자기지도학습 기반 음성 언어 모델을 이용한 자소 단위의 한국어 음성 인식

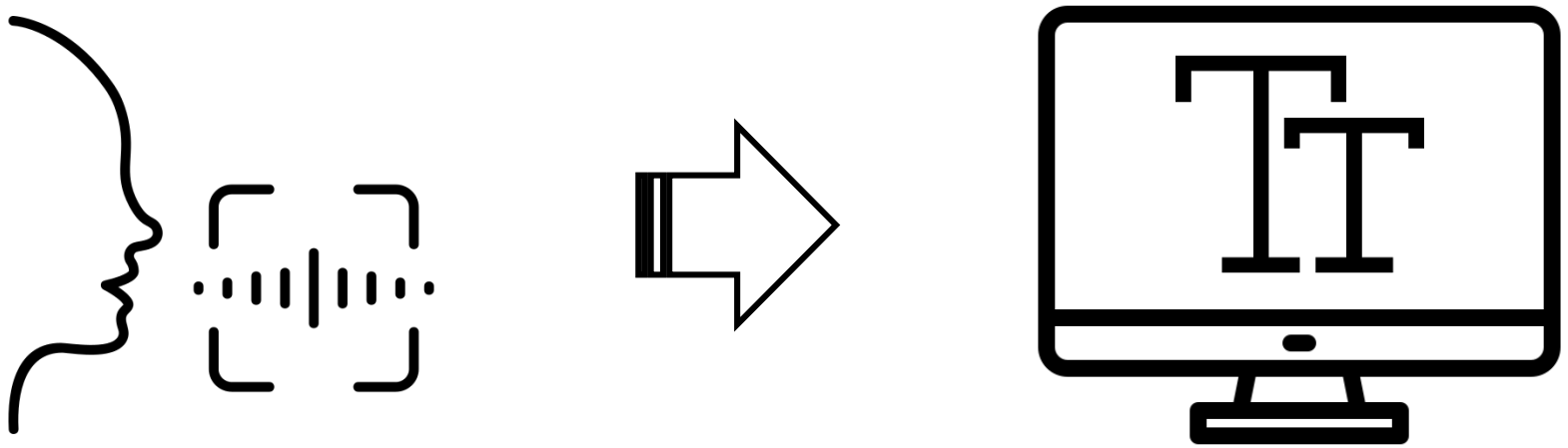
Grapheme-level Automatic Speech Recognition of Korean using Self-supervised Spoken Language Model

이 정°[1], 서민택[2], 나승훈[3], 나민수[4], 최맹식[5], 이충희[6]

전북대학교[1,2,3], ㈜엔씨소프트[4,5,6]

