클로바노트의 두뇌: 네이버의 End-to-End 음성인식 엔진 소개

정민규
NAVER Cloud Speech
김한규
NAVER Cloud Speech

NAVER DEVIEW 2023

CONTENTS

1.음성인식이란?

1.1 음성인식(ASR)이란?

NAVER
DEVIEW
2023

NAVER CLOUD PLATFORM

1.1 음성인식(ASR)이란?

NAVER
DEVIEW
2023

12시에 몇 분으로 예약해 드릴까요?

1.1 음성인식(ASR)이란?

NAVER
DEVIEW
2023

음성인식 (ASR)

자연어 이해 (NLU)

예약 문의

12시에 몇 분으로 예약해 드릴까요?

음성 합성 (TTS)

1.1 음성인식(ASR)이란?

NAVER
DEVIEW
2023

Automatic Speech Recognition (ASR) / Speech-To-Text (STT)

음성신호를 텍스트로 변환시키는 모델

음성신호에서 길이가 T인 feature를 추출: $\boldsymbol{x} = (x_1, x_2, \cdots, x_T)$

- frame 마다의 feature x_1는 D-dimensional vector. (MFCC, Mel spectrum, raw waveform, ...)

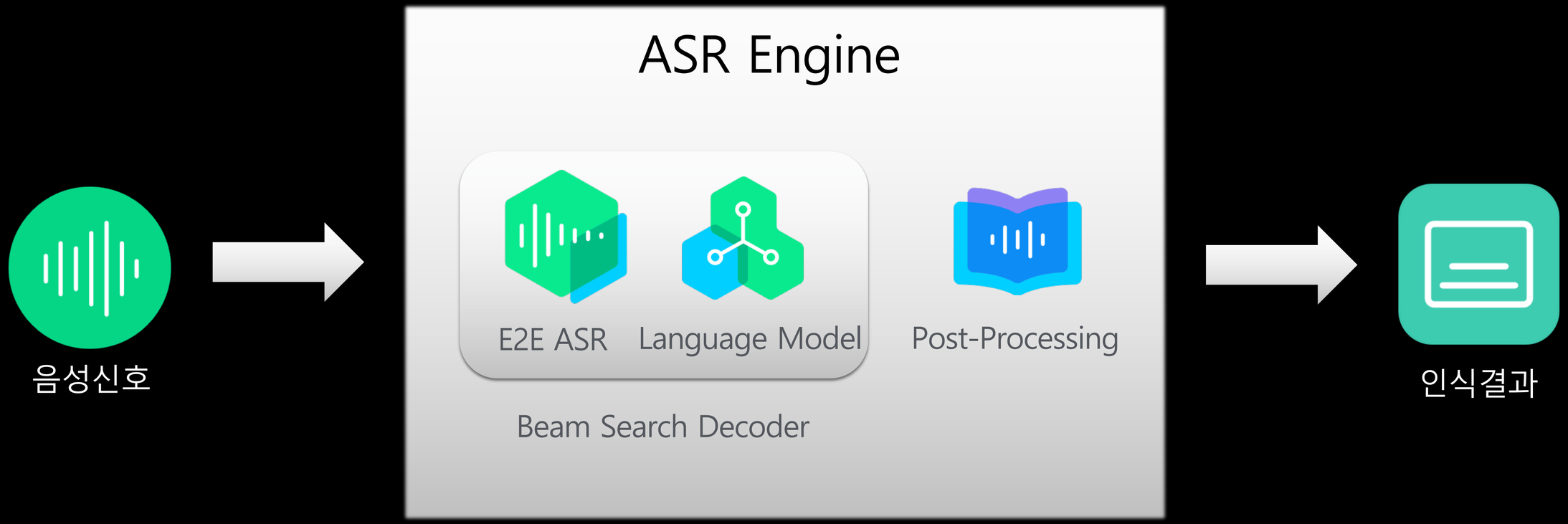
이 음성신호에 대응될 수 있는 문자열 $\boldsymbol{y} = (y_1, y_2, \cdots, y_U)$.

ASR 모델은 주어진 input $\boldsymbol{x}$와 확률적으로 가장 가까운 문자열 $\hat{\boldsymbol{y}}$을 찾는 모델:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P_{ASR}(\boldsymbol{y}|\boldsymbol{x})$$

1.2 NAVER End-to-End 음성인식 구조

NAVER
DEVIEW
2023



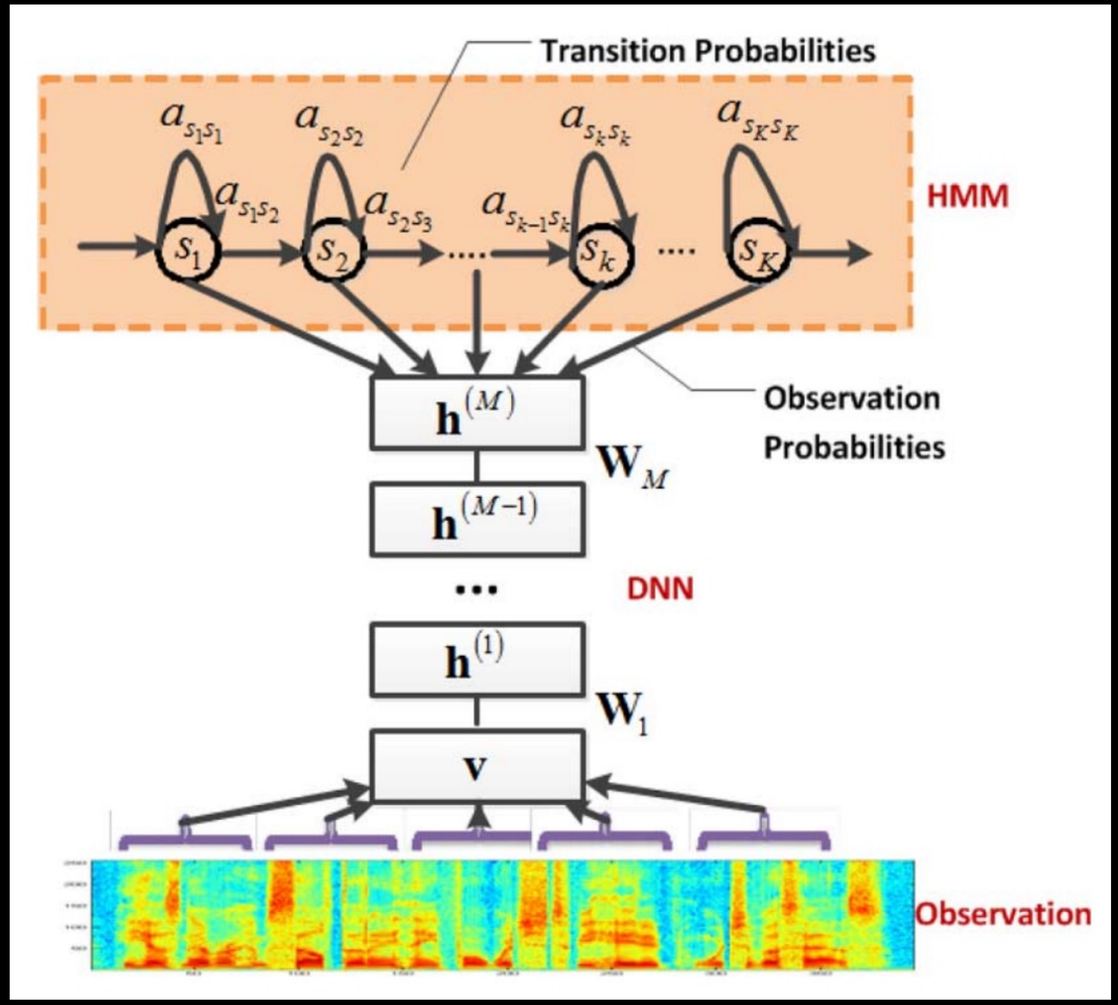
2.사용자의 목소리를 듣는 End-to-End ASR 모델

2.1 전통적인 ASR 모델

NAVER
DEVIEW
2023

전통적인 ASR: DNN-HMM 모델

- 신경망을 사용하기는 하는데...
- HMM이라는 알고리즘도 같이 써야 하고
- 발음기호로 변환도 해야 하고
- 학습 과정도 복잡하고
- ...



George E. Dahl et al., Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, IEEE Trans on Audio, Speech and Language Processing, 2012

2.1 ASR 신경망 모델

NAVER
DEVIEW
2023

1. 발음기호 변환 등 전처리 작업이 필요 없고
2. 여러 알고리즘 대신 신경망 모델만 사용하며
3. 복잡한 과정 없이 한번에 학습이 가능한

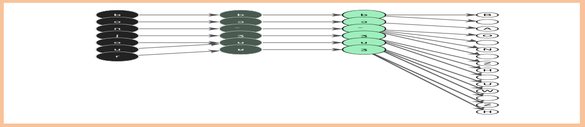
ASR 모델을 만들 수는 없을까?

→ End-to-End 신경망 모델!

