자연스러운 한국어 다중 화자 음성 합성을 위한 x-vector 기반 휴지 지속 시간 예측

Duration-aware Pause Prediction using x-vector
for Natural Korean Multi-speaker Speech Synthesis

이 정ㅇ[1], 나정호[2], 나승훈[3], 정정범[4], 최맹식[5], 이충희[6]
전북대학교[1,2,3], NCSOFT Corp.[4,5,6]
fhqlatm@jbnu.ac.kr

전북대학교 JEONBUK NATIONAL UNIVERSITY

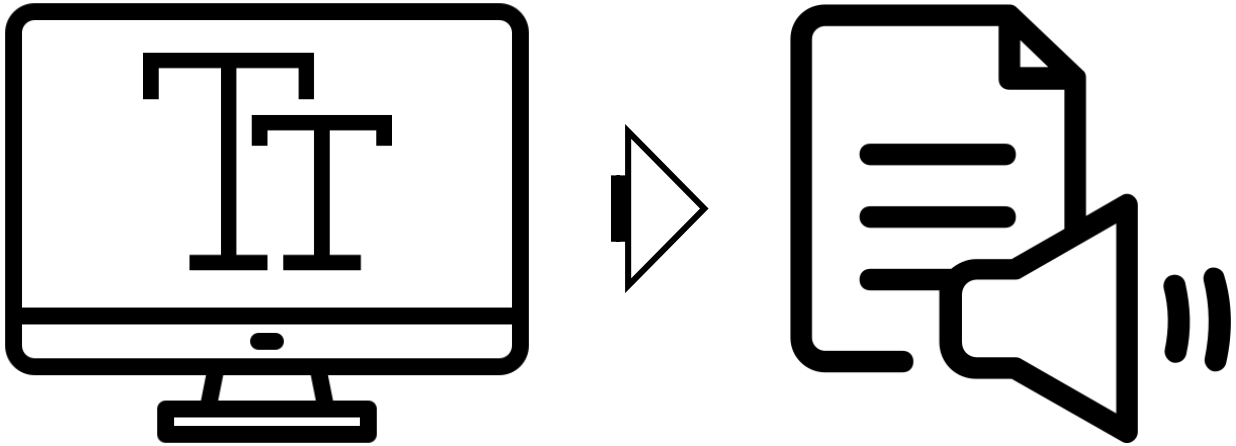
KSC 2023

NCSOFT

목차

- 서론
- 관련 연구
- 모델
- 실험
- 결론 및 향후 연구

서론



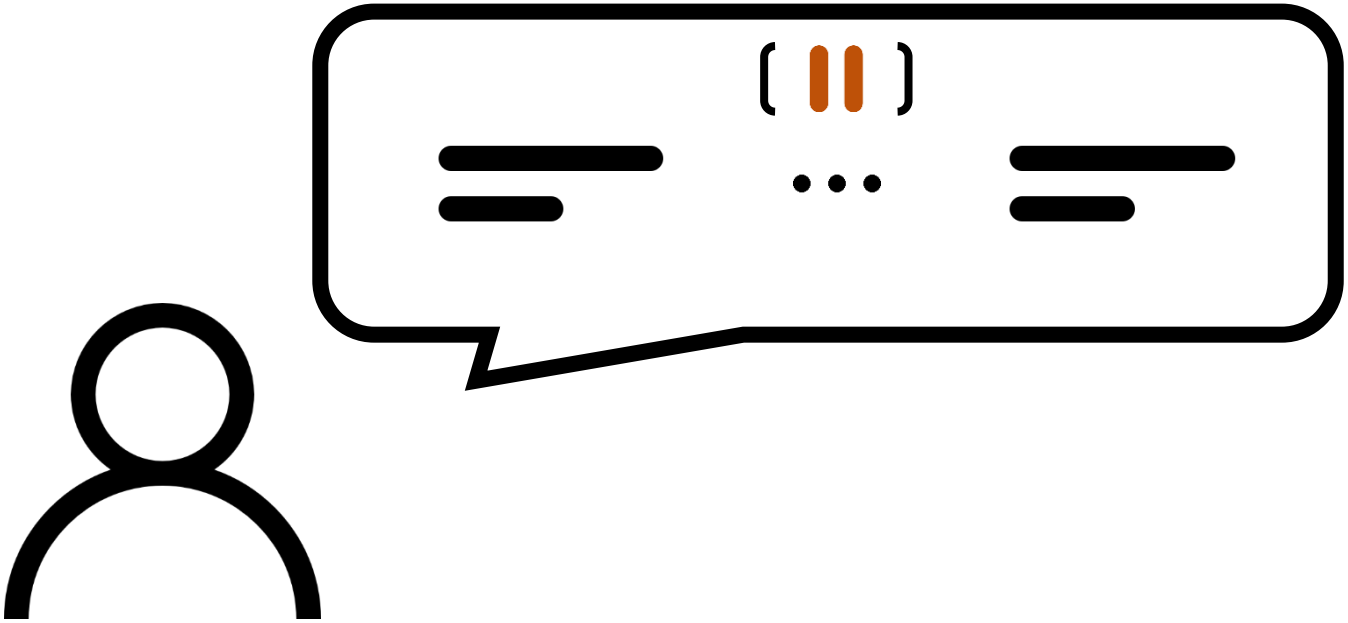
- **음성 합성(Speech Synthesis)**

  = Text-to-Speech (TTS)