▪ 목차

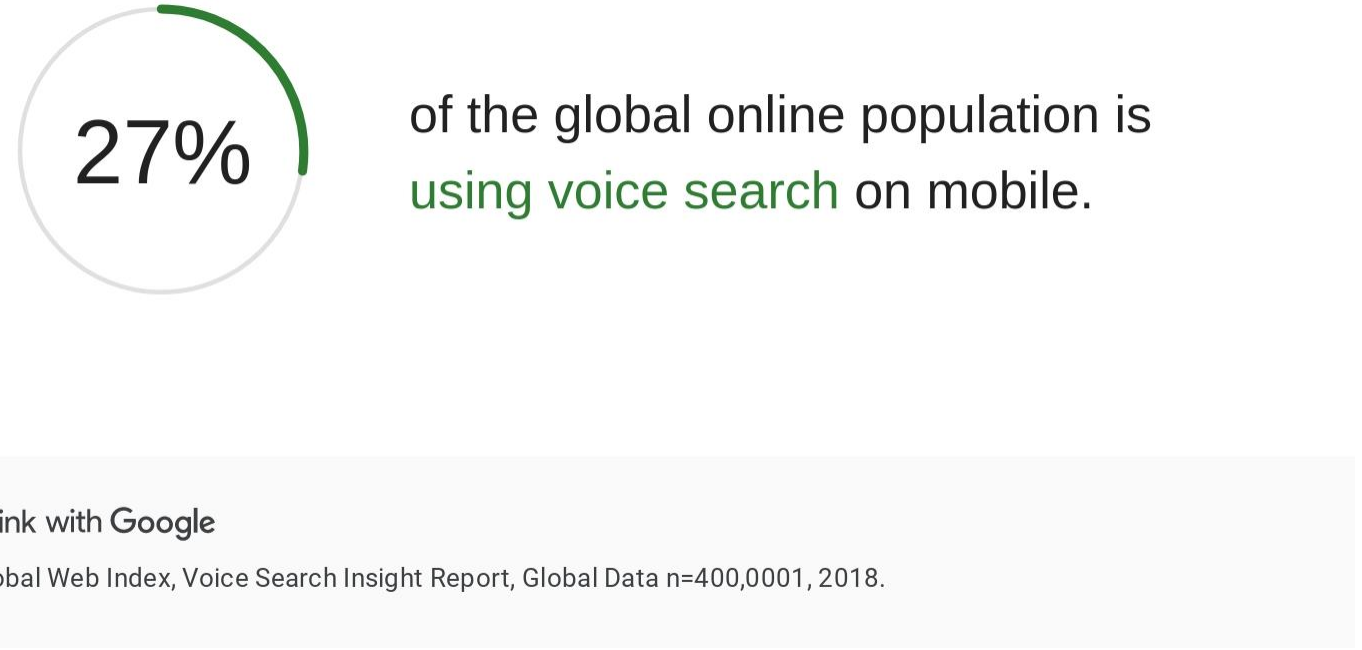
▪ 서론



- 자동 음성 인식 (Automatic Speech Recognition, ASR)
  = Speech-to-Text(STT)

사람이 말하는 음성 언어를 텍스트 데이터로 전환하는 일련의 처리나 기술
ASR 기술의 발전으로 음성 기반의 인터페이스를 통한 상호 작용이 확대되는 추세

▪ 서론



- **음성 기반 인터페이스를 통한 상호 작용**

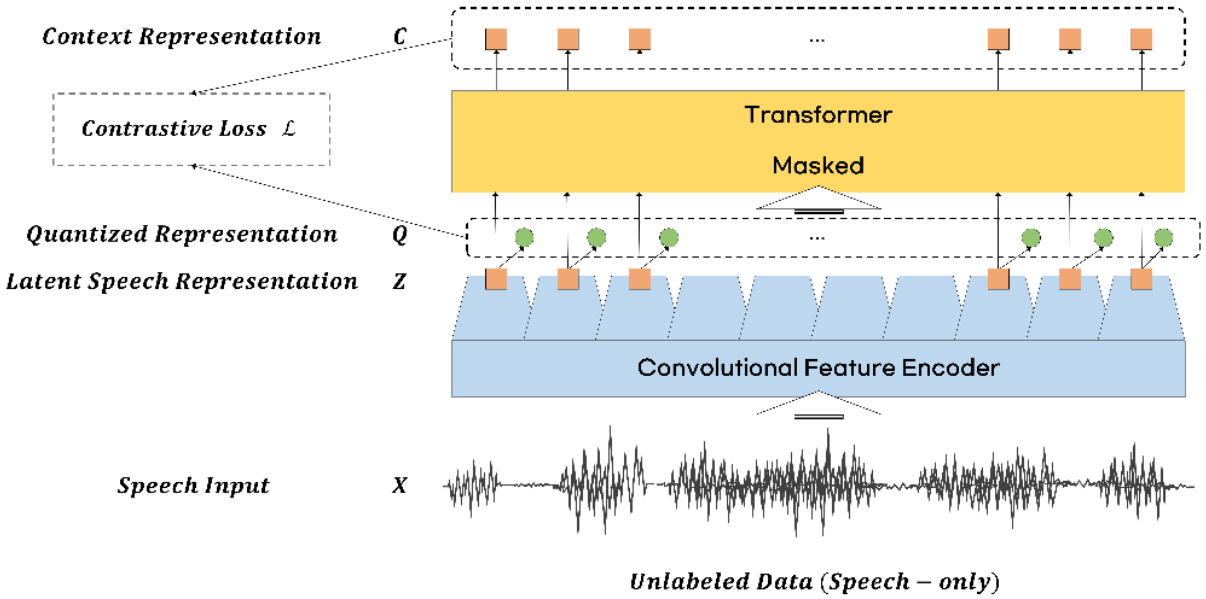
  구글이 발표한 자료에 따르면 모바일 검색에서 음성 검색을 활용하는 비율이 27%에 도달

  다양한 산업 전반에 걸쳐 적용되고 있는 음성 기반 인터페이스

  ➔ 최근 CNN 기반의 음성 특징 추출 기법을 포함한 딥러닝 기술과의 결합으로 성능 증대

https://www.thinkwithgoogle.com/marketing-strategies/search/voice-search-mobile-use-statistics/

▪ 모델 소개



- XLS-R (wav2vec 2.0)