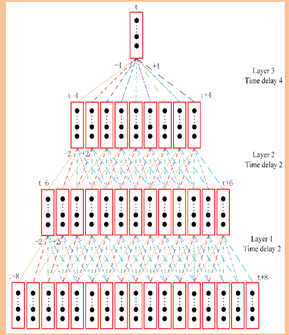
2.1 End-to-End ASR 신경망 모델

NAVER DEVIEW 2023

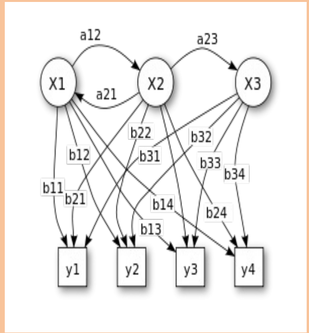
Grapheme to Phoneme



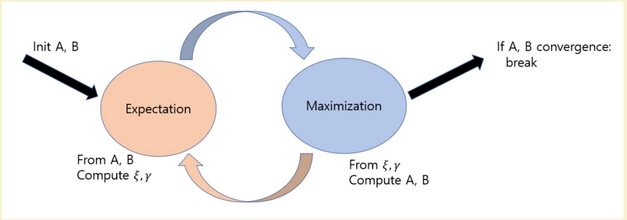
신경망 모델



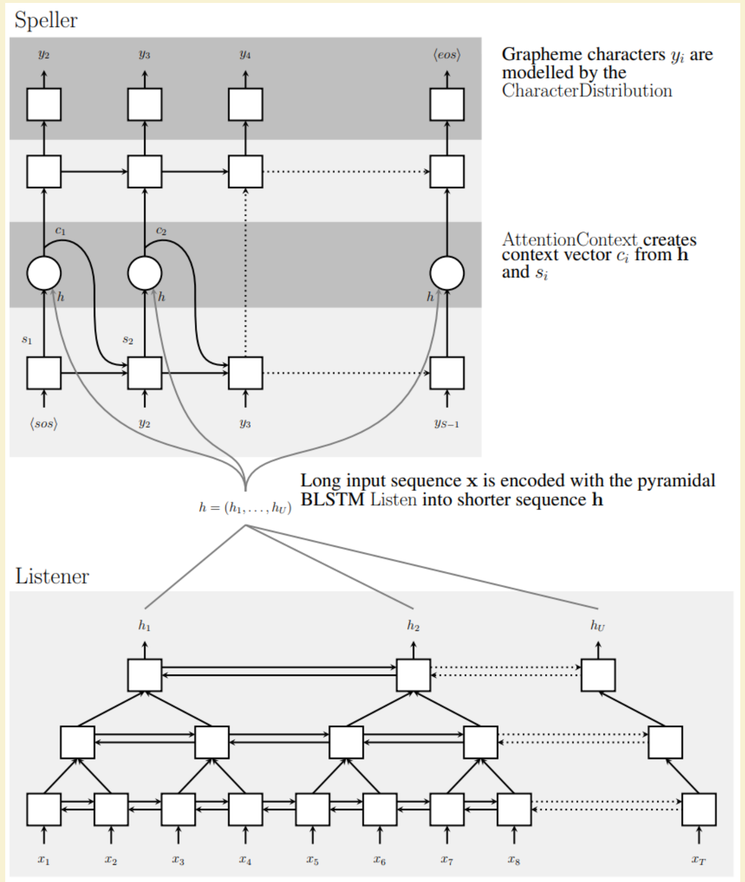
HMM



Baum-Welch 학습



End-to-End ASR 모델



William Chan et al., Listen, attend and spell, ICASSP 2015

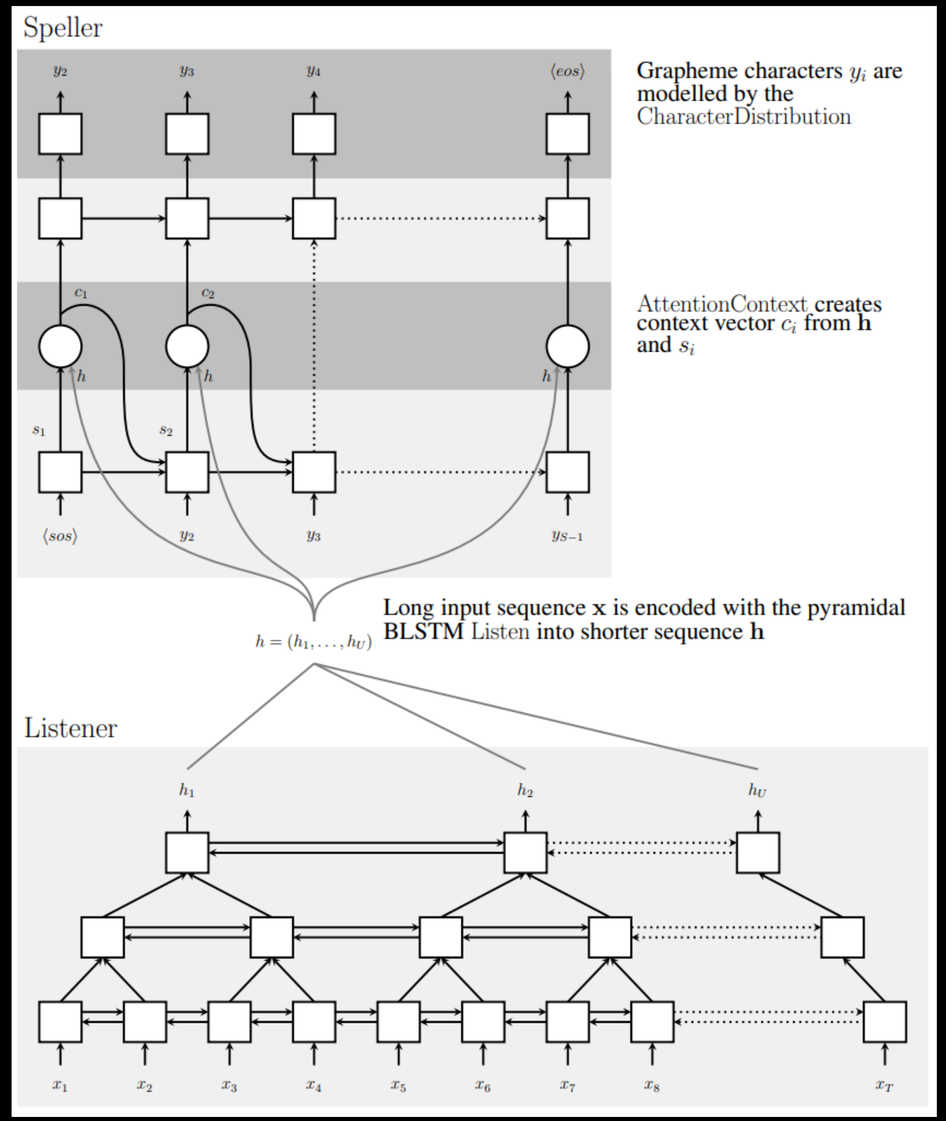
2.1 End-to-End ASR 신경망 모델

NAVER DEVIEW 2023

End-to-End ASR 모델의 장점

1. 단순한 구조로 인한 낮은 진입장벽
   - 전처리 X
   - 신경망 외 다른 알고리즘 X
   - 복잡한 학습 메커니즘 X

2. 사람의 개입없이, 오로지 데이터만으로 학습
   - 사람의 잘못된 bias로 인한 성능 저하가 없음

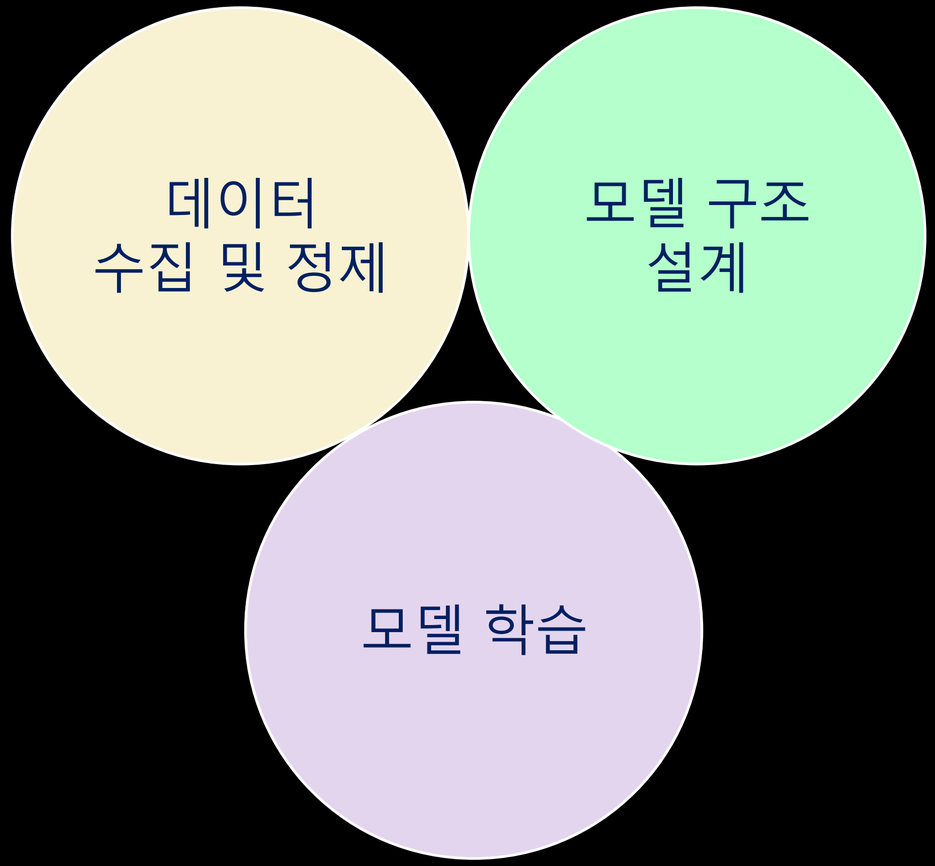


William Chan et al., Listen, attend and spell, ICASSP 2015

2.1 End-to-End ASR 신경망 모델

NAVER
DEVIEW
2023

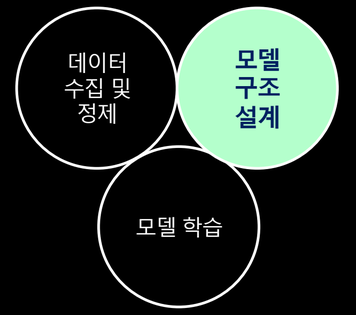
- audio-text pair
- 대량의 데이터
- 양질의 데이터



- Transformer
- LSTM
- …

- 지도학습
- Gradient Descent

2.1 End-to-End ASR 신경망 모델 구조



CTC

C L O O __ V A __

CTC Projection

Transformer Encoder

AED

C L O V A

Transformer Decoder

Transformer Encoder

RNN-Transducer

CL__OVA__

Joint

LSTM Predictor

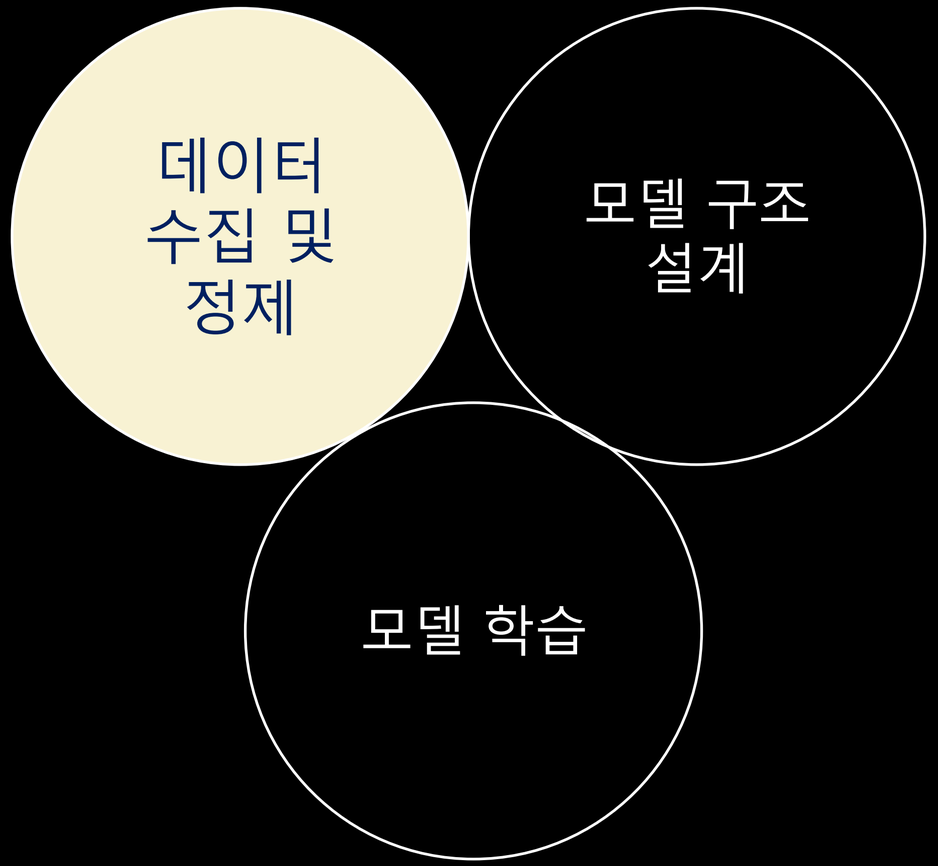
Transformer Encoder

2.1 End-to-End ASR 데이터

NAVER
DEVIEW
2023

End-to-End ASR모델을 위한 데이터

- Speech-text pair data
- 양질의 데이터 필요
  1) 실제 발화 환경과 동일 (잡음, 녹음 장비 등)
  2) 실제 발화 특징과 동일 (어투, 성별, 나이, 국적 등)
  3) 정확한 전사
- 대량의 데이터 필요
  - 음성을 말하고, 사람이 듣고 받아써야 함
  - 1시간을 전사하기 위해 10시간이 필요함!



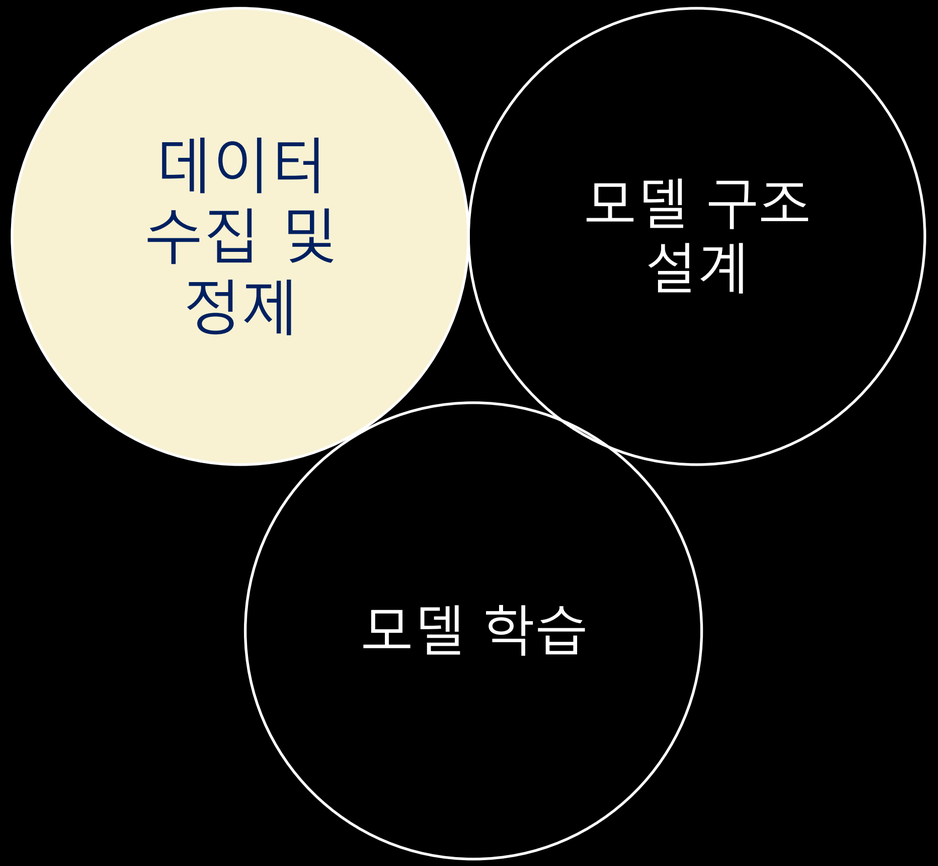
2.1 End-to-End ASR 데이터

NAVER
DEVIEW
2023

End-to-End ASR모델을 위한 데이터

- 양질의 + 대량의 + Speech-text pair data
  - 대본을 읽는 대신, 자유발화 음성을 녹음
  - 자유발화를 듣고 전사
  - 1시간을 전사하기 위해 22.5시간이 필요함!

음성 데이터 수집 (쉬움) → 음성 데이터 전사 (어려움)



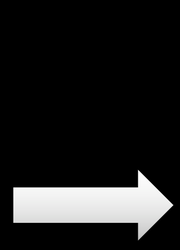
2.2 자기지도 학습

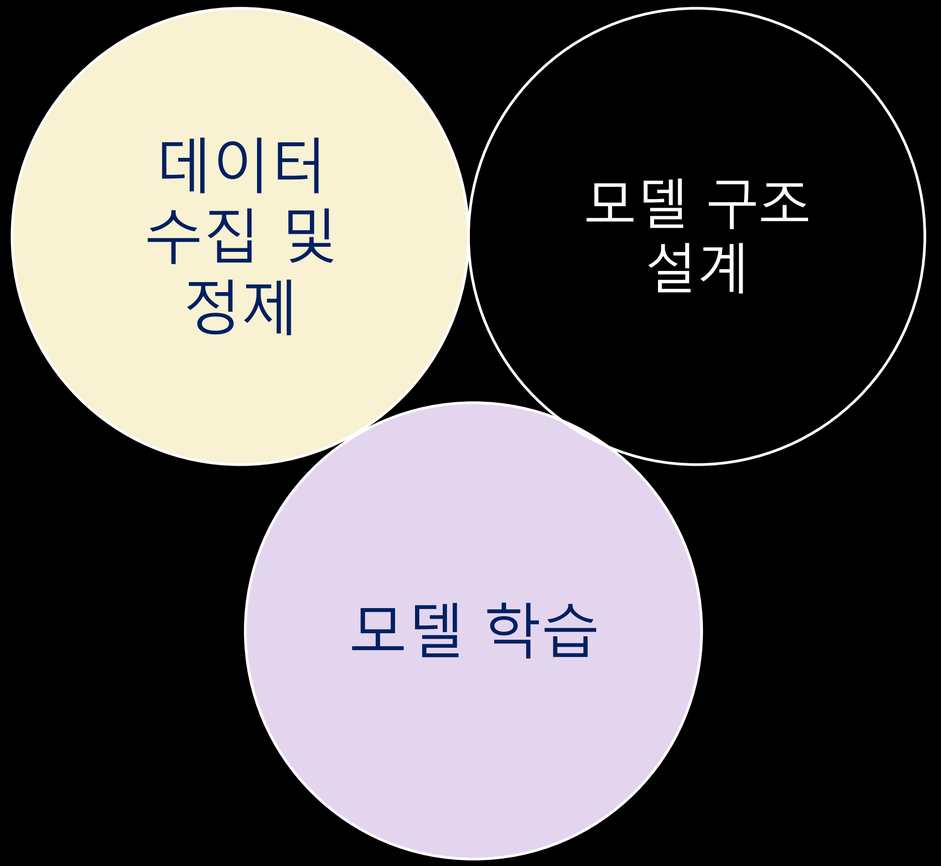
NAVER
DEVIEW
2023

End-to-End ASR모델 자기지도 학습 훈련

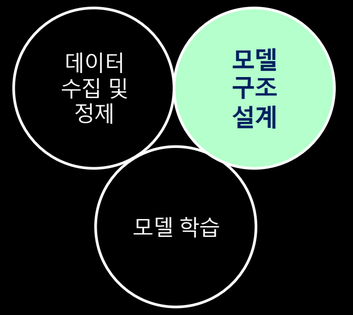
- Text가 없어도,
- Speech만으로 ASR모델 훈련이 가능하다고?
- 신경망 모델이 스스로 음성의 “특성” 을 배운다