  사람의 발화 음성에 해당하는 음파를 기계가 자동으로 합성하는 기술

  자연스러운 음성 발화를 모방하기 위해서는 발화의 운율, 억양, 휴지 및 강조 등을 포착하고 적절히 삽입하는 것이 중요

서론

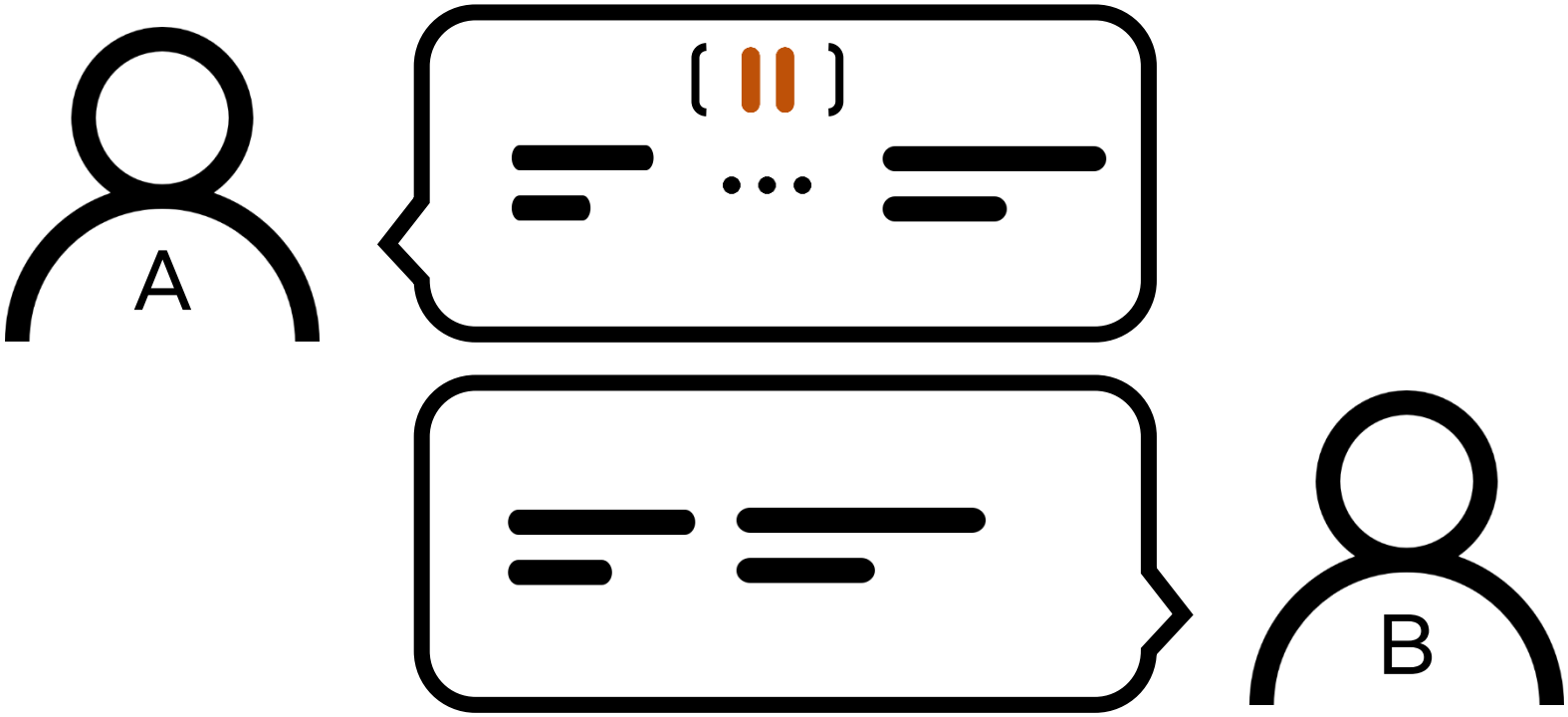


- **휴지(Pause)**

  음성 발화 내의 일시적인 정지

  - PIP: 구두점 등 문장 부호에서 발생하는 휴지(Punctuation-Indicated Pause)
  - RP: 긴 문장 내에서 발화 특성에 따라 나타나는 휴지(Respiratory Pause)

서론



- **다중 화자 휴지 예측**

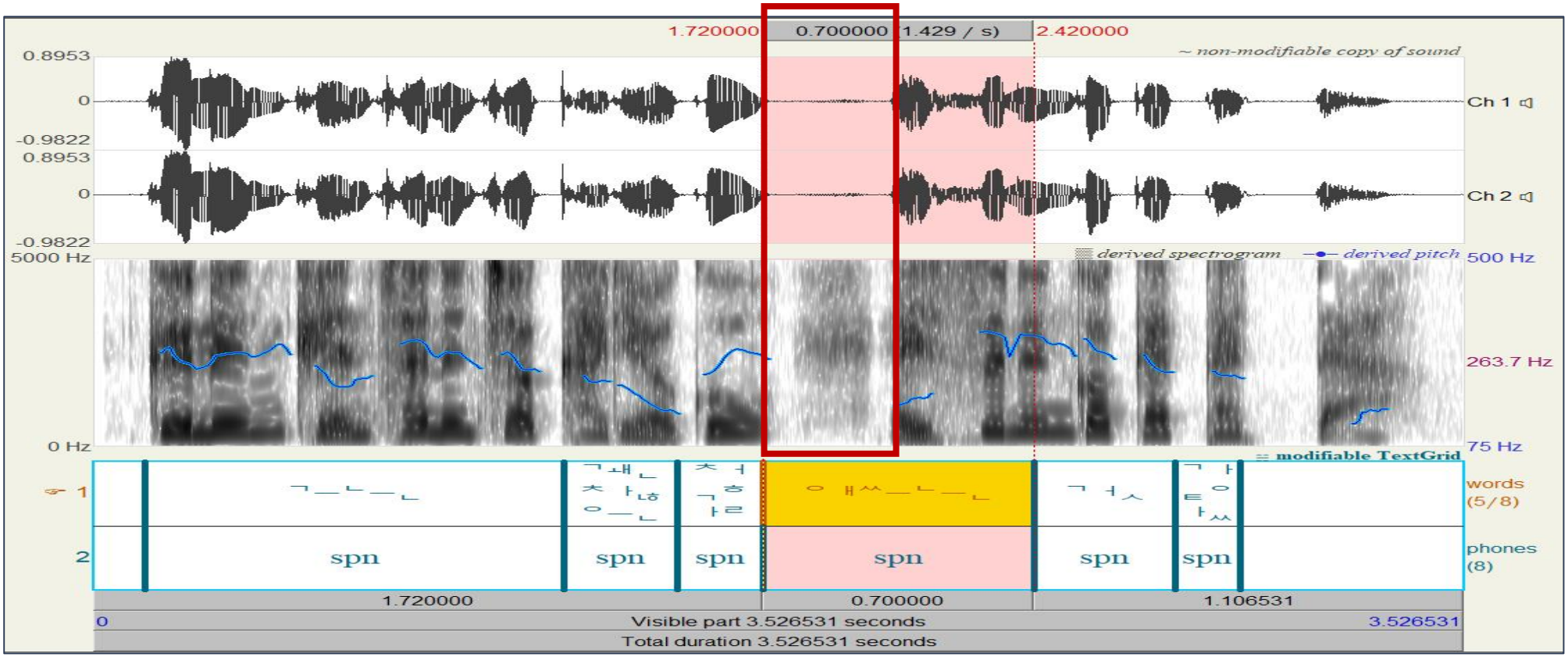
  각 화자의 발화 스타일이나 운율, 억양 등에 따라 휴지(RP) 삽입 방식이 상이함