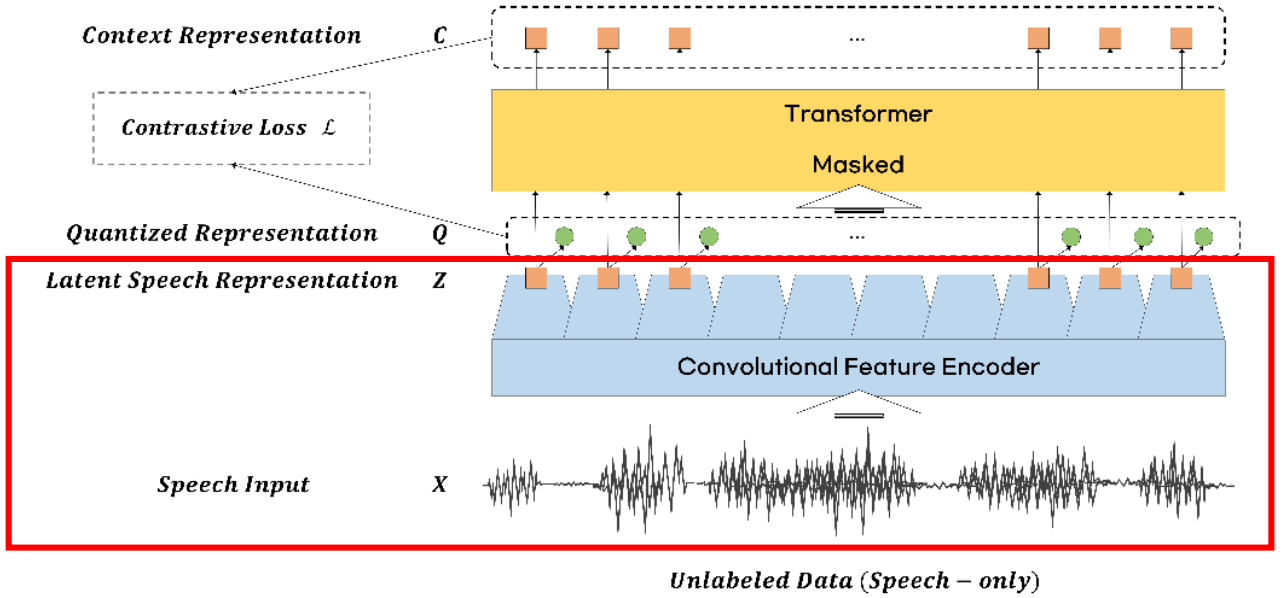
  자기지도학습을 음성 인식에 활용한 대표적인 모델 wav2vec 2.0을 기반으로 구성됨

  XLS-R 모델은 기존 wav2vec 2.0 대비 학습 데이터를 증대시켜 한국어 음성 데이터 61시간을 포함한 약 436,000 여 시간의 다국어 음성 데이터로 학습을 수행

  모델의 매개변수 또한 300M, 1B, 2B로 크게 증가함

XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale [Babu et al. `20]