음성 데이터
수집
(빠름)





2.2 자기지도 학습 모델 구조



CTC

C L O O __ V A __

CTC Projection

Transformer Encoder

AED

C L O V A

Transformer Decoder

Transformer Encoder

RNN-Transducer

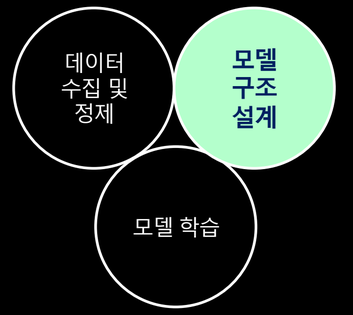
CL__OVA__

Joint

LSTM Predictor

Transformer Encoder

2.2 자기지도 학습 모델 구조



CTC

C L O O __ V A __

CTC Projection

Transformer Encoder

AED

C L O V A

Transformer Decoder

Transformer Encoder

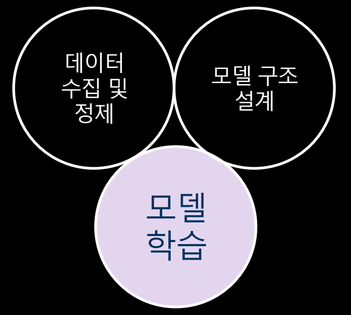
RNN-Transducer

CL__OVA___

Joint

LSTM Predictor

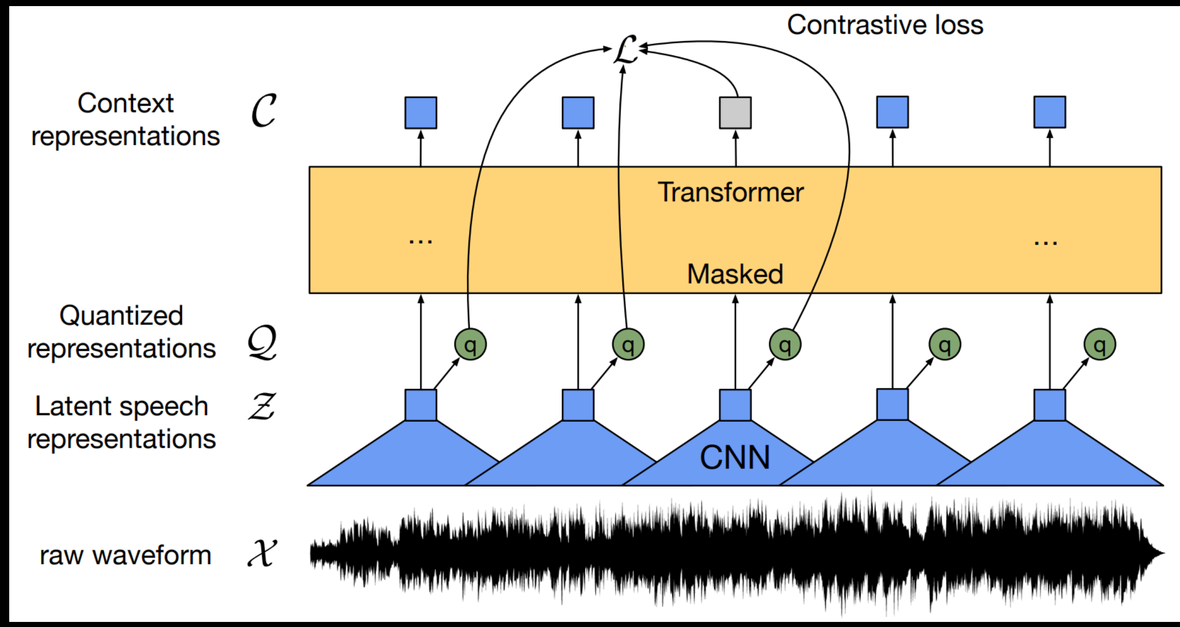
Transformer Encoder



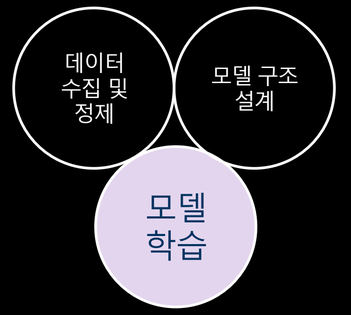
2.2 자기지도 학습

End-to-End ASR모델 자기지도 학습 훈련

- Transformer를 text없이 audio 데이터만으로 학습
- 하지만, 자기지도 학습? 어떻게? Label이 없는데?
  1) 일부 audio를 지우고
  2) 주위의 audio로 지운 audio를 예측하도록 훈련
  3) 데이터를 왕창 넣고 오래 기다리면 완성
- 마치 사람이 학습하는 방식과 흡사