  화자 정보를 포함하는 휴지 예측 모델 제안

  → 한국어 발화문 임베딩에 화자 특징 벡터를 추가하여 휴지 지속 시간 예측

관련 연구

https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner



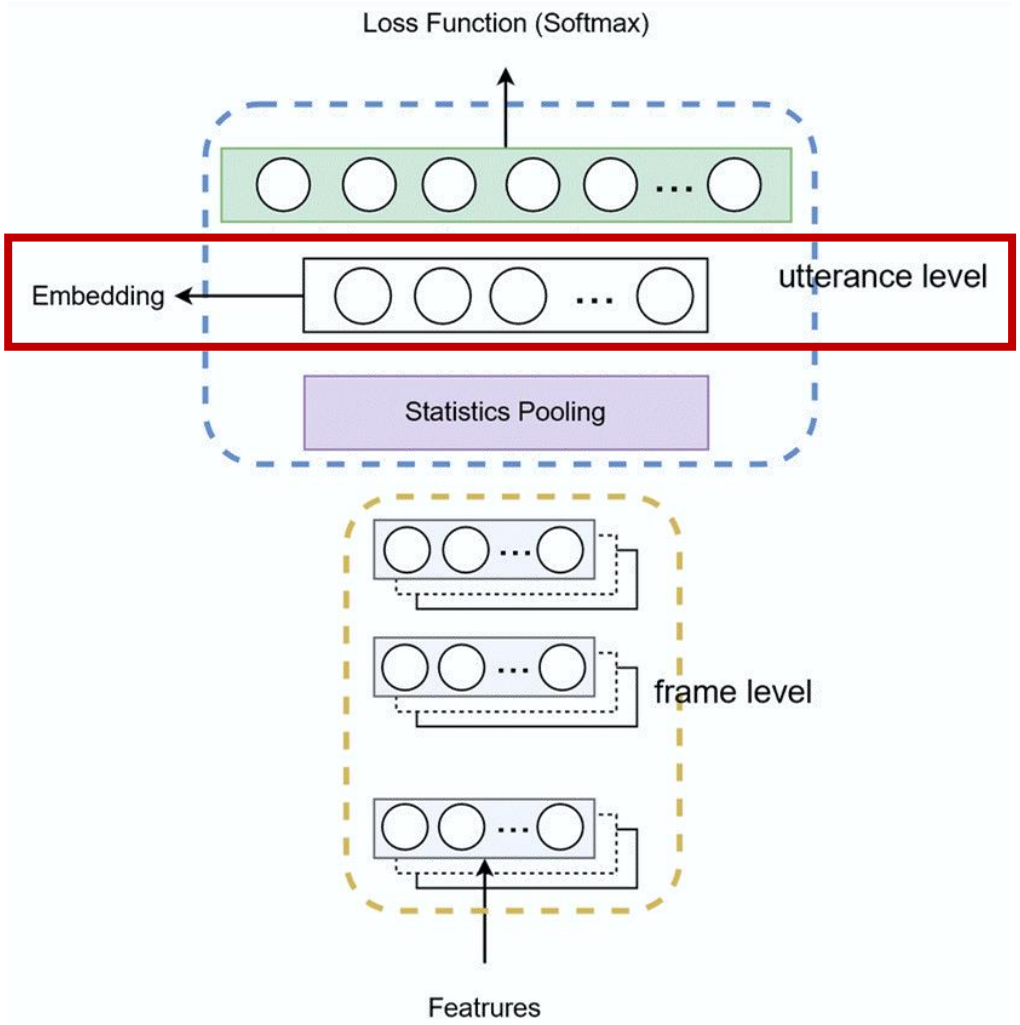
- **Montreal Forced Aligner (MFA)**

  입력된 텍스트-음성 데이터 쌍으로부터 음소 혹은 어절 단위의 정렬 제공

  음소 간의 휴지 또한 포착하여 정렬 가능

관련 연구

DOI:10.1109/ISCSLP.2018.8706575

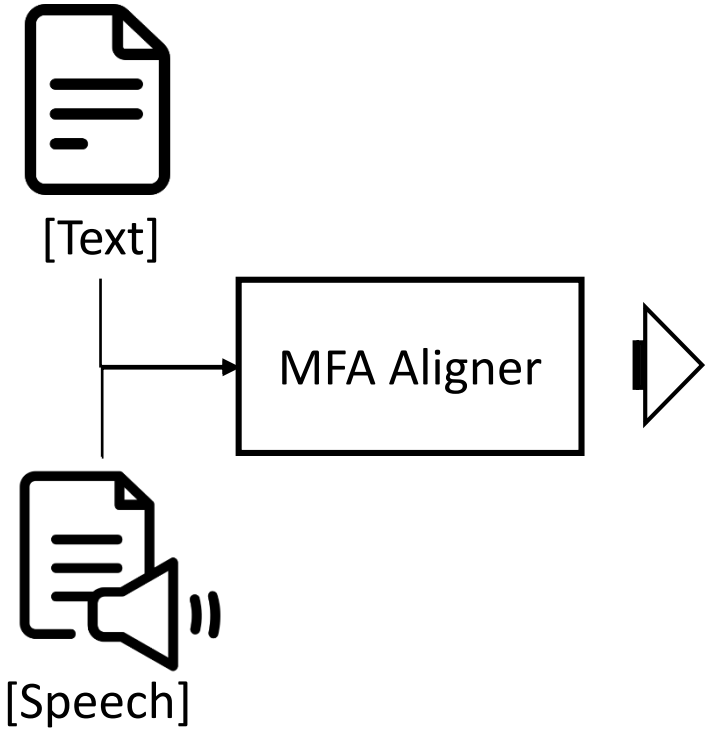


- **화자 특징 벡터**

  딥러닝 기반의 화자 특징 벡터 추출 모델, x-vector 모델 활용

  시간 지연 신경망(TDNN)의 마지막 은닉층을 통계적 풀링(Statistics pooling)