▪ 모델 소개

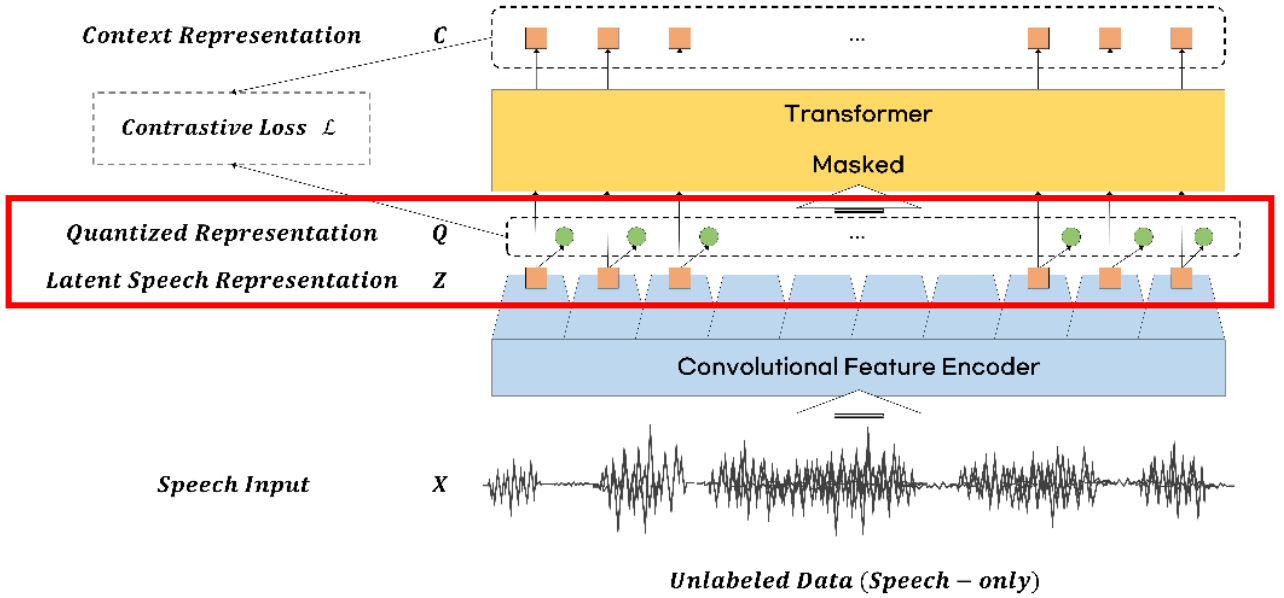


- **모델 구조: $X \mapsto Z$**

  Convolutional Feature Encoder에 Speech Input X을 입력하여 전체 Timestep T에 대한 Latent Speech Representation $Z = [z_1, z_2, \dots, z_T]$ 획득

XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale [Babu et al. `20]

▪ 모델 소개



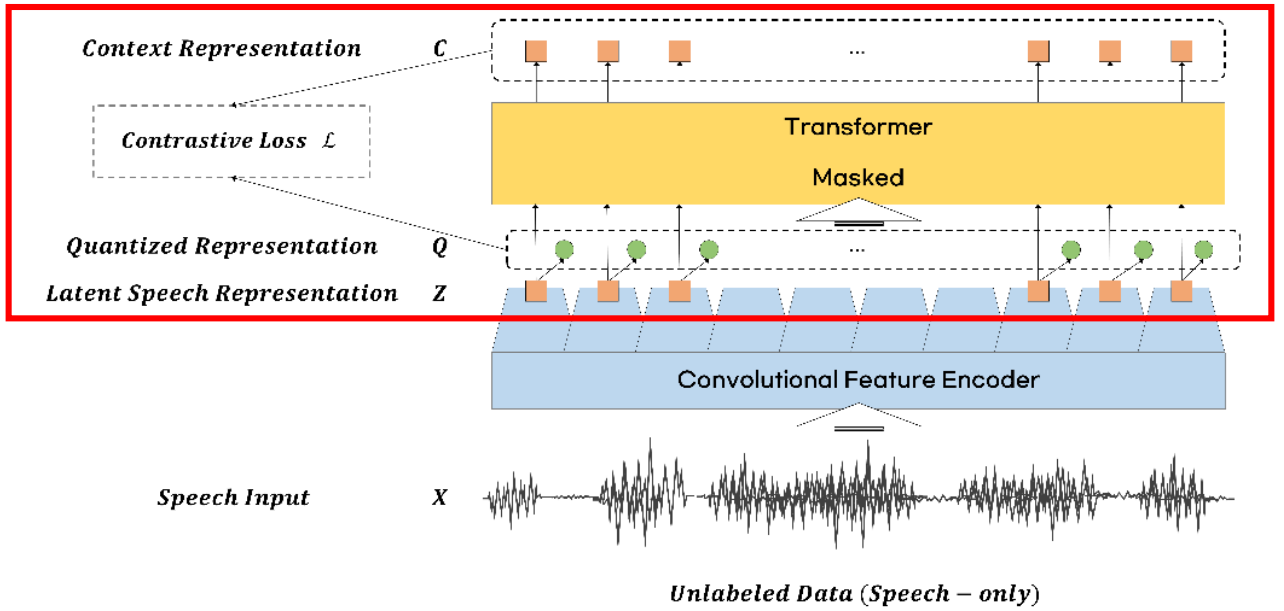
- **모델 구조: $Z \mapsto Q$**

  G개의 Codebook으로부터 Codeword vector $e_1, \dots, e_G \in \mathbb{R}^{d/G}$를 추출하고 이들을 연결하여 얻은 벡터 $e_t \in \mathbb{R}^d$ 에 선형변환($\mathbb{R}^d \mapsto \mathbb{R}^f$)을 수행 (Product Quantization)

  Latent Speech Representation Z 로부터 한정된 집합의 Quantized Representation Q를 얻음

XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale [Babu et al. `20]