A. Baevski et al. Wav2vec 2.0: A framework for self-supervised learning of speech representations, NeurIPS 2020



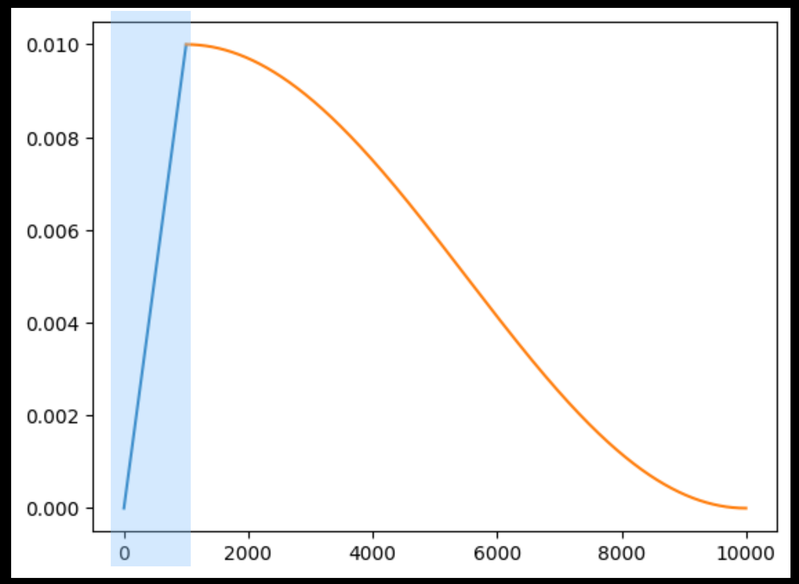
2.2 자기지도 학습

End-to-End ASR모델 자기지도 학습 훈련

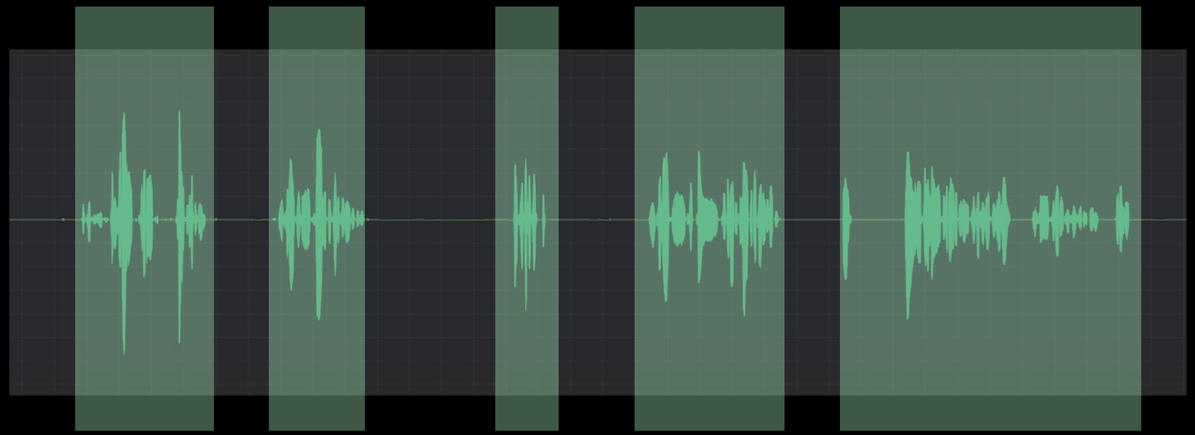
- Label이 없이 학습하기 때문에, 학습이 불안정함

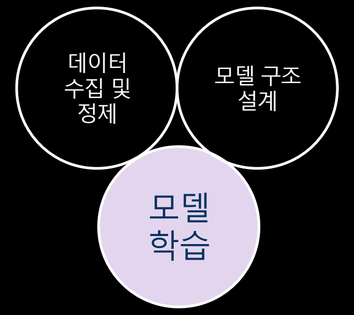
1) Warm-up Scheduler 사용

학습 초기에 learning rate 를 천천히 끌어올려 안정적으로 학습을 시작함



2) Voice Activity Detection을 사용하여 speech data만 선별하여 훈련에 사용

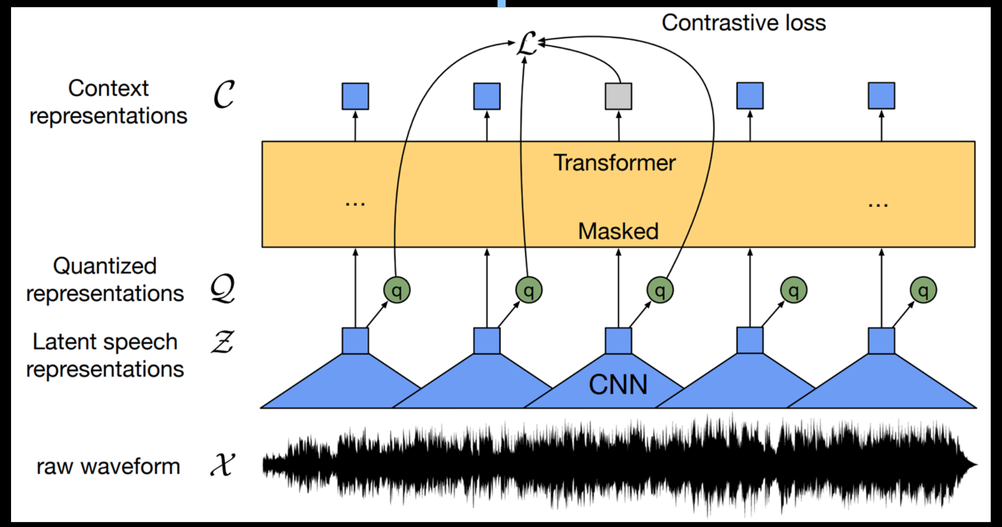


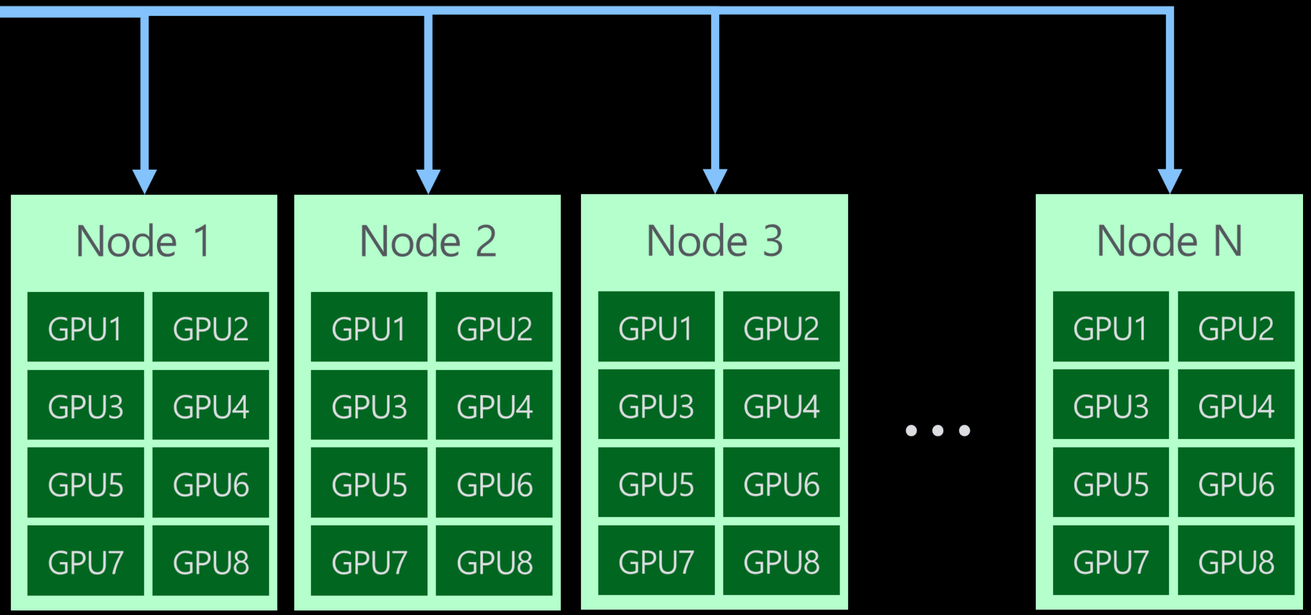


2.2 자기지도 학습

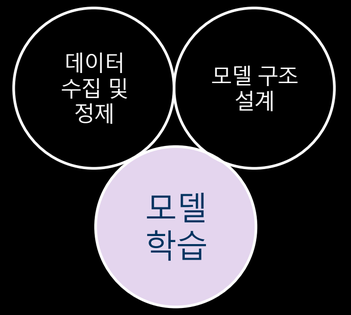
End-to-End ASR모델 자기지도 학습 훈련

- 제한된 시간안에 많은 데이터를 처리하고 큰 모델을 학습하기 위해
- multi-node multi-gpu 분산학습기법 사용





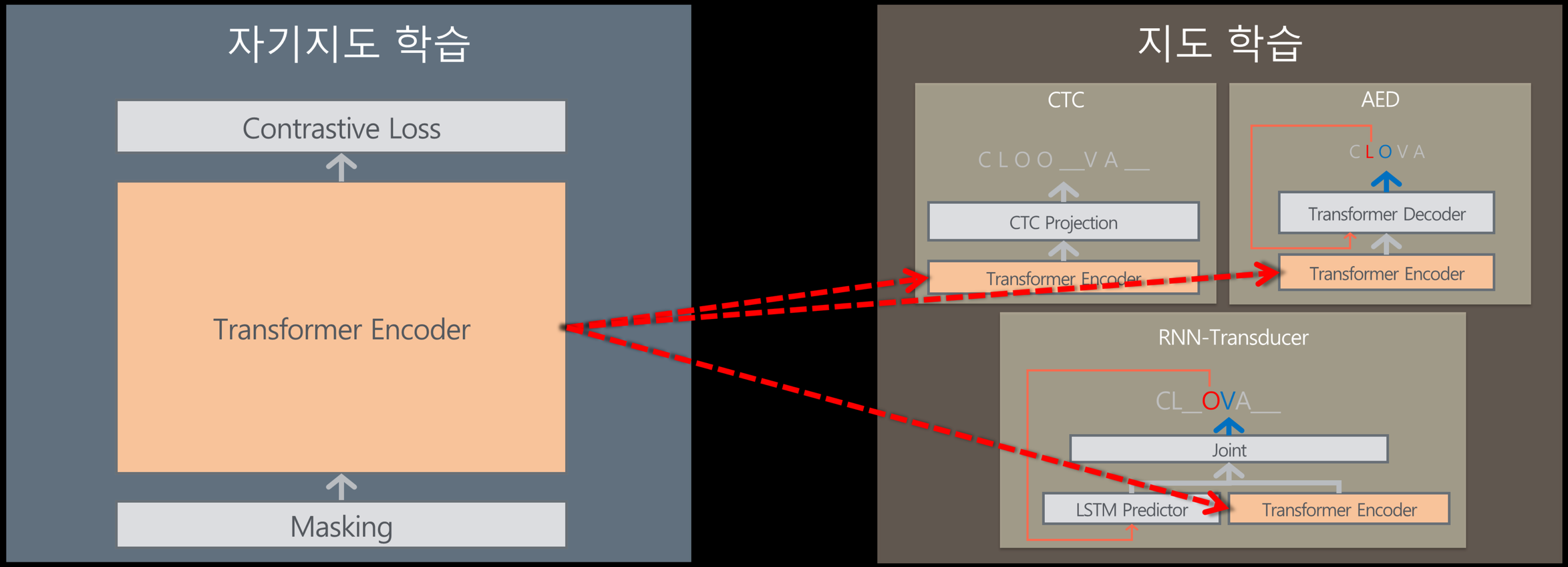
A. Baevski et al. Wav2vec 2.0: A framework for self-supervised learning of speech representations, NeurIPS 2020

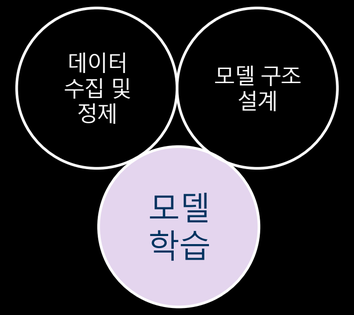


2.2 자기지도 학습

End-to-End ASR모델 자기지도 학습 훈련

- 자기지도 학습만으로는 transformer encoder만 훈련 가능
- 최종 ASR을 위해서는, speech-text pair 데이터를 사용한 지도학습이 필요함

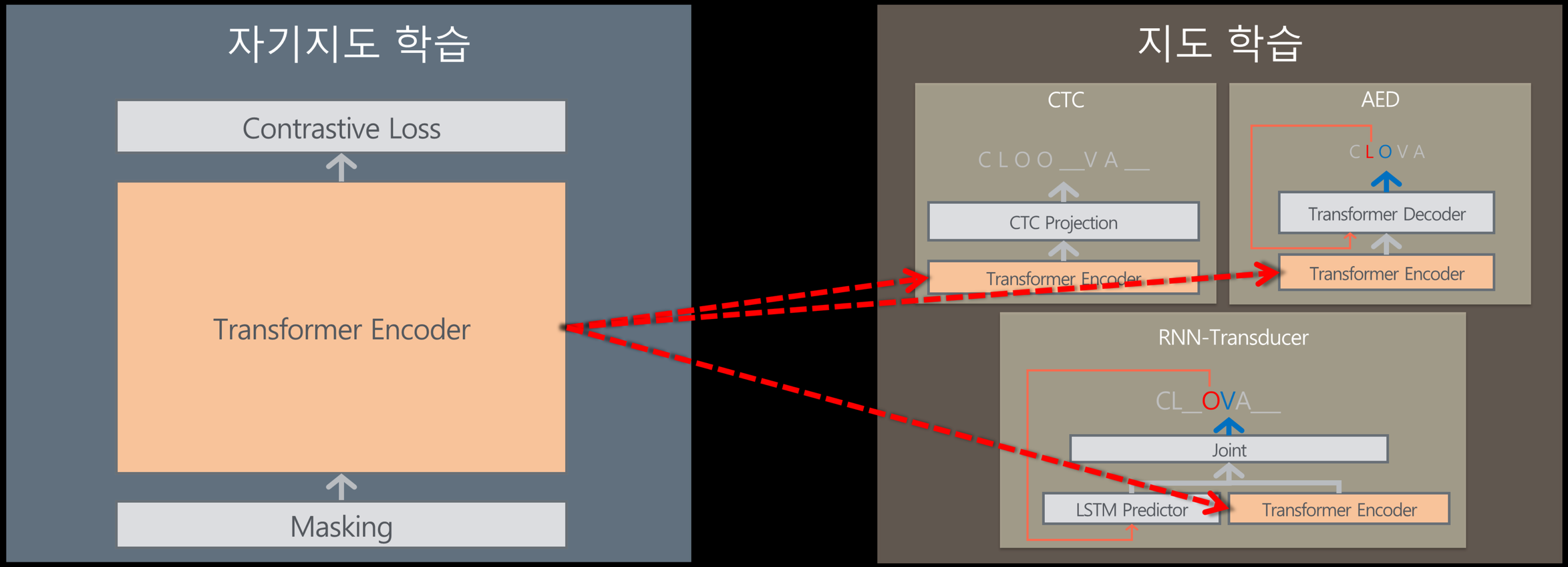




2.2 자기지도 학습

End-to-End ASR모델 자기지도 학습 훈련

- 자기지도 학습 이후라면, 아주 소량의 speech-text pair데이터 만으로도 좋은 성능의 ASR모델이 가능

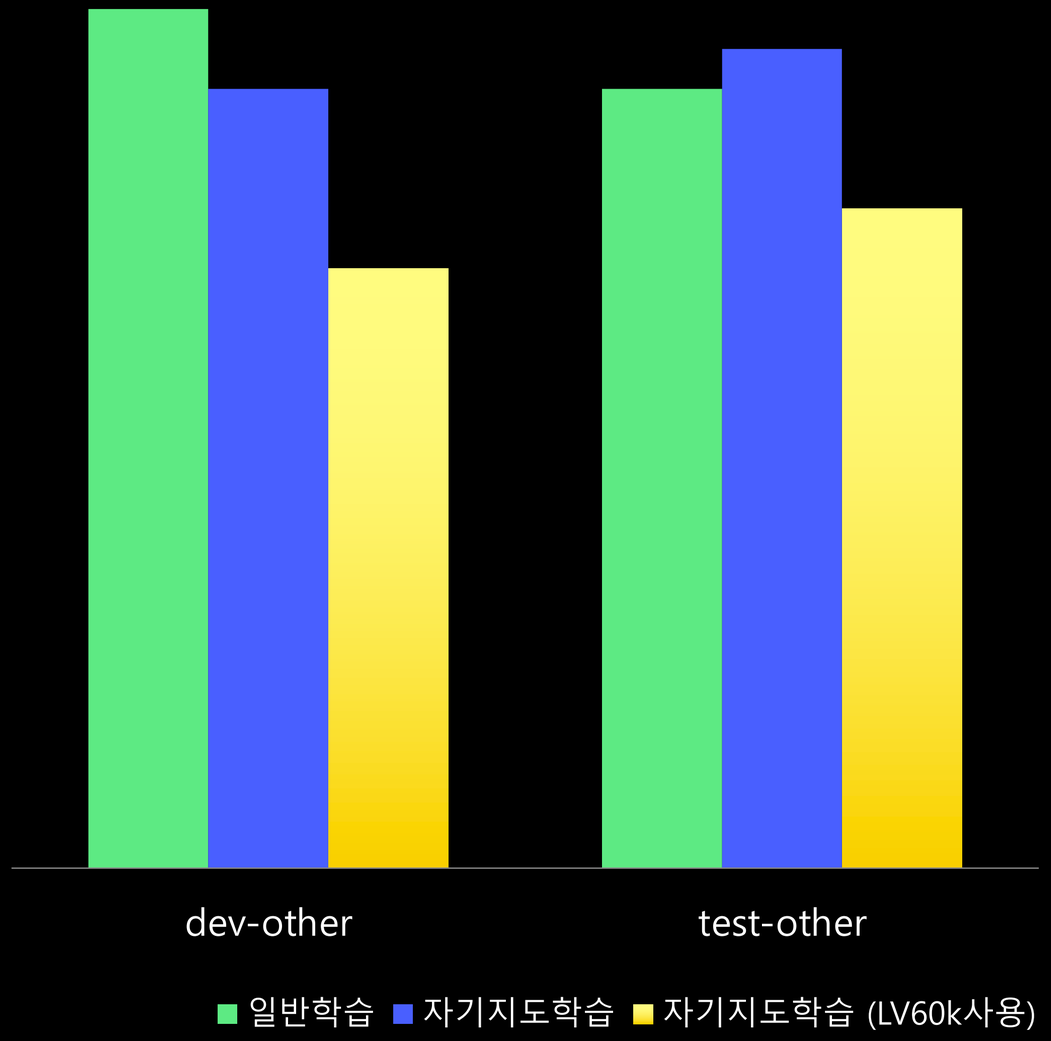


2.2 자기지도 학습

NAVER
DEVIEW
2023

자기 지도 학습의 성능

- 평가지표: WER (Word error rate)
- 훈련 데이터:
  - Librispeech 960h (전사 데이터)
  - LibriVox 60k (비전사 데이터)
- 평가 데이터: Librispeech dev-other/test-other



A. Baevski et al. Wav2vec 2.0: A framework for self-supervised learning of speech representations, NeurIPS 2020

3. 사용자의 마음을 이해하는 언어모델

3. 언어모델

NAVER
DEVIEW
2023

3. 언어모델

NAVER
DEVIEW
2023

내추럴/뇌출혈

내추럴 커피 한잔 나왔습니다~

뇌출혈로 인한 사망이 크게
증가했습니다.

앞뒤 문맥이나 배경지식이 없다면
같은 소리도 다르게 들린다!

3. 언어모델

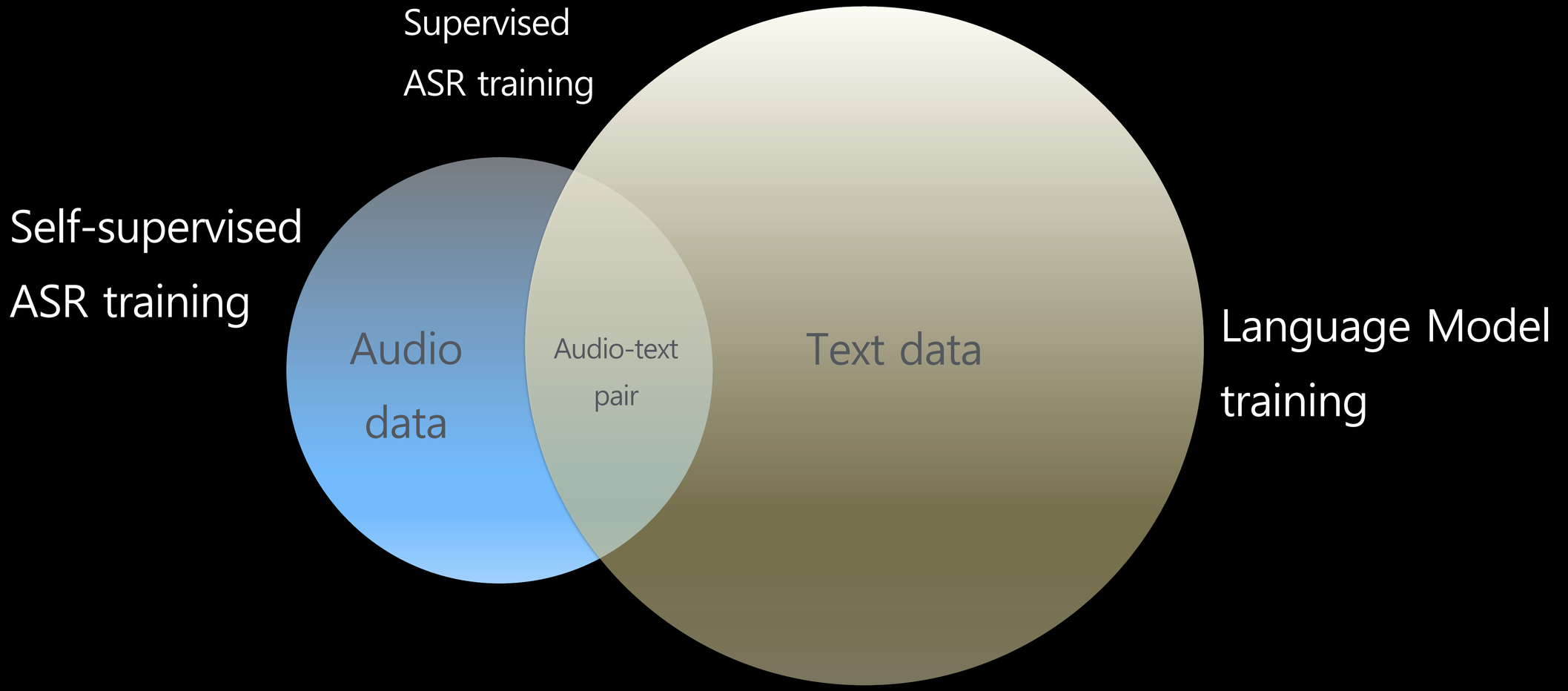
NAVER
DEVIEW
2023

- 언어의 문법, 단어의 빈도, 문장 구조 등을 이해하여 음성인식 결과를 보완하고, 음성에서 인식한 단어들이 문맥에 맞는지 검증
- 텍스트 데이터만으로도 학습 가능
  (텍스트 데이터는 음성 데이터에 비해 훨씬 많고 쉽게 구할 수 있음)

언어적으로 좀 더 의미가 통하게 만들어서 정확도를 향상!

