자연스러운 한국어 음성합성을 위한 휴지 예측 모델

Predicting Pause Position for Natural Text-to-Speech Synthesis of Korean

전북대학교 JEONBUK NATIONAL UNIVERSITY

나정호 · 이 정 · 나승훈 (전북대학교), 정정범, 최맹식, 이충희 (엔씨소프트)

{ jhna2023, fhqlatm, nash } @ jbnu.ac.kr, { jbjeong, mschoi, forever73 } @ ncsoft.com

Ⅰ. 서 론

연구 목적

휴지는 조음활동에서 호흡, 발화의 강조, 청자의 주목을 끌거나 청자가 화자의 발화를 이해하기 위한 시간을 제공하는 등의 다양한 기능을 수행한다. 이러한 휴지를 컴퓨터나 인공지능을 통한 음성합성에 삽입하면 보다 자연스러운 음성을 얻을 수 있을 것으로 예상되지만, **Respiratory Pause (RP)**는 화자의 발화스타일 등으로 인해 동일한 문장일지라도 삽입 여부와 위치, 기간 등에서 차이를 보인다. 따라서 자연스러운 음성합성을 위해 휴지의 위치를 예측하고 이를 적절한 위치에 삽입하기 위한 체계화된 방법이 필요하다. 이에 본 논문에서는 화자 임베딩과 음성특징을 활용한 한국어 휴지 예측 모델을 제안하였다.

Ⅱ. 실험 설계

데이터셋

Raw Text	유명한 관광 명소가 아주 많아요.								
Subwords	유명	##한	관광	명소	##가	아주	많	##아	##요	.
RP Label	O	O	O	O	RP	O	O	O	O	O

그림 1: 데이터 전처리 및 라벨링 예시

표 1: 한국어 자연언어 감성 분류 예측 정확도

	Training	*Validation*	*Test*
Sentences	96399	6025	18075
Tokens	1160228	72711	216974
Speakers	49	49	49
RP	69790	4465	18890

한국어의 휴지 예측을 위해서는 토크나이징된 발화문과 발화별 음성특징, 화자 임베딩 등 다양한 요인이 필요하다. 이를 위해 다음과 같은 데이터 수집 및 전처리 과정을 수행하였다. 첫째, 모델 학습에는 **Ai-hub**에서 제공하는 '감성 및 발화스타일별 음성합성 데이터'를 사용하였다. 둘째, **Montreal Forced Aligner(MFA)**를 사용하여 음성과 음소를 정렬하고 휴지와 휴지 기간을 추출하였다. 셋째, 그림 1과 같이 **Klue RoBERTa**의 토크나이저를 사용하여 모든 문장을 서브워드(Subwords)로 분리하고 휴지가 나타나지 않은 문장은 'O'로 RP가 나타난 서브워드는 'RP'로 라벨링하였다. 마지막으로 화자 임베딩 추출을 위해 **Kaldi**의 'SRE16 xvector Model'을 사용하여 음성데이터로부터 **x-vector**를 추출하였고, **python**의 **librosa** 라이브러리를 사용하여 음성데이터로부터 **Mel Spectrogram**을 추출하였다.

이러한 데이터 전처리를 통해 총 120,499건의 학습데이터를 구축하였다. 최종적으로 모델학습에 사용한 문장(Sentences), 토큰(Tokens), 화자(Speakers), RP의 분포는 표1과 같다.