▪ 모델 소개

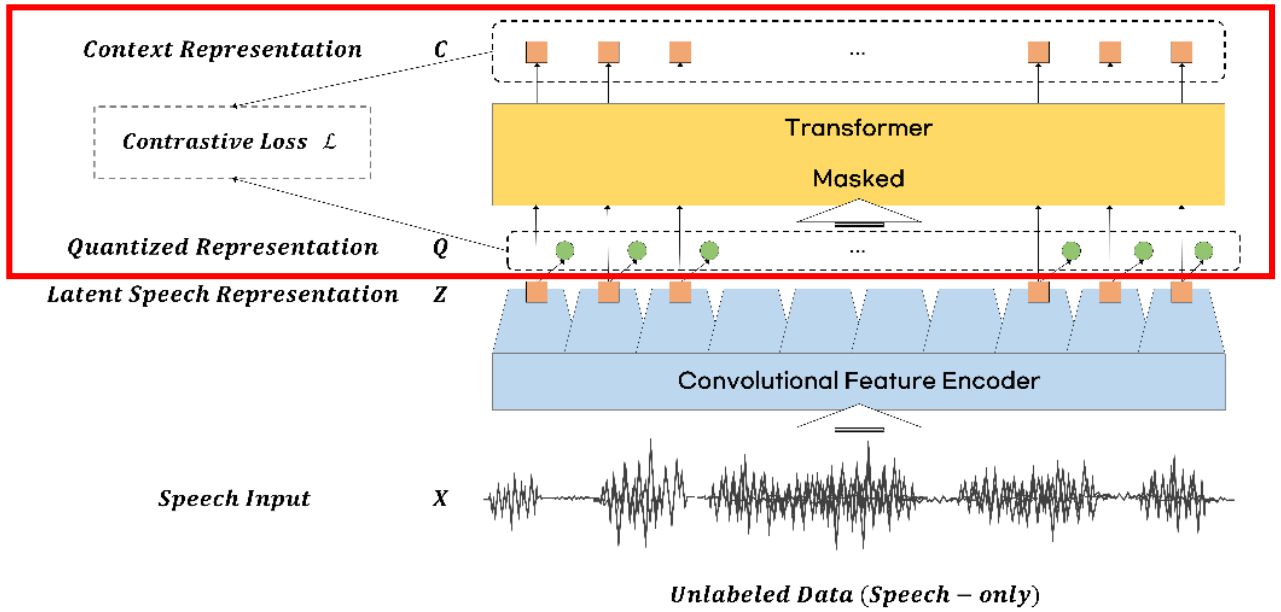


- **모델 구조: $Z \mapsto C$**

  Transformer Encoder를 거쳐 Latent Speech Representation Z 로부터 Context Representation $\mathbf{C} = [c_1, c_2, \dots, c_T]$를 얻도록 학습

XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale [Babu et al. `20]

▪ 모델 소개



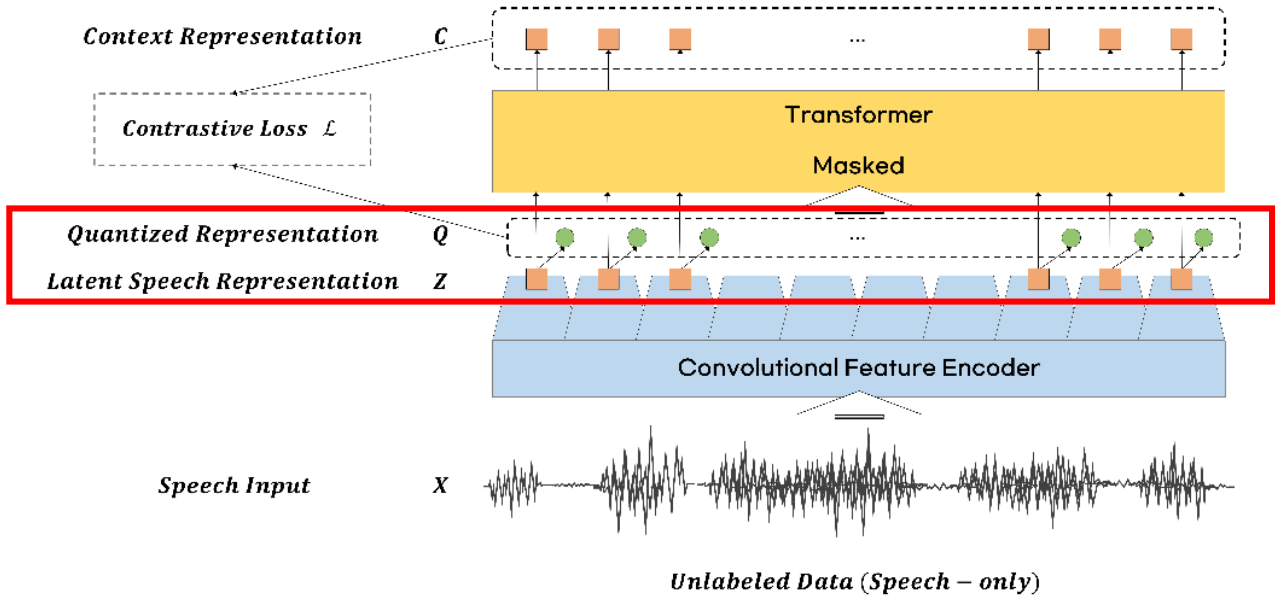
$$\mathcal{L}_m = -log \frac{\exp(sim(c_t, q_t)/\kappa)}{\sum_{\tilde{q} \sim Q_t} \exp(sim(c_t, q_t)/\kappa)}$$