3. 언어모델

NAVER
DEVIEW
2023



3.1 언어모델의 구조

NAVER
DEVIEW
2023

Text sequence에 대한 joint probability model

Autoregressive

- n-gram, GPT

Bidirectional

- ELMo, BERT

3.2 Beam Search with LM

NAVER
DEVIEW
2023

Shallow Fusion (First-pass rescoring)

- ASR model 확률과 LM의 확률을 합쳐서 Beam Search
- Autoregressive 구조의 모델만 사용할 수 있음

$$P_{LM}(Y_1, \cdots, Y_T) = \sum_{t=1}^{T} P_{LM}(Y_t | Y_{1:t-1})$$

- Shallow fusion

$$Score_{ASR}(Y|X) = \log P_{AM}(Y|X) + \lambda \log P_{LM}(Y)$$

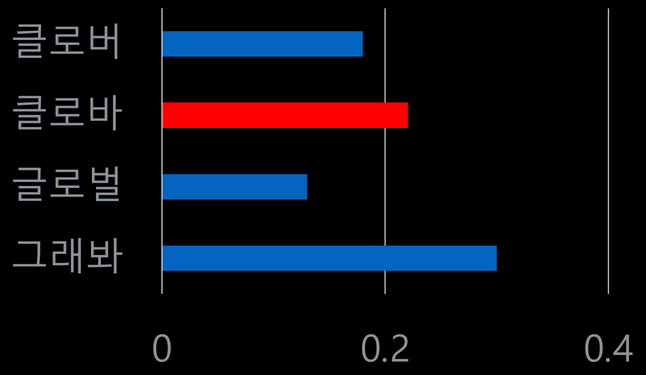
- Beam search step

$$Score_{ASR}(Y_{1:t-1}|X) \mathrel{+}= \log P_{AM}(Y_t|X, Y_{1:t-1}) + \lambda \log P_{LM}(Y_t|Y_{1:t-1})$$

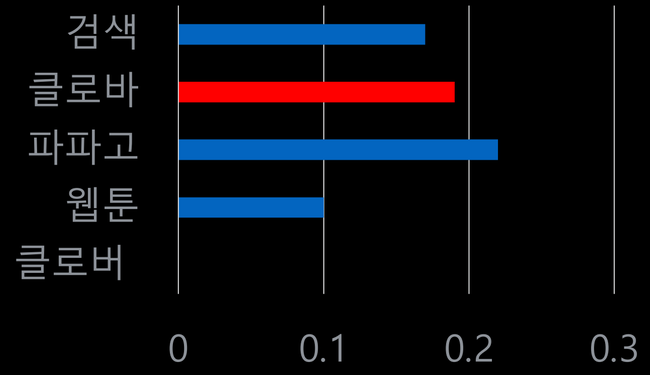
3.2 Beam Search with LM

NAVER
DEVIEW
2023

사용자의 마음을 이해하는 네이버 OOO



$P_{AM}(Y_t|X, Y_{1:t-1})$



$P_{LM}(Y_t|Y_{1:t-1})$

3.2 Beam Search with LM

NAVER
DEVIEW
2023

- Multi-level Beam Search

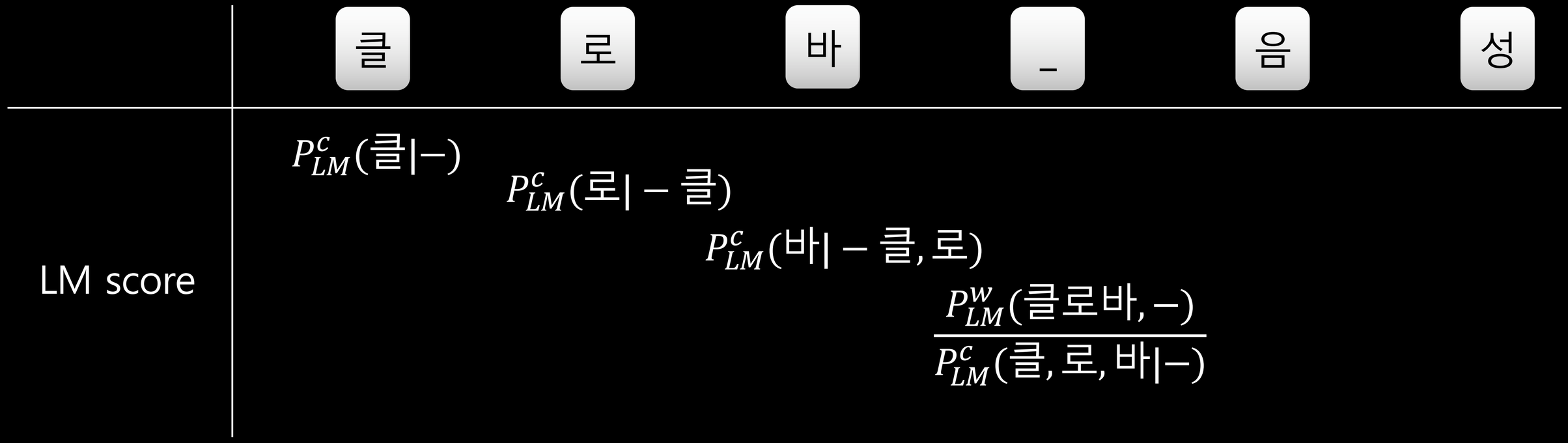
AM의 vocab unit (character) 과 LM의 vocab unit (word) 이 달라서 shallow fusion이 어려운 경우가 있음.

→ character 단위의 LM으로 beam search를 하다가 단어가 완성되는 경우 word LM 점수로 치환해서 사용.

3.2 Beam Search with LM

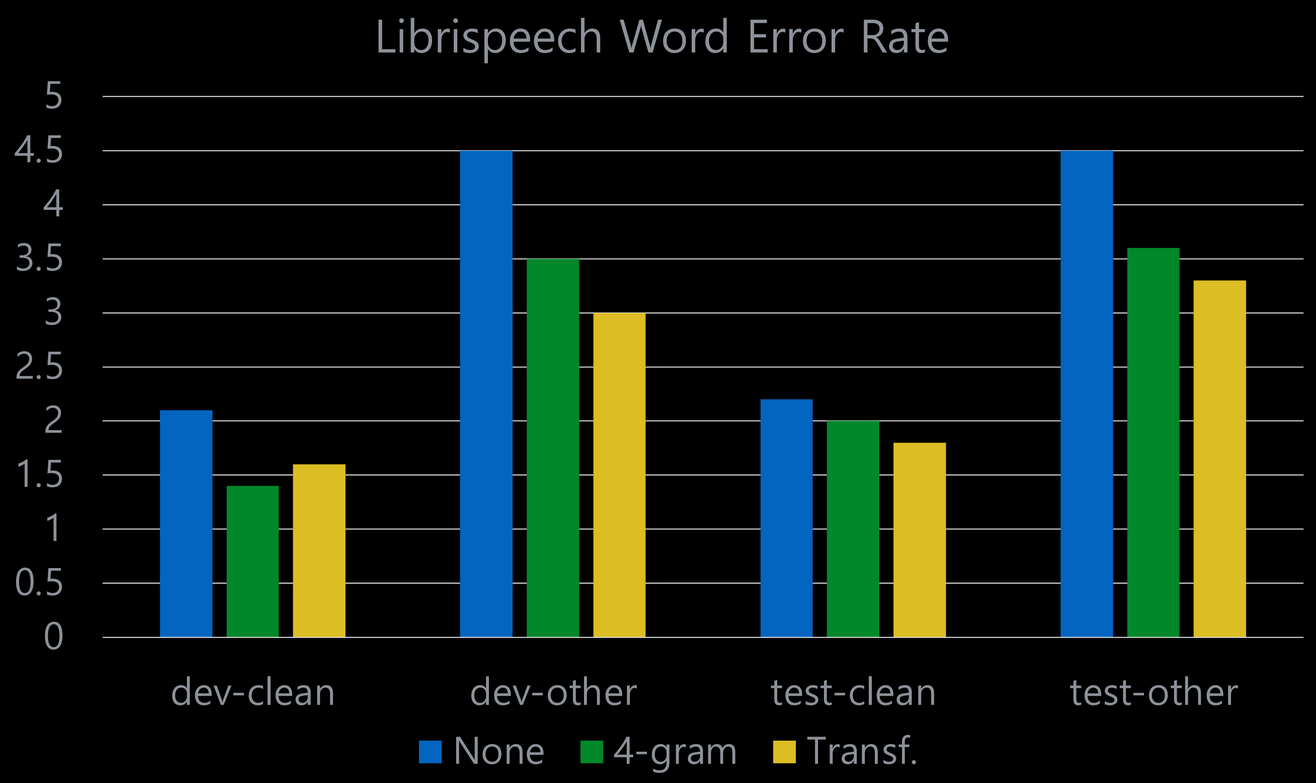
NAVER DEVIEW 2023

- Multi-level Beam Search



3.2 Beam Search with LM

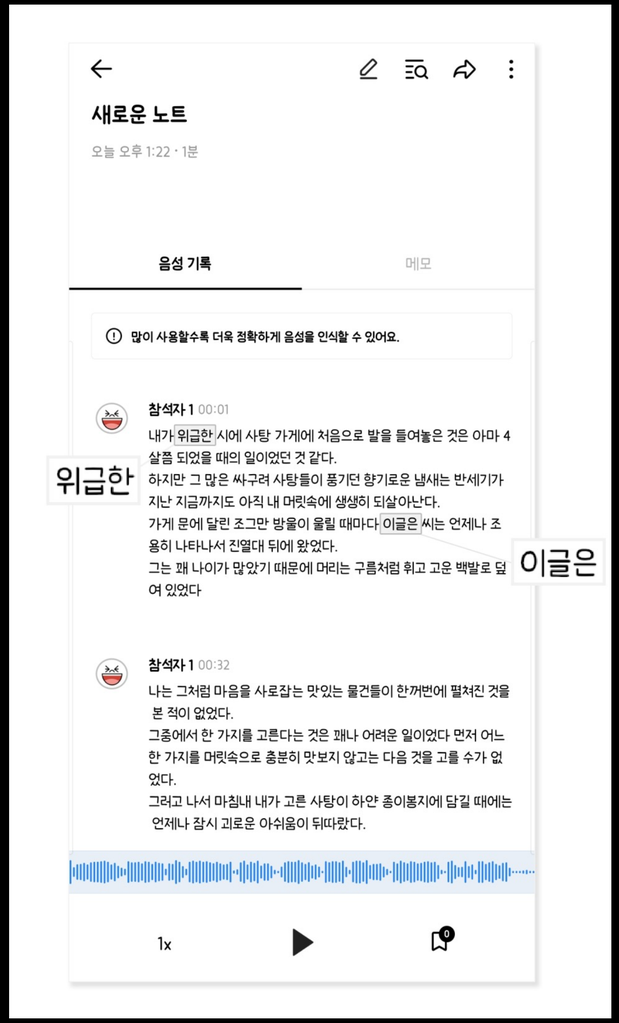
NAVER
DEVIEW
2023



4. NAVER E2E 인식기만의 특별한 기능

4.1 Keyword Boosting

NAVER
DEVIEW
2023



인식률이 기대보다 훨씬 좋아서 깜짝 놀랐어요.
고유 명사인 ***위그든 씨*** 의 이름은 인식하지 못했고,
희고 고운 -> 휘고 고운 으로 받아적은 것 외에는
전부 완벽하게 텍스트로 변환되었어요.

[출처] 네이버 클로바노트, 음성 텍스트 변환 성능은 어떨까?| **작성자** 나루

사람이름, 전문용어를 미리 학습하지 않고
그때그때 받아서 잘 인식되게 할 수 없을까?