모델 및 평가 방법

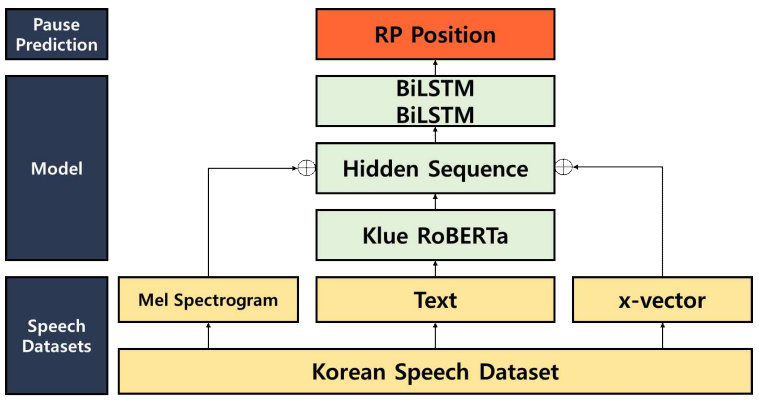

그림 2: 휴지 예측 모델 구조

본 연구에서 사용한 모델은 Yang et al. (2023)의 모델을 베이스라인으로 하여 그림2와 같이 확장하였다. 먼저, 한국어 발화문에 나타난 구문 특징을 추출하기 위해 한국어 데이터를 대상으로 사전학습된 'klue/roberta-large'모델을 파인튜닝하였다. 그 다음 2개의 BiLSTM 층을 디코더로 사용해 Klue RoBERTa의 출력값, Mel Spectrogram, 화자 임베딩을 입력으로 받아 주어진 발화문에서 RP의 위치를 예측하는 방식으로 모델을 학습하였다. 모델 학습을 하는 동안 배치크기는 64, 학습률은 2e-5로 설정하였고 옵티마이저는 AdamW를 사용하여 15 epochs동안 학습을 수행하였다. 그 후, 검증 과정에서 가장 높은 성능을 보인 모델을 선정하고 정밀도(Precision)와 재현율(Recall), F_1 점수(F_1 score)를 사용해 성능을 비교하였다.

Ⅲ. 평가 결과

결과 분석

표 2: 한국어 자연언어 감성 분류 예측 정확도

Model	Precision	Recall	F_1 Score
Klue RoBERTa	0.7152	0.6426	0.6770
Klue RoBERTa + BiLSTM	0.7309	0.6200	0.6709
Klue RoBERTa + BiLSTM W/ x-vector	0.7755	0.7174	0.7453
Klue RoBERTa + BiLSTM W/ Mel Spectrogram	0.8151	0.7869	0.8007
Klue RoBERTa + BiLSTM W/ x-vector + Mel Spectrogram	**0.8221**	**0.8194**	**0.8208**

본 논문에서 제안한 모델의 성능 평가 결과는 표 2와 같으며 모델은 **Klue RoBERTa** 단일 모델 혹은 **Klue RoBERTa + BiLSTM** 모델 대비 가시적인 성능 향상을 보인다. 특히, 화자 임베딩을 추가한 모델은 동일한 구조의 **Klue RoBERTa + BiLSTM** 모델보다 F1 점수 기준 약 7%p의 성능이 올랐으며, **Mel Spectrogram**을 추가한 모델은 약 13%p, 두 가지 특징 벡터를 모두 추가한 모델은 약 15%p의 성능이 향상된 것을 확인할 수 있었다. 이와 같은 결과는 휴지 예측에 있어 **x-vector**에 기반한 화자 임베딩과 **Mel Spectrogram**에 기반한 음성특징의 필요성과 효용성을 증명한다.

Ⅳ. 결 론

결 론

본 논문에서는 자연스러운 한국어 음성합성을 위해 화자 임베딩과 음성특징에 기반한 휴지 위치를 예측한다. 기존 모델과 대비하여 한국어에 적합한 **Klue RoBERTa** 모델을 사용하였으며, 다양한 화자의 발화스타일에서 화자 임베딩을 추출하고 각 발화별로 음성특징을 뽑아내 주어진 발화문에서 휴지의 위치를 예측하는 모델을 설계하였다. 제안된 모델은 동일한 구조의 다른 모델에 비해 높은 성능을 보여 휴지 예측 연구에서 화자 임베딩과 음성특징의 활용 가치를 증명하였다.

향후 연구로는 제안한 모델을 최근 자연언어처리 분야에서 높은 성능을 보이는 대규모 언어 모델(**Large Language Model**)로 확장하여 성능을 평가하고, 확장된 모델을 활용해 휴지 위치를 포함한 한국어 음성합성 데이터를 생성하고 정성적·정량적 평가를 통해 생성된 데이터의 효용성을 확인하고자 한다.