- **Contrastive Loss**

  각 Timestep t에서 얻은 Quantized Representation q_t와 Context Representation c_t에 따른 Contrastive Loss ($\mathcal{L}_m$)를 위 수식과 같이 계산

  - sim: 코사인 유사도 함수

▪ 모델 소개



$$H(X) = \sum_{k} p(X = k) log\, p(X = k)$$

$$\mathcal{L}_d = \frac{1}{GV} \sum_{g=1}^{G} -H(\tilde{p}_g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \tilde{p}_{g,v} log\, \tilde{p}_{g,v}$$

- **Diversity Loss**

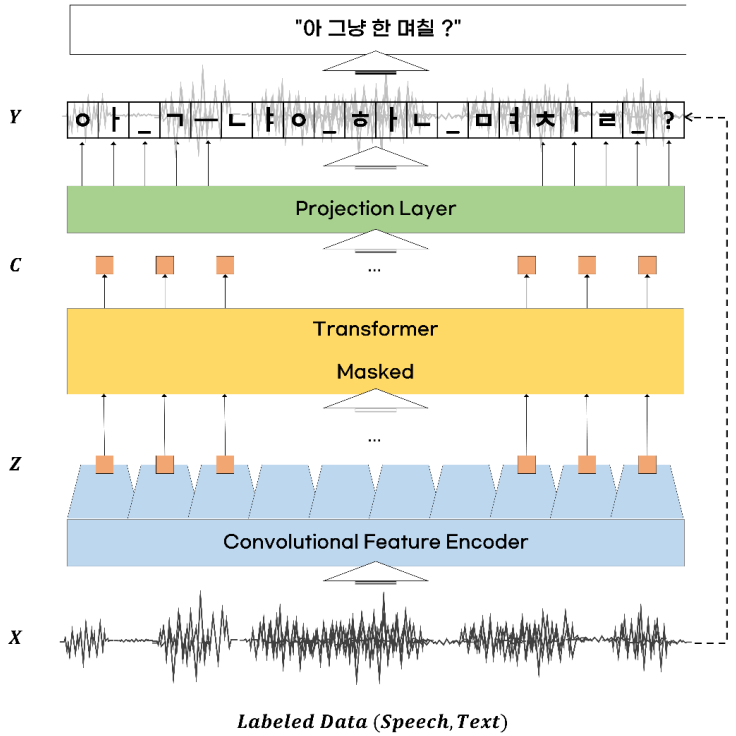
  Product Quantization을 수행하는 과정에서 Codebook 내의 다양한 Codeword가 균등하게 선택될 수 있도록 정보 엔트로피 $H(X)$를 도입하여 Diversity Loss 계산

  모델의 손실함수 $\mathcal{L} = \mathcal{L}_m + \alpha\mathcal{L}_d$ 계산

  - G, V: 전체 Codebook과 Codebook 내의 Codeword 수
  - α: 모델의 하이퍼 파라미터

XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale [Babu et al. `20]

▪ 한국어 음성 인식



Labeled Data (Speech, Text)