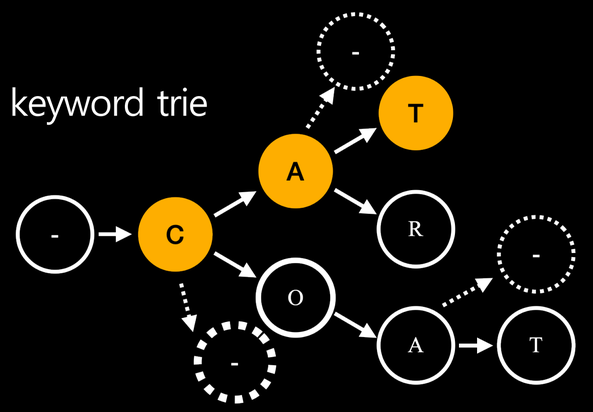
4.1 Keyword Boosting

NAVER
DEVIEW
2023

어려운 단어 인식을 위한 Keyword Boosting

- 사전에 주요 단어들 리스트를 받아서 Beam search 과정에서 해당 단어들이 포함된 path에 대해서 점수를 올려주는 방식
- 키워드들은 미리 token 단위의 trie(prefix tree)로 만들어둬서 ASR beam search와 함께 사용
- ICASSP 2022에 모델 발표

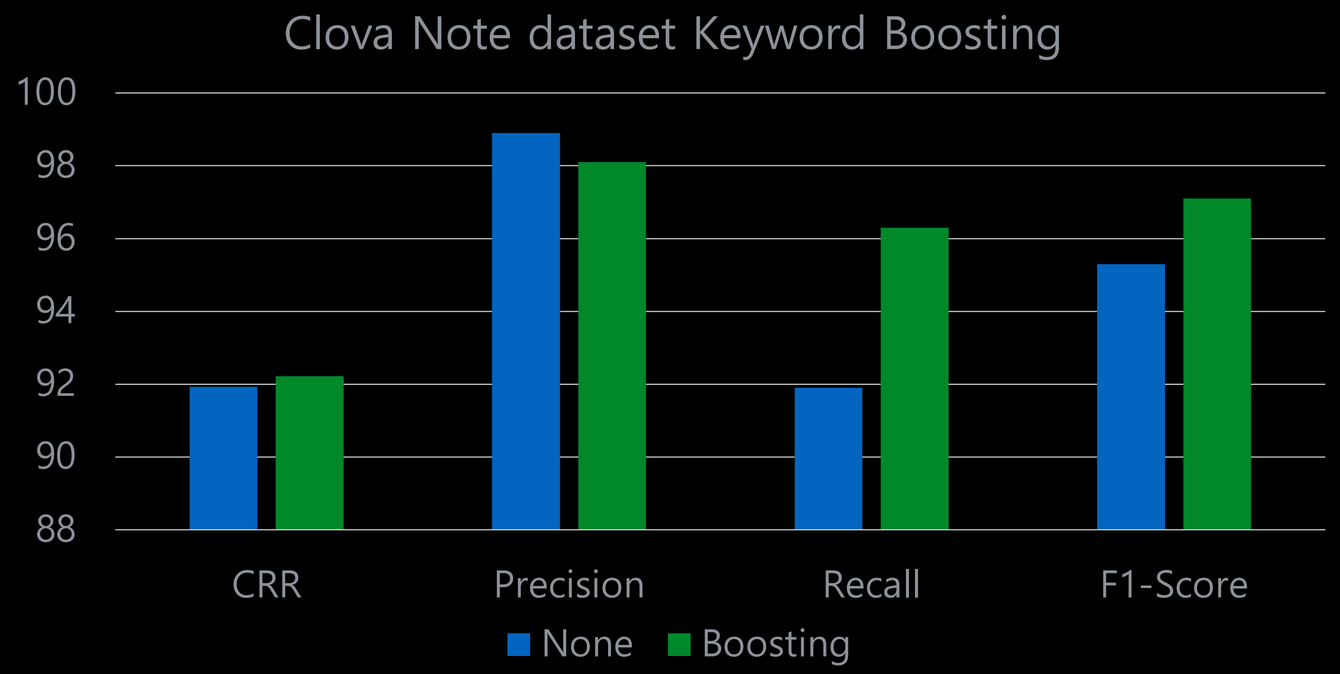
CAT? CAN? CAP?



CAT!

4.1 Keyword Boosting

NAVER
DEVIEW
2023



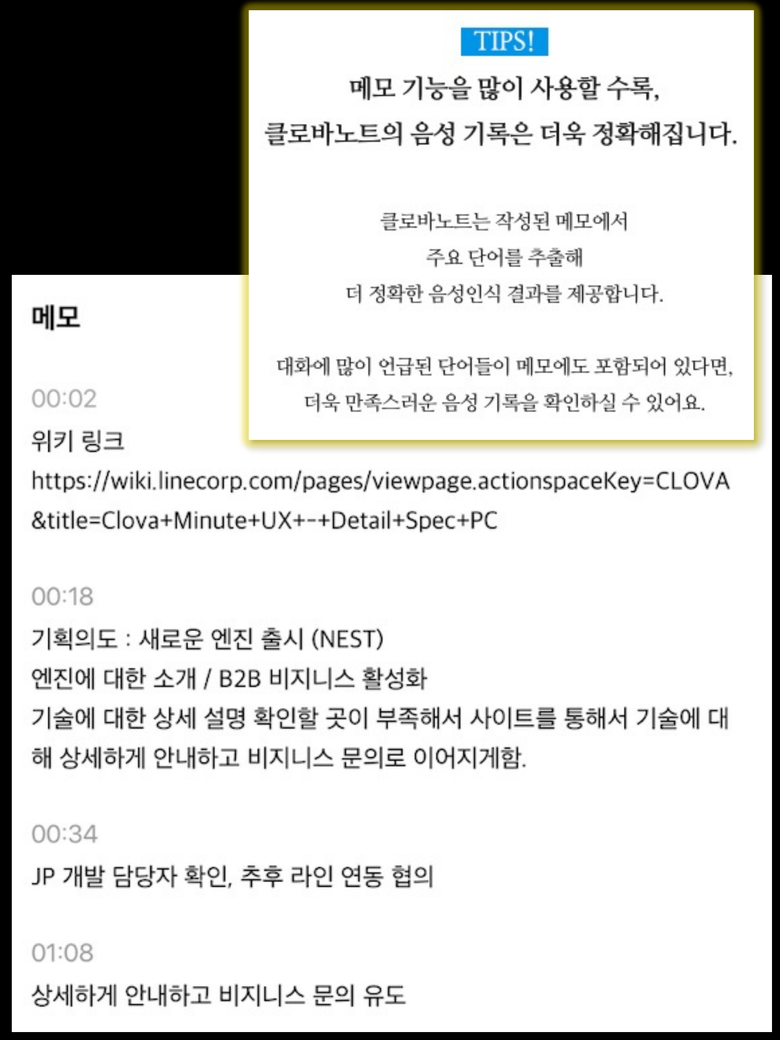
정확도는 큰 차이 없지만 키워드 Recall이 유의미하게 향상!

약간의 Precision loss 발생

4.1 Keyword Boosting

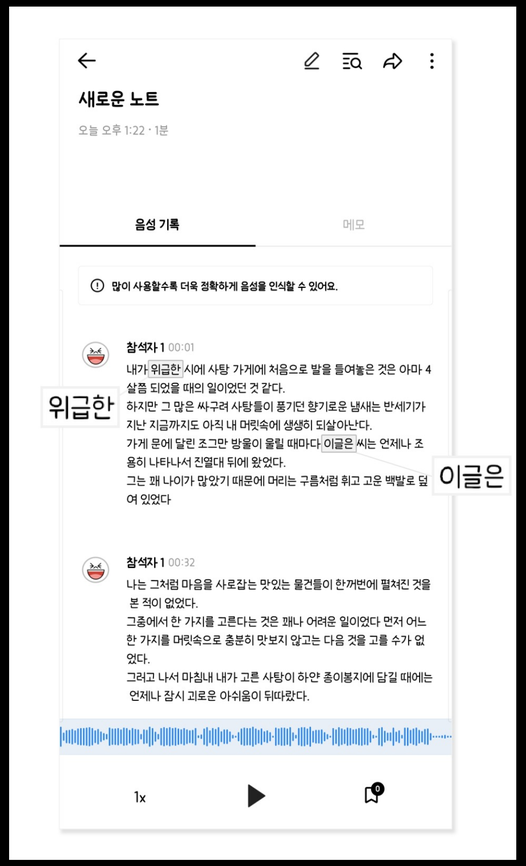
NAVER
DEVIEW
2023

클로바노트에 키워드 부스팅 적용!

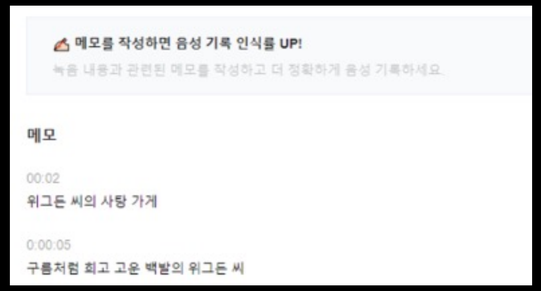


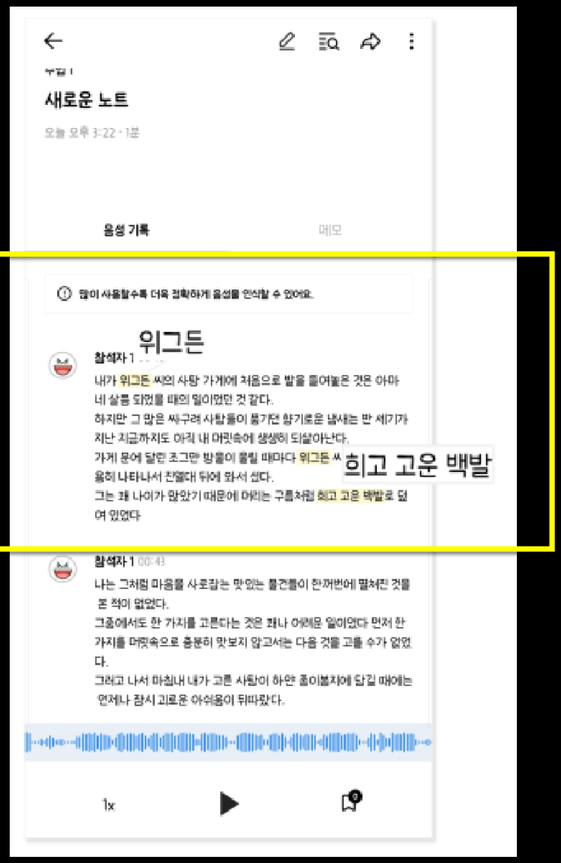
4.1 Keyword Boosting

NAVER
DEVIEW
2023



+





메모

위그든 씨의 사탕 가게

구름처럼 희고 고운 백발의 위그든 씨

4.2 Post-Processing

NAVER
DEVIEW
2023

가독성 향상을 위한 후처리 기술들

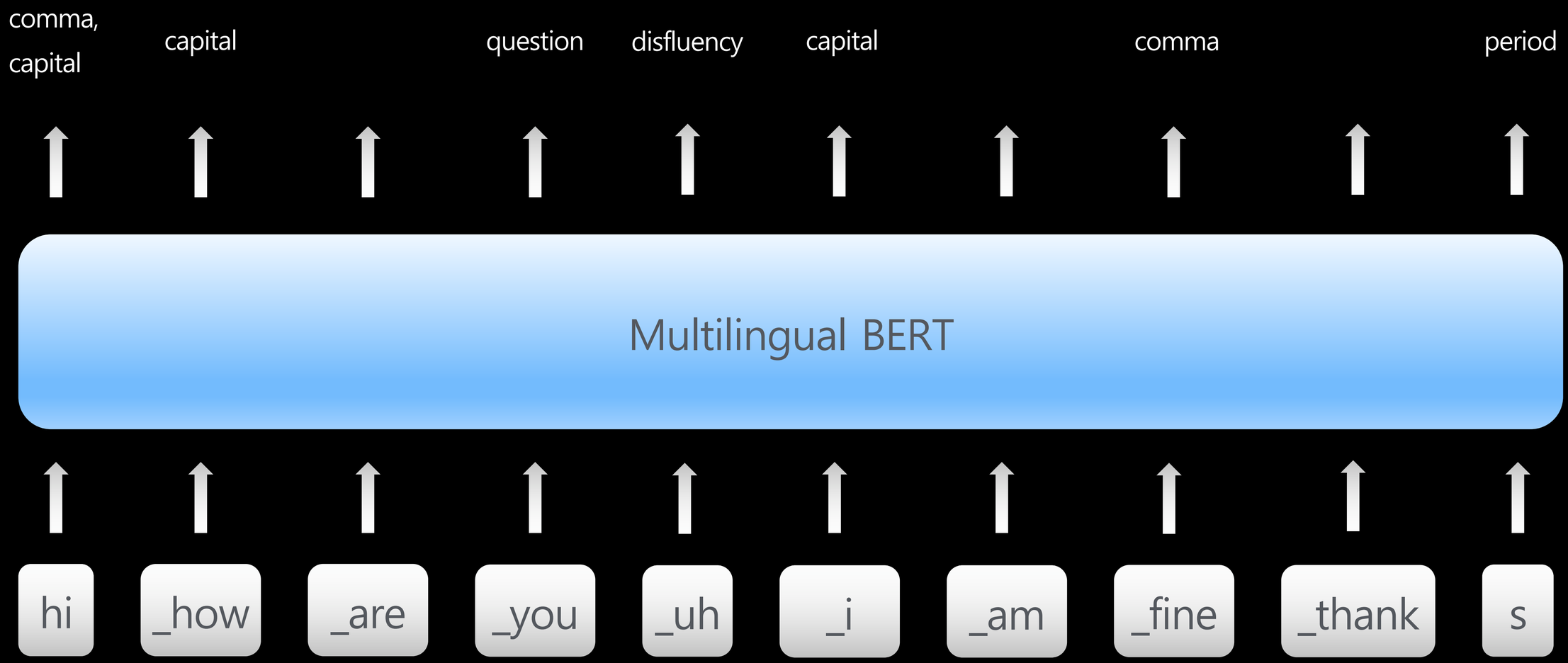
- Punctuation
- Truecasing (Capitalization)
- Disfluency Removal

argentina defeat france on penalties to win the 2022 world cup this game was everything today with the entire world watching and two of the sport's biggest stars and club teammates at paris saint germain squaring off head to head for the title, these teams delivered arguably the best world cup final of all time and one that ended fittingly with the game's greatest player lionel messi finally winning a world cup

Argentina defeat France on penalties to win the 2022 World Cup. This game was everything today. With the entire world watching and two of the sport's biggest stars and club teammates at Paris Saint-Germain squaring off head-to-head for the title, these teams delivered arguably the best World Cup final of all time, and one that ended fittingly with the game's greatest player, Lionel Messi, finally winning a World Cup.