모델



Raw text	졸업을 하고 모든 것을 누릴 수 있어						
Tokenized Input	졸업	##을	하고	모든	…	수	있어
Position of RP (Pause Probability)	0	0	1	0	…	0	1
Duration of RP (Pause Category)	0	0	1	0	…	0	2

- **데이터 전처리**

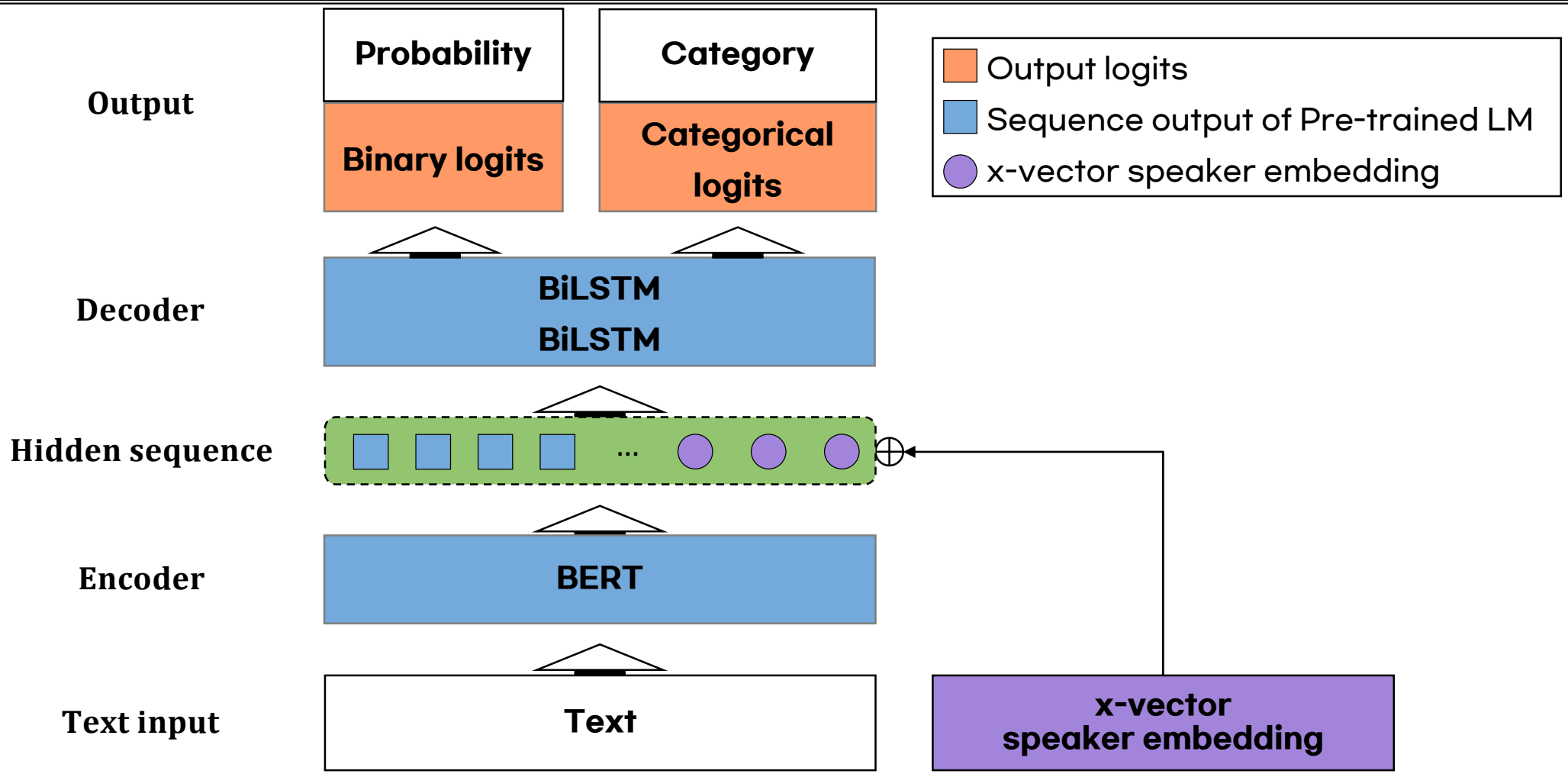
  지속 시간에 따라 RP (Respiratory Pause) category 분류

  ('RP1': <300ms, 'RP2': 300-700ms, 'RP3': >700ms)

