soft-label로 활용함으로써 보다 세밀한 억양 조절이 가능함을 확인하였다. 이는 one-hot 방식이 방언 강도의 차이를 반영하지 못하는 단점을 보완할 수 있는 가능성을 제시한다.

그러나 soft-label 방식이 방언의 미세한 억양 차이나 발화 속도 차이를 묘사하는 데에는 일정한 효과가 있었으나, 지역별 방언 고유의 언어적 특징을 뚜렷하게 반영한다고 느끼기는 어려웠다. 특히, 경상도와 표준어의 억양 차이는 비교적 명확하게 드러났지만, 충청도, 전라도, 강원도 등 다른 지역의 방언 특성은 청각적으로 미묘하게 표현되어 사용자의 직관적인 인식과 차이가 있었다.

또한, 본 연구에서는 방언 TTS의 성능을 평가할 수 있는 정량적인 지표가 부재하다는 점이 한계로 작용하였다. 방언의 발화가 자연스러운지, 실제 특정 지역 방언처럼 들리는지를 평가할 수 있는 지역성 인식 평가 기준이나 청취 기반 방언 구분 정확도와 같은 지표가 필요하다고 판단된다.

향후 연구에서는 데이터셋 품질과 모델 구조를 개선하는 한편, 방언 표현력의 객관적 평가를 위한 새로운 정량적 지표를 설계하고 적용할 계획이다. 또한, 음성 합성 결과에 대한 사용자 평가 기반의 정성적 실험도 함께 진행할 예정이다.

제 6 장 Conclusion

본 논문에서는 한국어의 다양한 지역 방언을 지원하는 최초의 음성 합성 프레임워크를 제안하였다. 또한, 입력 텍스트에 대해 방언의 정도를 조절할 수 있는 phoneme-based soft-encoded label을 이용한 한국어 방언 TTS를 구현하였으며, classification network의 추론 결과를 활용하여 방언의 정도를 나타낼 수 있는 Dialect-score를 새롭게 정의하였다.

여러 실험의 결과, soft-label 방식을 통한 TTS 모델이 억양의 변화나 발화 속도 차이를 표현하는 데에 있어 기존 one-hot 방식보다 유의미한 개선을 보였지만, 여전히 지역별 방언의 언어적 특성 자체를 명확히 반영하지는 못하는 한계가 있었다. 또한 방언 TTS 성능을 정량적으로 평가할 수 있는 기준이 없다는 점은 연구의 객관성을 확보하는 데 어려움을 주었다.

향후 연구에서는 이러한 한계를 보완하기 위해 Dialect-score 산출 방식에 대한 추가 연구를 진행하고, 지역성에 대한 정량적 평가 지표의 설계에 중점을 둘 예정이다. 이와 함께, 실제 사용자 청취 실험을 기반으로 방언 표현력에 대한 주관적 평가도 병행하여 연구의 외연을 확장할 계획이다.

본 연구에서 제안한 방법은 향후 한국어 다중 방언 TTS 기술의 기반이 될 수 있으며, 이를 통해 보다 다양한 지역 사용자들에게 친숙한 음성 인터페이스를 제공하는 데 기여할 수 있을 것이다. 최종 TTS 합성 결과는 demo page[18]에서 확인할 수 있다.

제 7 장 References

[1] Arik, Sercan, et al. "Deep voice 2: Multi-speaker neural text-to-speech." arXiv preprint arXiv:1705.08947 (2017).

[2] Ren, Yi, et al. "Fastspeech 2: Fast and high-quality end-to-end text to speech." arXiv preprint arXiv:2006.04558 (2020).

[3] Yang, Jinhyeok, et al. "GANSpeech: Adversarial training for high-fidelity multi-speaker speech synthesis." arXiv preprint arXiv:2106.15153 (2021).

[4] Cho, Hyunjae, et al. "SANE-TTS: stable and natural end-to-end multilingual text-to-speech." arXiv preprint arXiv:2206.12132 (2022).

[5] Kim, K. H., & Kwon, C. H. (2022). A Korean Multi-speaker Text-to-Speech System Using d-vector. *Journal of the Convergence on Culture Technology*, 8(3), 469 - 475. https://doi.org/10.17703/JCCT.2022.8.3.469

[6] Kairess. *torch-hybrid-tacotron2*. GitHub repository. Available: https://github.com/kairess/torch-hybrid-tacotron2

[7] Casanova, Edresson, et al. "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone." International conference on machine learning. PMLR, 2022.

[8] https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=542

[9] https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=118

[10] https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=119

[11] https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=122

[12] https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=121

[13] https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=120

[14] Harmlessman. *g2pkk: Grapheme-to-Phoneme Conversion for Korean*. GitHub repository. Available: https://github.com/harmlessman/g2pkk

[15] Desplanques, Brecht, Jenthe Thienpondt, and Kris Demuynck. "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification." arXiv preprint arXiv:2005.07143 (2020).

[16] Gulati, Anmol, et al. "Conformer: Convolution-augmented transformer for speech recognition." arXiv preprint arXiv:2005.08100 (2020).

[17] Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." IEEE/ACM transactions on audio, speech, and language processing 29 (2021): 3451-3460.

[18] https://sungjin20.github.io/SATURI-demo/

Abstract

In this paper, we propose SATURI, a Korean text-to-speech (TTS) system that enables controllable synthesis of regional dialects. Unlike conventional multi-speaker Korean TTS systems, SATURI utilizes a dialect-score vector to naturally reflect the unique prosodic and phonetic features of each dialect in the generated speech. In particular, the phoneme-based soft encoding technique applied to the dialect-score vector effectively captures and represents regional linguistic characteristics. As the first TTS framework to support multiple Korean dialects, SATURI offers significant research value and opens new directions for speech synthesis in underrepresented linguistic varieties.

SATURI: A Controllable Korean Text-to-Speech Model for Regional Dialect Synthesis

Kim Sung Jin

Civil and Environmental Engineering

Seoul National University

This study presents the first Korean TTS framework supporting regional dialects, introducing a Dialect-score to represent dialect intensity using a phoneme-based classification model. Soft-labeling improved prosody and speech rate control compared to one-hot encoding but fell short in capturing distinct linguistic traits of each dialect. Future work will refine Dialect-score inference and propose quantitative evaluation metrics.

keywords : Korean TTS, Multi-dialect TTS, YourTTS