4.2 Post-Processing

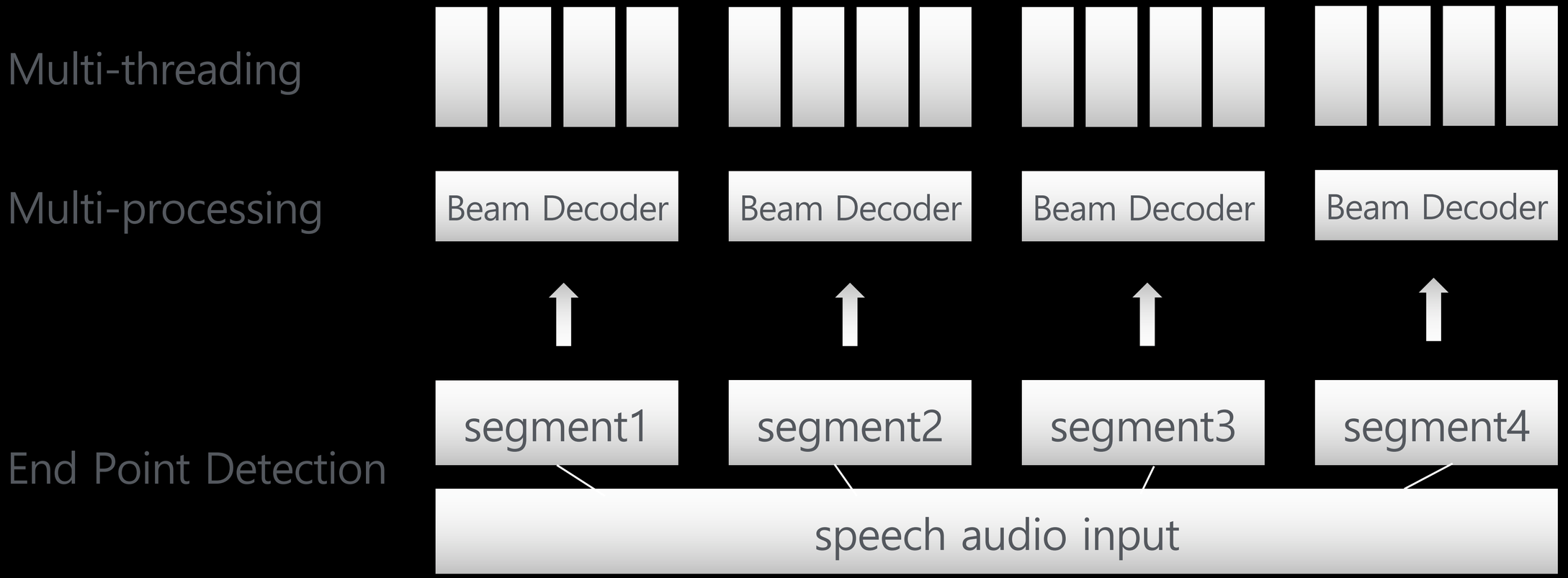
NAVER DEVIEW 2023



Hi, How are you? I am fine, thanks.

4.3 Parallel Decoding

NAVER
DEVIEW
2023

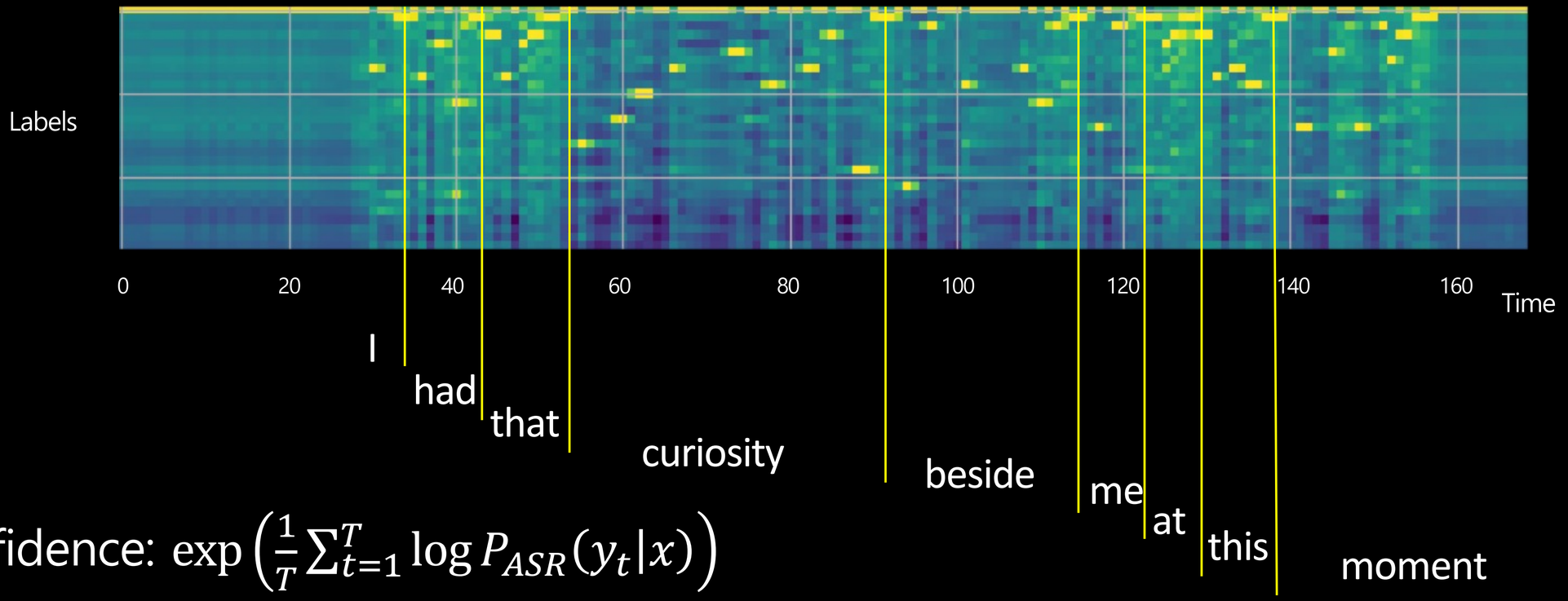


4.4 Time Alignments / Confidence

NAVER DEVIEW 2023

Time Alignment: CTC output에서 단어별 시간 정보를 찾아서 사용자가 자막 등에 활용

Confidence: ASR 확률을 통해 인식 결과에 대한 신뢰도 제공 및 노이즈 여부 판별



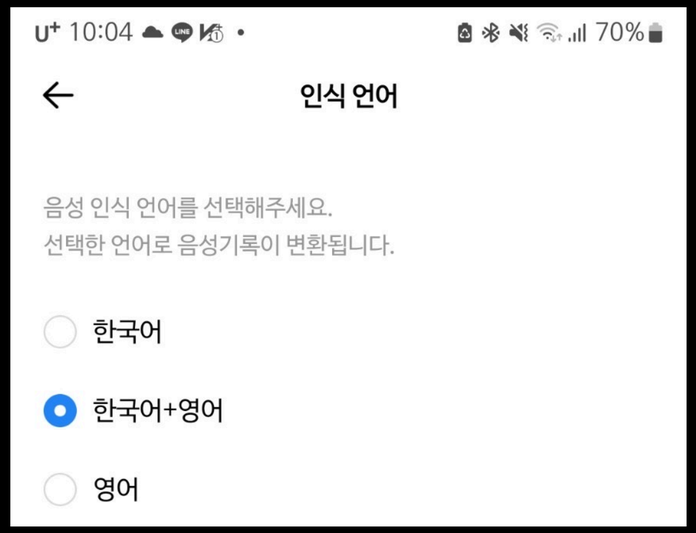
* Confidence: $\exp\left(\frac{1}{T}\sum_{t=1}^{T}\log P_{ASR}(y_t|x)\right)$

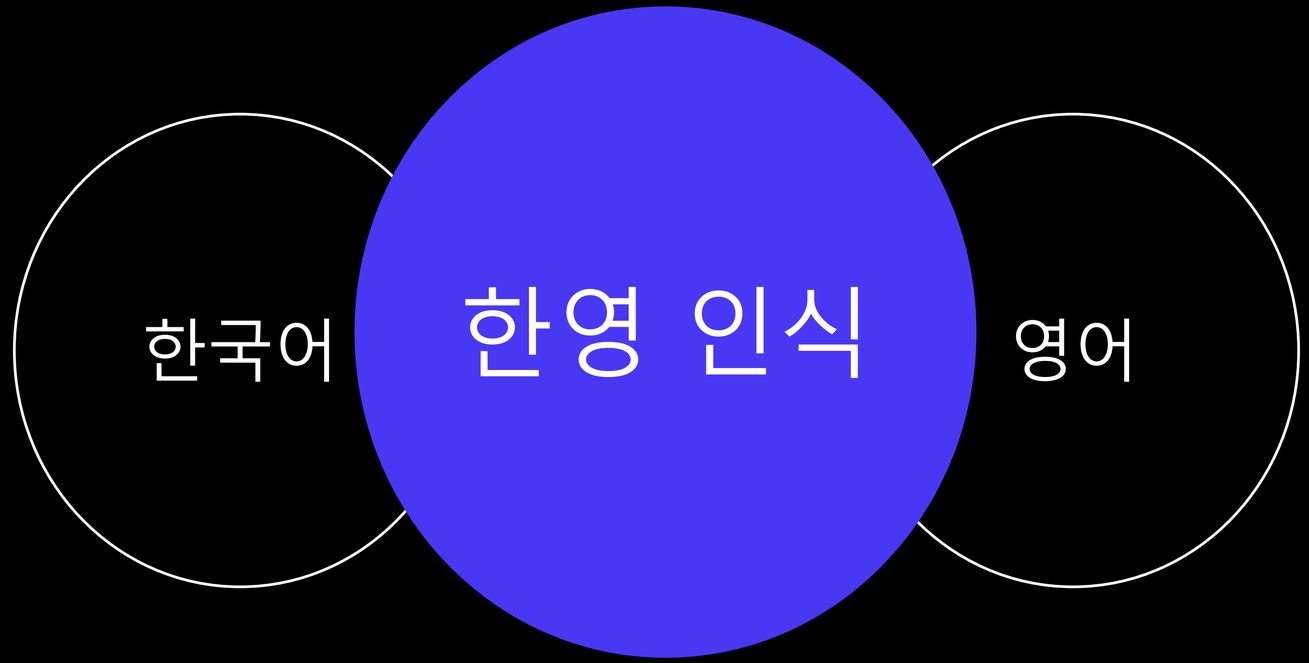
4.5 다국어 동시 인식

NAVER
DEVIEW
2023

한영 동시 인식 모델

- 한국어/영어 두가지 언어를 동시에 인식
  1) 한/영 언어 전환 불필요 → 편의성 향상
  2) 한국어/영어 동시 대화 인식 가능





4.6 서비스 맞춤형 음성 인식

NAVER
DEVIEW
2023

전문성이 필요한 도메인의 인식을 위한, 도메인 전용 음성인식

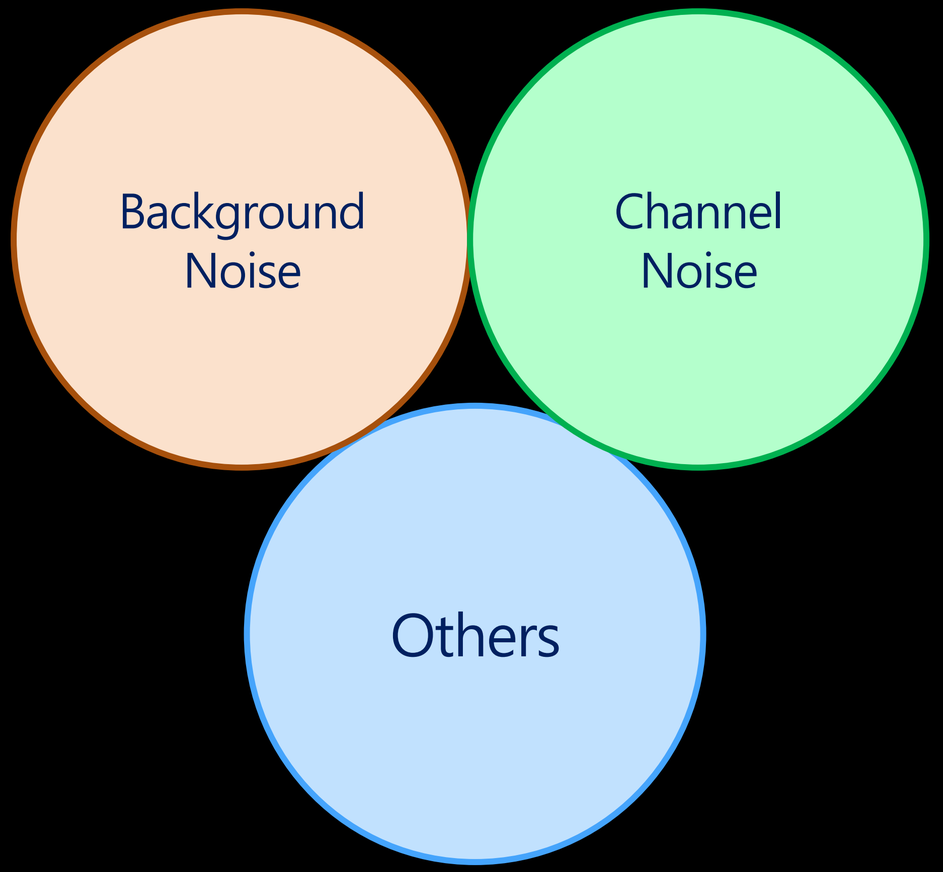
- VoiceEMR: 의료 도메인 전용 인식기
  - 병명, 수술명 등 의료도메인 전문용어를 잘 인식하도록 훈련
- 쇼핑 라이브 서비스 전용 인식기
  - 숫자, 제품명, 브랜드명 등을 잘 인식하도록 훈련
- 미래에셋대우 전화상담 전용 인식기
  - 숫자, 종목명, 주식 트레이드 관련 용어 등을 잘 인식하도록 훈련

4.7 Data Augmentation

NAVER
DEVIEW
2023

실제 소음 환경 시뮬레이션

- 일반 배경잡음
- 채널잡음
  - 전화망, 마이크 특정 등
- 기타 augmentation
  - speed perturbation,
  - spec augment
  - …