  Pause position & category label

  → 발화문 텍스트의 직전 토큰에 위치하도록 조정

모델

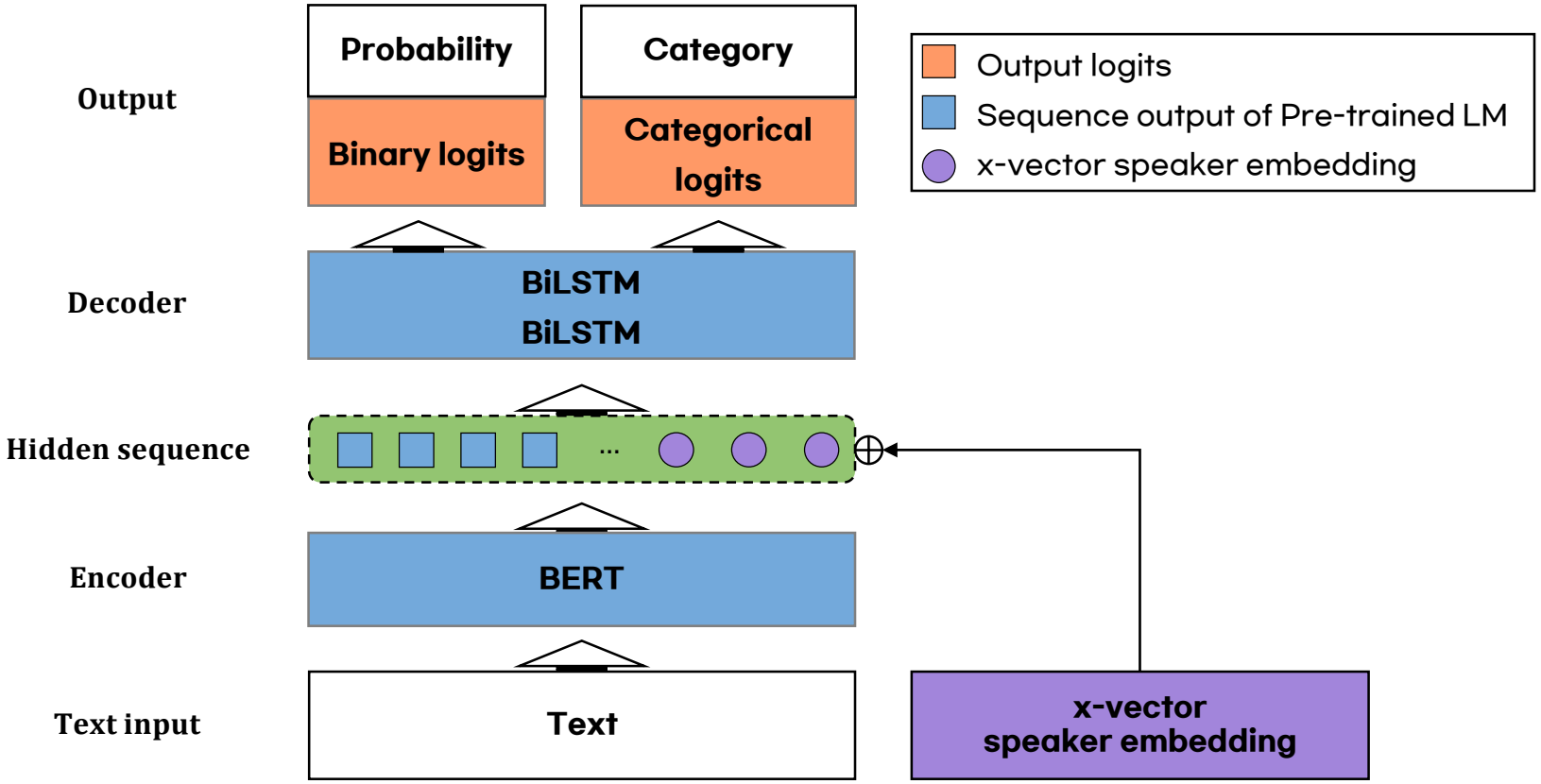


- **모델 구조**

  Pre-trained BERT + BiLSTM

  → BERT-base-multilingual-uncased 모델 Fine-tuning

모델



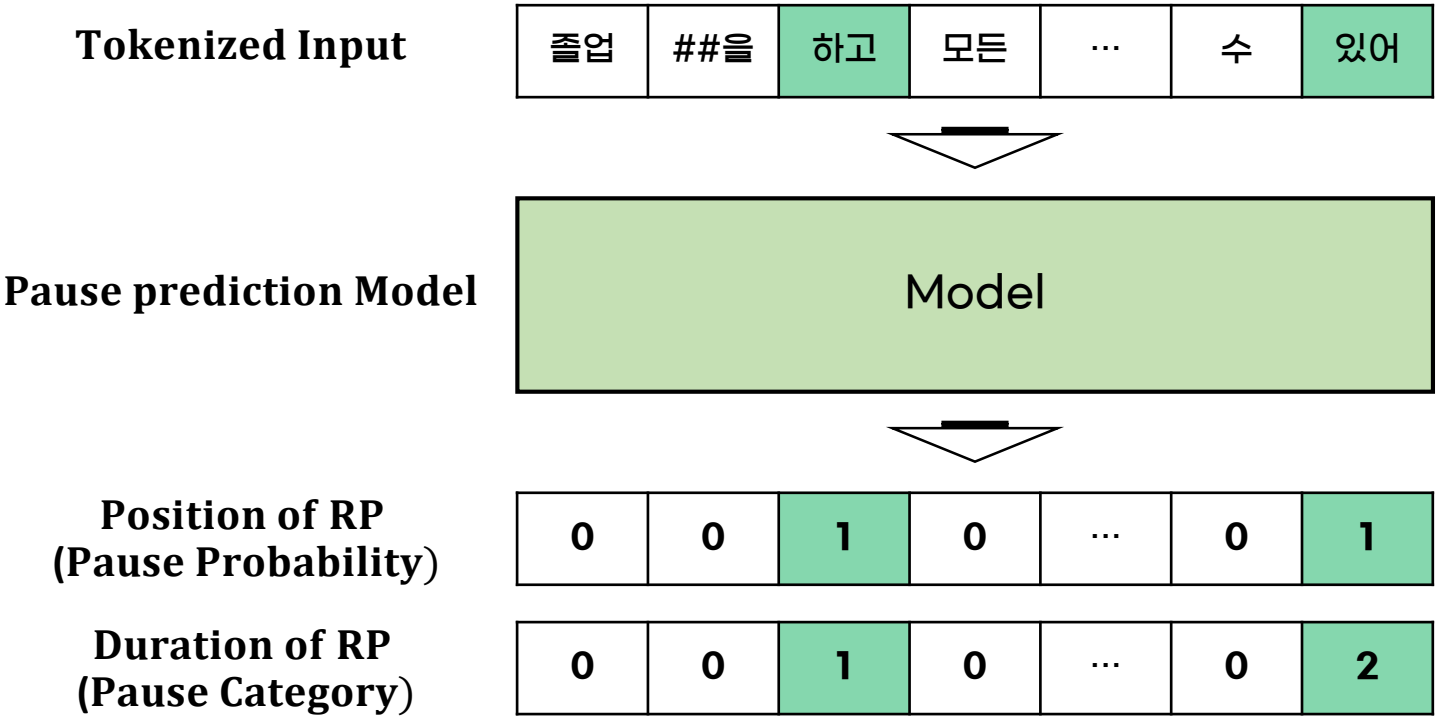
- **모델 구조**

  x-vector 추출 모델로 사전 학습된 Kaldi의 'SRE16 x-vector Model 1a' 모델 활용

  BERT의 Hidden sequence 출력과 x-vector 기반 화자 특징 벡터 연결(concatenate)

  Tokenized text에 기반하여 각 토큰별 휴지 발생 확률과 지속 시간 예측

모델



- **모델 출력**

  카테고리 0를 zero probability로 추정하는 것은 부정확한 예측 결과

  토큰별 휴지 발생 확률을 예측한 뒤 휴지 지속 시간 예측 결과에 반영

실험

https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=466

- **데이터셋**

AI-Hub '감성 및 발화 스타일별 음성합성 데이터' 내의 학습용 데이터셋 활용