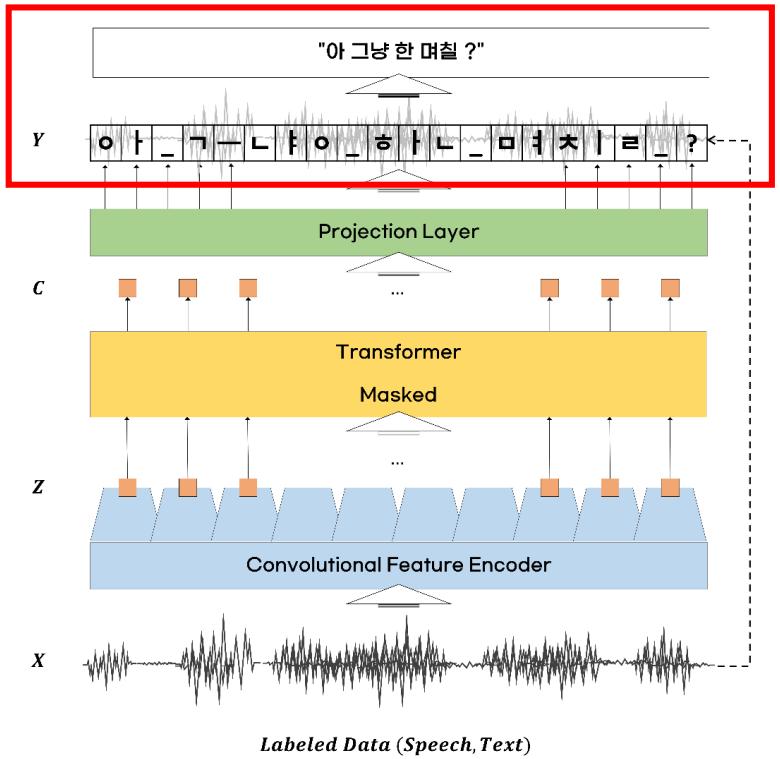
- **데이터 셋**

  모델의 학습 및 평가를 위해 AI Hub의 KsponSpeech (한국어 음성, 1000h) 데이터 셋 활용

  음성 데이터와 음성 전사 텍스트 데이터 입력

  텍스트 데이터의 경우 자소 단위로 분절하여 각 음성 프레임 단위로 예측하도록 학습

▪ 한국어 음성 인식



Labeled Data $(Speech, Text)$

<1 epoch 음성 인식 성능 비교, WER>

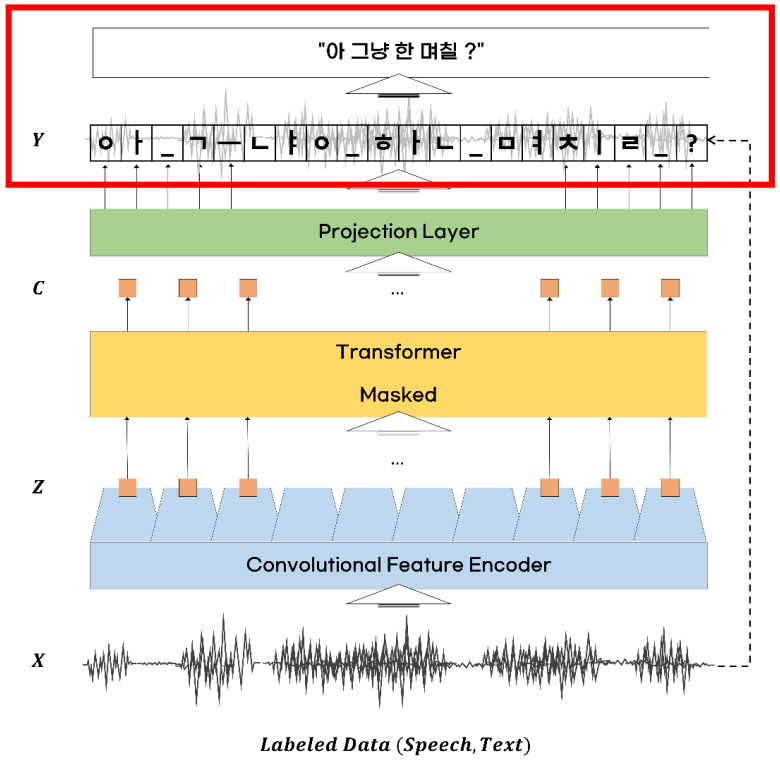
Vocab	Character-level	Grapheme-level
WER	0.459	**0.413**

- **실험 결과**

  음절 단위와 자소 단위의 한국어 음성 인식 모델 성능 (WER, Word Error Rate) 비교

  자소 단위의 모델의 경우 디코딩 수행 후 띄어쓰기를 포함한 원본 문장으로 재구성

▪ 한국어 음성 인식



Labeled Data (Speech, Text)

<학습 데이터 비율 조정에 따른 성능, WER>

Data	30%	50%	100%
WER	0.722	0.520	**0.413**

<학습 시간에 따른 성능, WER>

Epoch	1	3
WER	0.413	**0.355**

- **실험 결과**

자소 단위의 한국어 음성 인식 모델 성능 (WER)