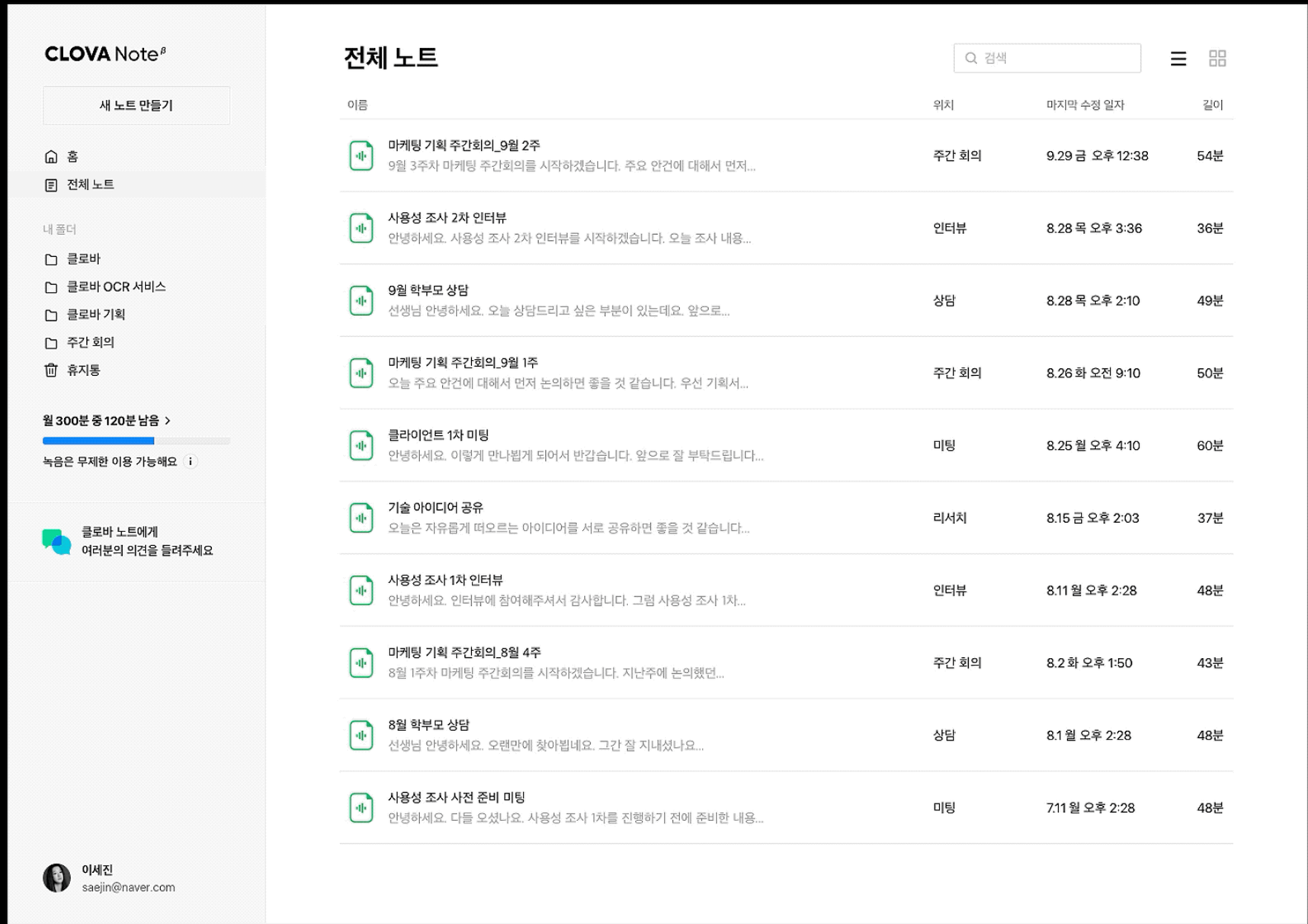
5. 서비스 적용 사례

5.1 클로바 노트

NAVER
DEVIEW
2023

음성 회의록 서비스

- 음성 인식
- 화자 인식
- 주요 키워드
- 회의록 요약



5.1 클로바 노트

NAVER
DEVIEW
2023

- Apple App Store 오늘의 앱 선정 (2021.04)
- Google Play 2021 Best of Awards
  ‘올해의 인기 앱 & 올해를 빛낸 일상생활 앱’ 2관왕 수상 (2021.12)
- 22년 11월 **300만 다운로드** 달성

5.2 클로바 케어콜

NAVER
DEVIEW
2023

세상을 이롭게 하는 AI

- 독거노인/1인가구 돌봄 전화 서비스

11월 부산 해운대구에서 베타 서비스를 시작했으며, 이후 대구, 인천, 서울로 도입을 이어가며 현재 전국 20개 지자체에서 시범 운영 중이다.

클로바 케어콜과의 상호작용이 독거 어르신의 외로움과 사회적 고립을 완화시키는 효과도 확인됐다. 부산 해운대구에서 클로바 케어콜 베타 서비스에 참여한 어르신들을 대상으로 설문 조사를 시행한 결과, 약 90%의 어르신 사용자는 서비스 이용 후 위로를 느꼈다고 답했다. 또한 약 95%의 응답자는 계속 서비스를 이용할 의향이 있다고 답해 높은 서비스 만족도를 보였다. 네이버는 이러한 결과를 바탕으로 다음 달 열리는 노인의학 분야 최고 권위 국제 학

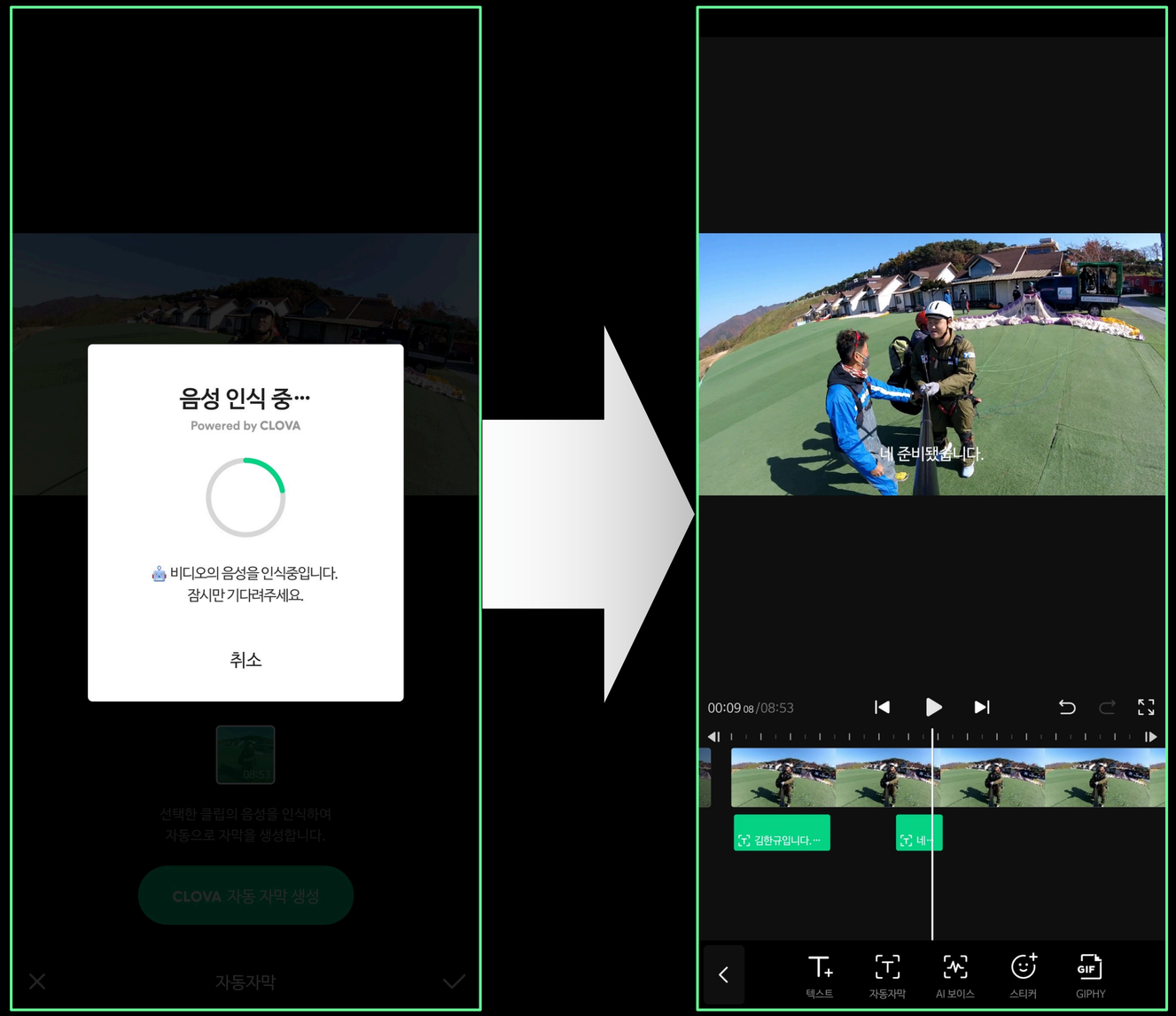
5.3 동영상 자동자막 서비스

NAVER
DEVIEW
2023

- VITA
  - 동영상 자동 자막 생성 서비스
  - 많은 크리에이터들이 사용하고 있음

- Weverse (구 VLive)
  - 23년 1분기 안에 서비스 예정

- 쇼핑라이브 자동자막



5.3 동영상 자동자막 서비스

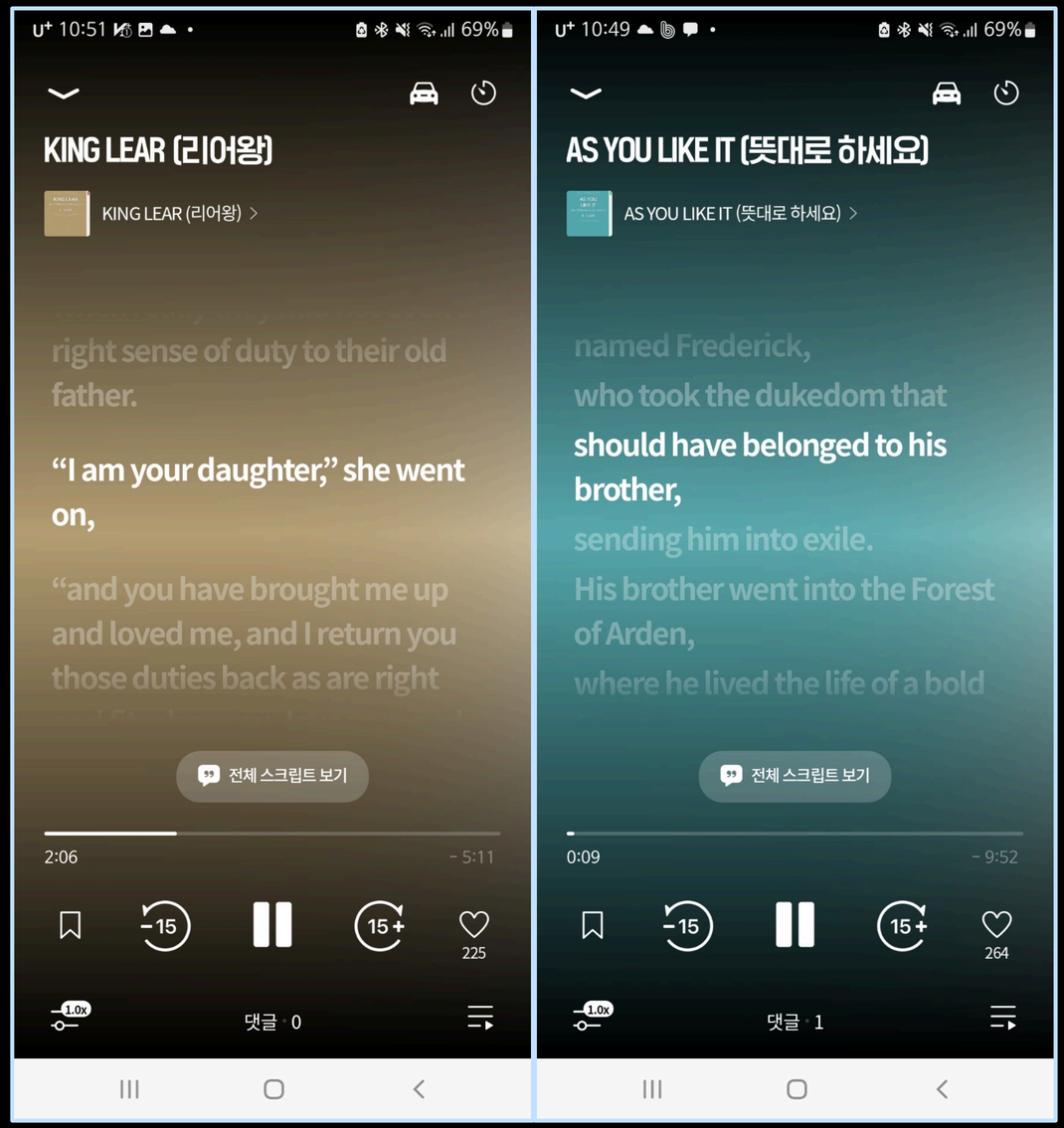
NAVER
DEVIEW
2023

네이버 뉴스

- 음성인식을 사용한 자동 자막 생성

오디오 클립

- 음성인식결과와 원본 도서 내용을 비교하여
- 자동 alignment 수행



5.4 B2B 서비스

NAVER
DEVIEW
2023

미래에셋대우 고객상담 STT서비스

- 콜센터 고객상담 통화에 대해 음성인식 딕테이션 진행
- 효율적인 고객 상담 이력 관리 → 상담 서비스 품질 향상

순천향대학교 중앙의료원 Voice EMR 서비스

- 의사/간호사의 목소리를 딕테이션 하여 진료 차트로 변환
- 의무기록 업무 효율화 → 의료 서비스 품질 향상

6.1 앞으로 고민해야할 문제들

NAVER
DEVIEW
2023

- 더 빠르고 효율적인 디코더 / 서빙 구조
- 어려운 도메인들에 대한 음성인식 모델 학습
- 음성인식 오류를 자동으로 고쳐주는 교정 모델 개발
- 개인화된 인식 모델
- 데이터 수집 및 전사의 자동화
- 다국어 동시 지원
- 효율적인 모델 학습