약 1,000여 시간, 66만여 건의 데이터를 학습 및 검증, 평가용으로 각각 9:0.5:0.5 비율로 분할

데이터 불균형 문제를 고려한 가중치 설정

카테고리 1-2: 카테고리 0의 수를 전체 토큰의 수로 나눈 값을 가중치로 설정,

카테고리 3: 1.0의 가중치 설정

학습 데이터의 10%에 해당하는 6만여 건의 화자-음성 쌍 데이터로부터 화자별 x-vector 추출

실험

표 1: 휴지 지속 시간 예측 모델 간 성능 비교

Model	Precision	Recall	F_1 score
BERT	0.444	0.409	0.426
BERT+BiLSTM	0.453	0.394	0.422
BERT+BiLSTM w/ speaker embedding	**0.522**	**0.465**	**0.492**

- **실험 결과**

  휴지 카테고리의 수가 불균형한 점을 고려하여 Micro 평균으로 성능 평가

  기존 BERT 혹은 BERT+BiLSTM 모델 대비 큰 성능 향상

  특히 화자 특징 벡터를 추가한 것만으로 F_1 score 기준 약 7%p의 성능 증가

실험

표 2: 화자별 휴지 지속 시간 예측 성능 평균

Model	Avg of Samples	Avg of F_1 score
BERT+BiLSTM	518	0.422
BERT+BiLSTM w/ speaker embedding		0.479