자기지도학습 방식의 이점을 확인하기 위하여 학습 데이터 셋의 크기 별 모델 성능 비교

또한 Epoch 수를 늘렸을 때 성능이 확연히 향상된 모습을 보여 추가적인 학습을 통한 성능 증진 기대

▪ 결론

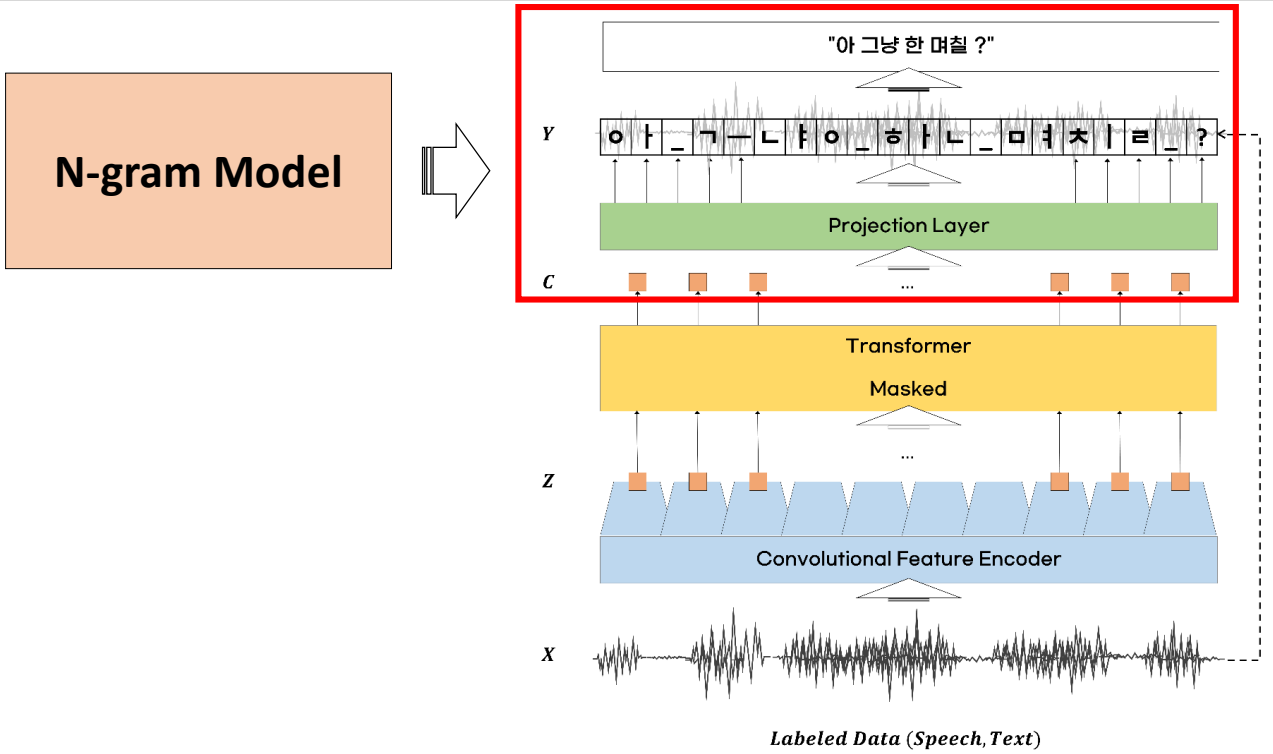
<음성 인식 결과 예시 및 WER 비교>

	Sentence	WER
Ref.	"지나칠수가 없지"	0.666
Pred.	"지나칠 수가 없지"	
Ref.	"어쩔 수 없어 음 그럼 언제 가냐고"	0.428
Pred.	"어쩔수 없어 그럼 언제 가냐고"	
Ref.	"농사 짓고 막 그랬잖아"	0.25
Pred.	"농사 지고 막 그랬잖아"	

- **결과 분석**

소규모 학습 데이터 및 자원을 활용하여 학습한 모델도 일정 수준의 음성 인식이 가능함을 보여 자기지도학습에 기반한 자소 단위의 한국어 음성 인식 모델의 효용성을 입증함.

▪ 결론



- **Future works**

현재 모델은 Context Representation C 로부터 가장 높은 확률의 Output을 취하는 CTC greedy decoding을 수행하나 자소 단위의 음성 인식 모델 특성 상 N-gram 모델 등에 기반한 Joint Decoding 도입하여 성능 증진 예정

감사합니다