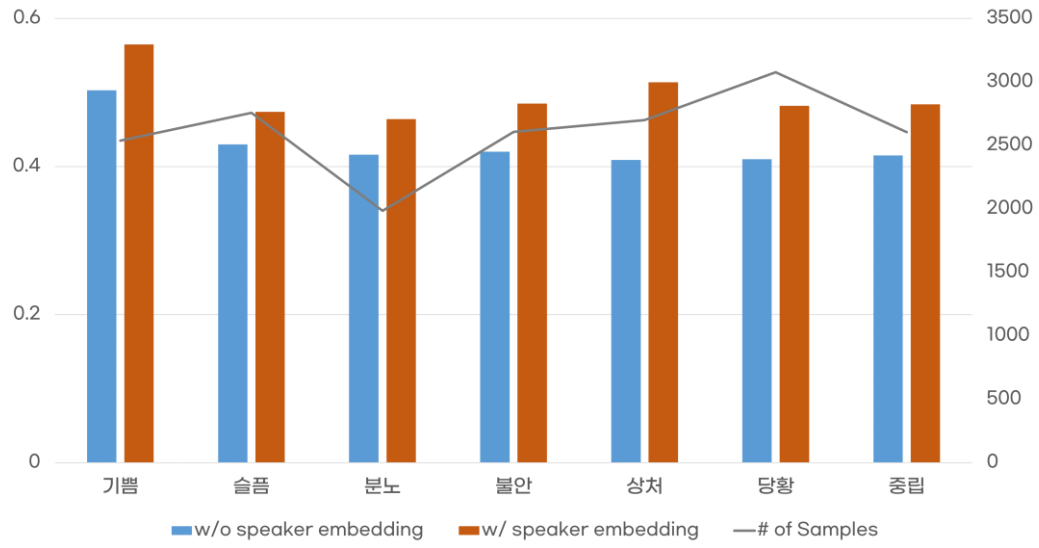
- **실험 결과**

  평가용 데이터: 전체 49인의 화자 당 평균 518건의 발화로 구성

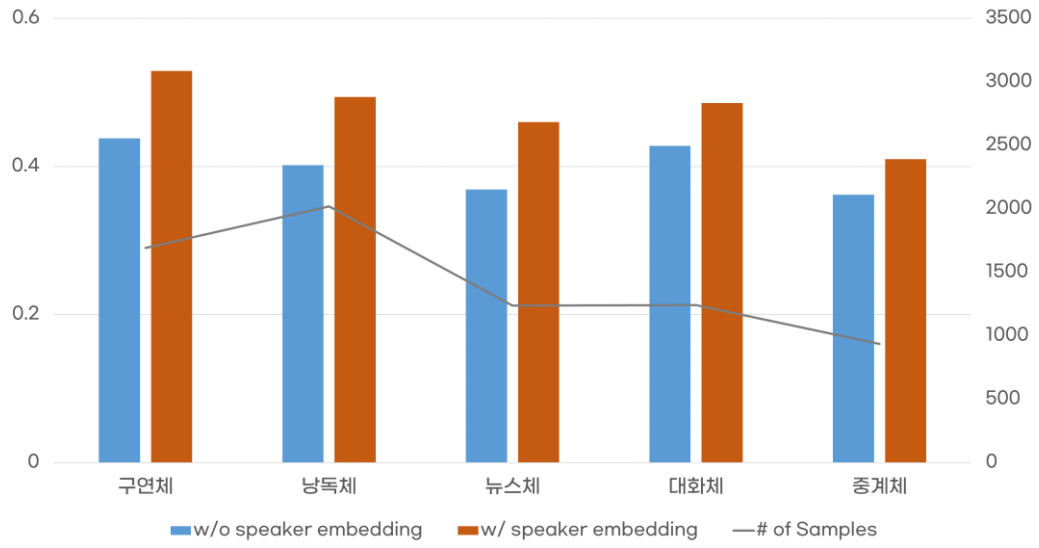
  화자별 모델의 휴지 지속 시간 예측 성능 평균

  화자별 F_1 score 평균값 기준 약 5%p의 성능 증가

실험



발화 감성별 휴지 지속 시간 예측 성능



발화 스타일별 휴지 지속 시간 예측 성능

- **실험 결과**

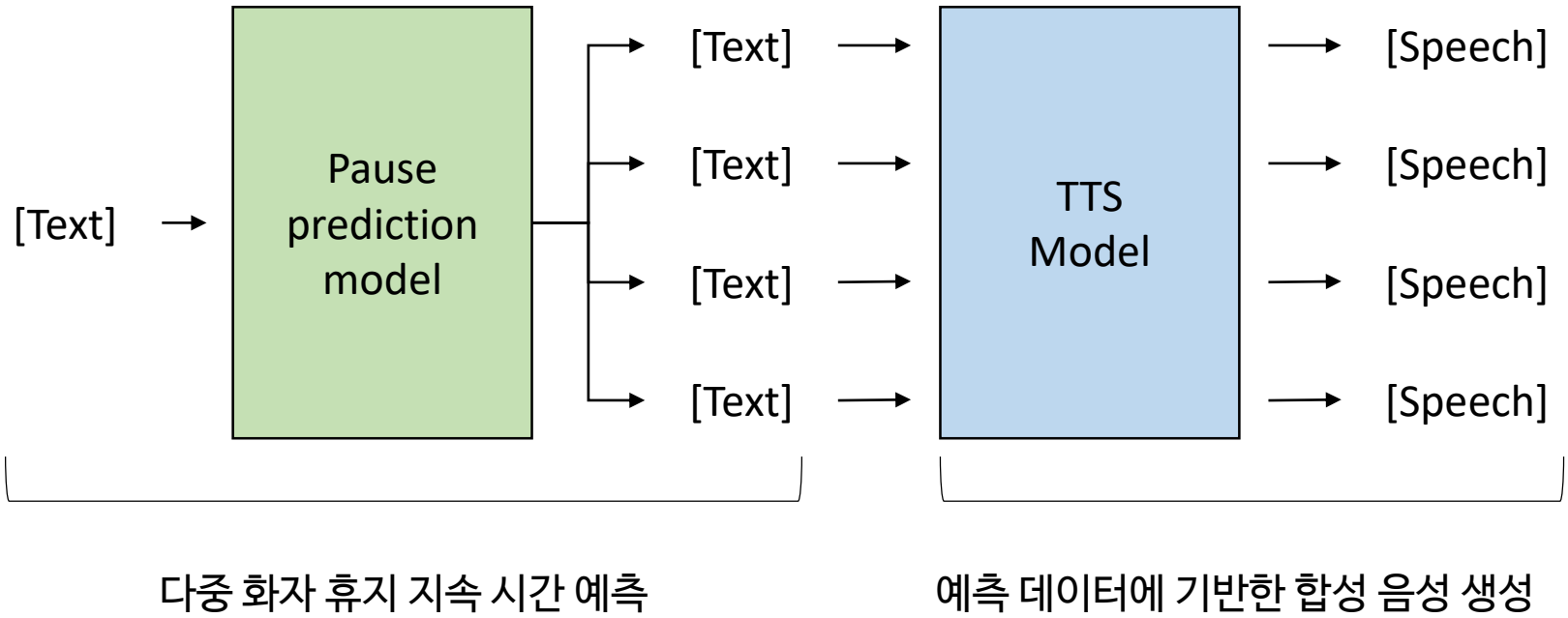
감성 및 스타일별 휴지 지속 시간 예측 결과

7가지 감정: [기쁨, 슬픔, 분노, 불안, 상처, 당황, 중립]

5가지 발화 스타일: [구연체, 낭독체, 뉴스체, 대화체, 중계체]

데이터 양이 증가할수록 성능 향상 폭이 증가하는 경향 보임

실험



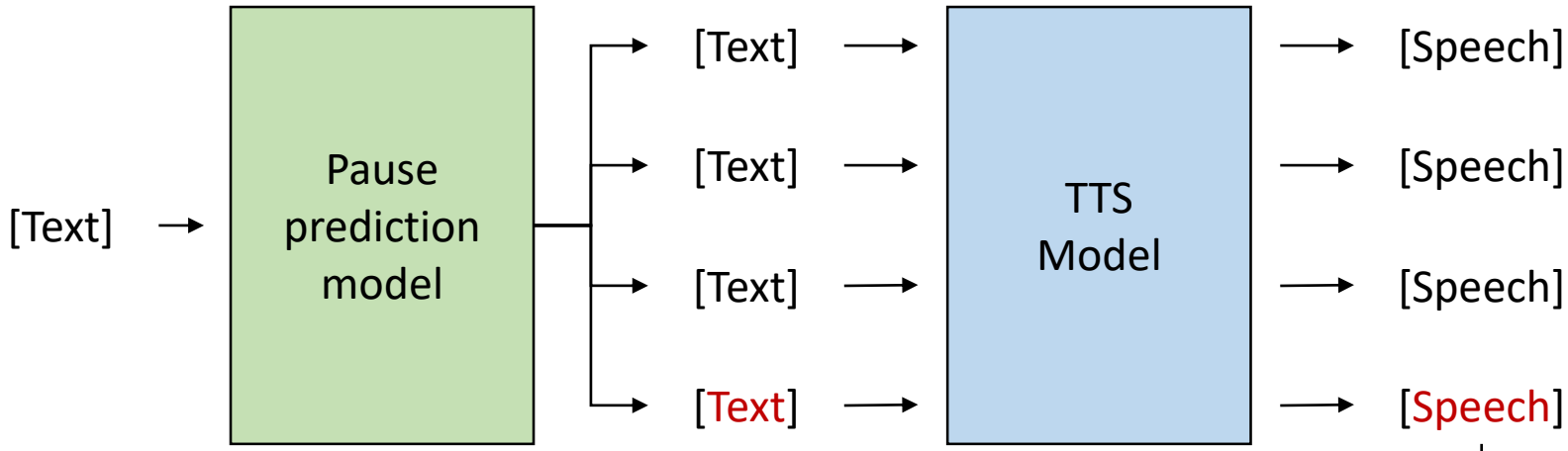
- **TTS 모델 연계 (2nd stage)**

  제안한 모델을 통해 생성된 데이터를 한국어 TTS 모델에 입력 → 합성 음성 생성

  Korean Single Speaker Speech Dataset(KSS) 데이터셋으로 학습한 Fastspeech2 모델 테스트

  = 약 12시간 정도의 소규모 음성-텍스트 데이터셋

실험



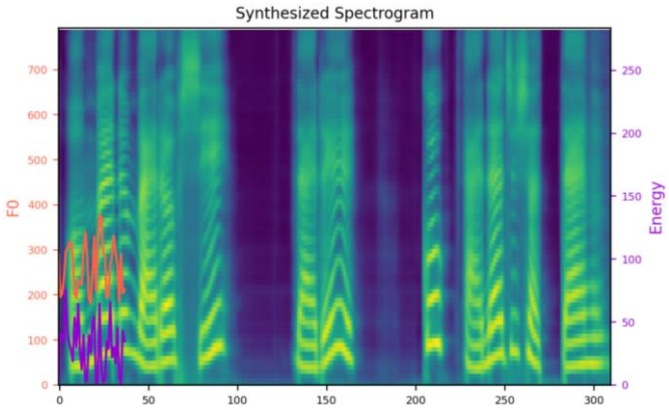
- **음성 합성 예제**
  - 엄마는 말없이 나를 꼭 안아주셨다.

    ↓

  - 엄마는 말없이 [sp3] 나를 [sp2] 꼭 [sp1] 안아주셨다.

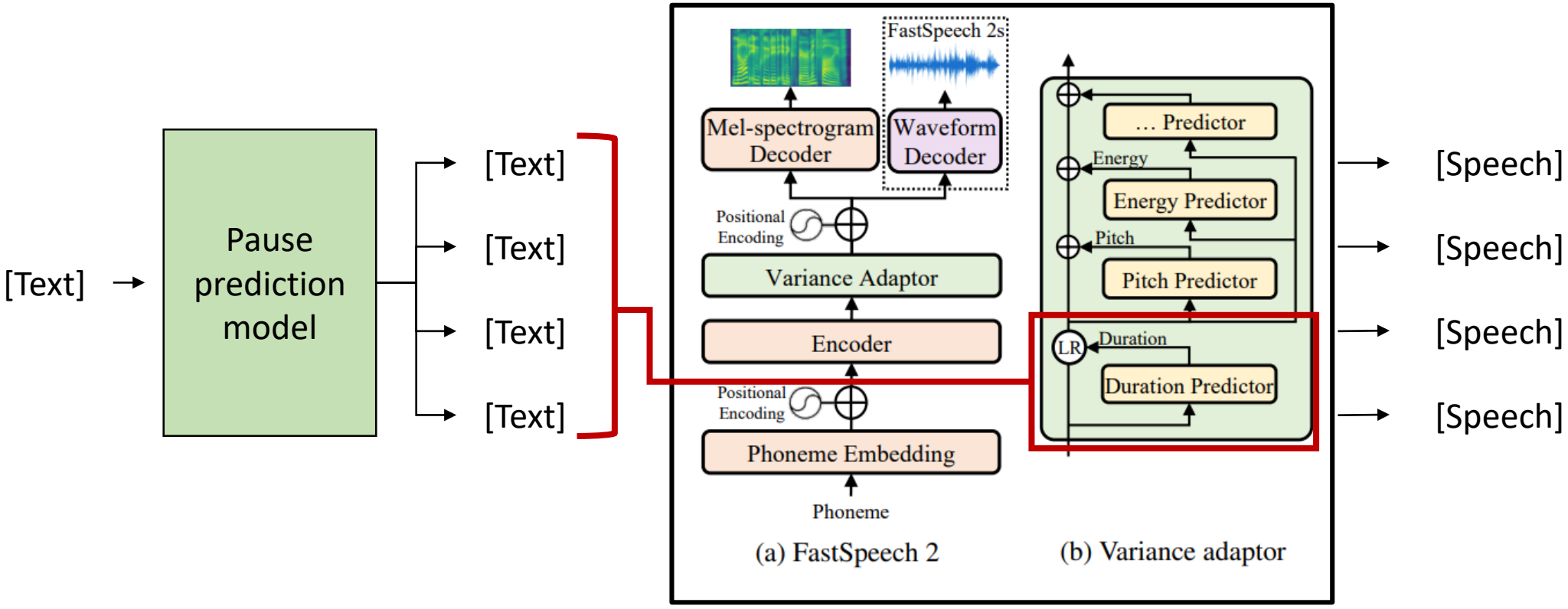
    ↓

  - Synthesized spectrogram:



결론 및 향후 연구

https://arxiv.org/abs/2006.04558



TTS model

- **LLM 확장 및 TTS 모델 연계 강화**

  화자 특징 벡터의 효용성과 휴지 예측 성능 확인함

  대규모 언어 모델(Large Language Model)로 확장하여 모델 성능 증진 기대

  Fastspeech2 모델 내부의 Duration predictor로의 injection을 통해 TTS 모델과의 연계 강화

QnA

감사합니다