DEVIEW 2020

네이버지도를 위한 음성인식 개발기

오명우 NAVER

CONTENTS

본 세션은 이런 분들께 적합합니다.

- 음성인식 엔진을 개발중인 분들
- 음성인식 서비스를 운용하거나, 운용할 예정이신 분들
- 음성인식 관련 서비스를 준비하시는 분들
- 음성 데이터를 분석하시거나, 음성 기반 머신러닝을 연구개발하시는 분들

본 세션 청강을 위해 필요한 기본 지식입니다.

- Machine learning / Pattern classification
- Digital signal processing / Audio processing
- Probability / Statistics

CONTENTS

1. 네이버 음성인식 서비스

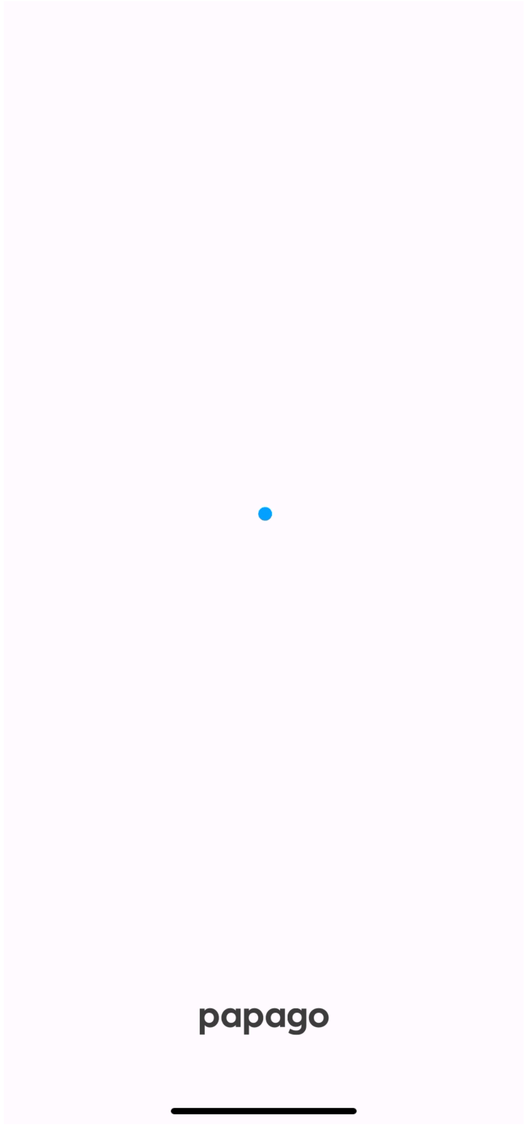
NAVER

네이버앱

2011.01~현재
국내기업 최초의
음성검색 서비스 출시

파파고

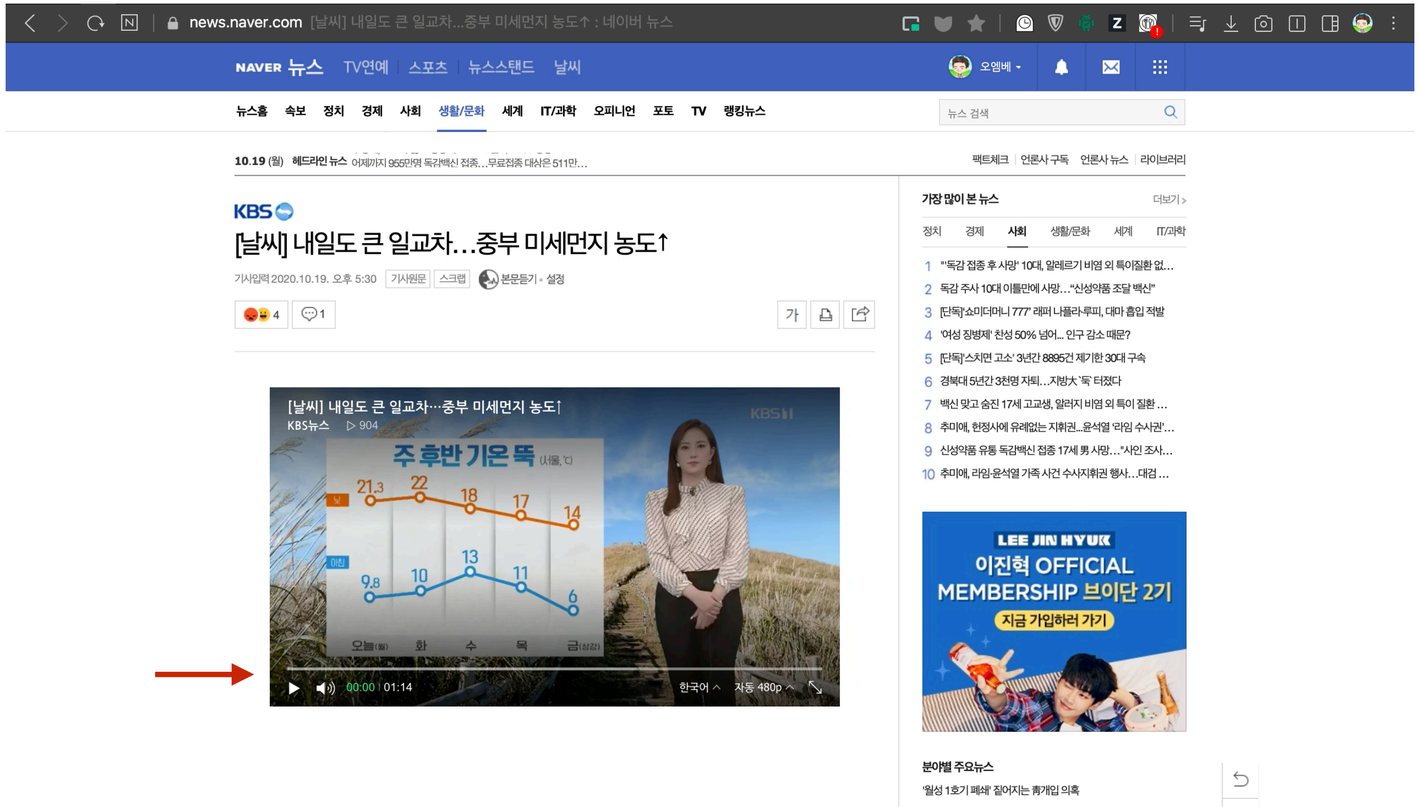
2016.08~현재
자동 통번역 앱
6개국 음성인식기 자체탑재



Clova

2017.07~현재

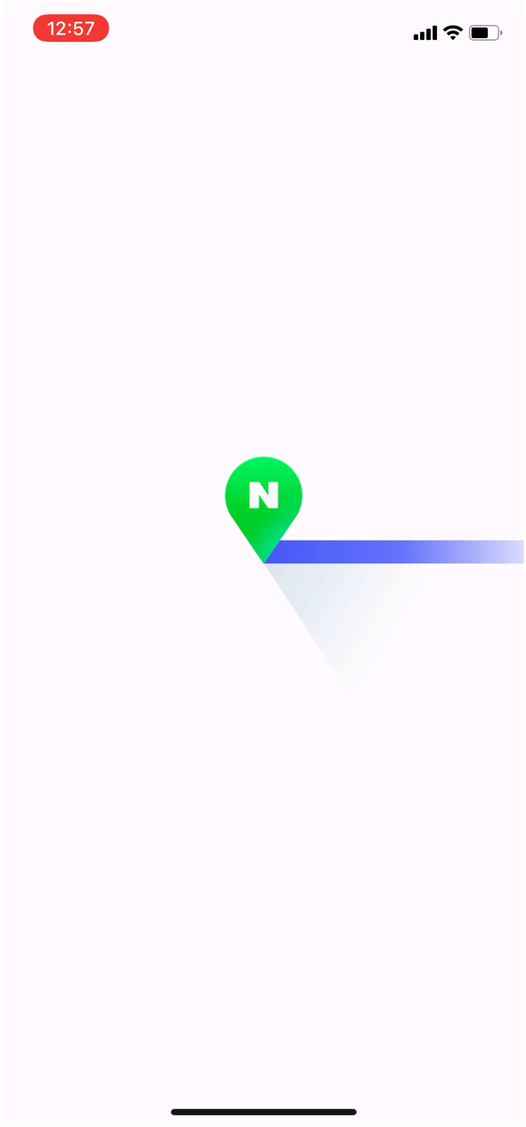
한국어 / 일본어 음성 어시스턴트



네이버뉴스

2020.04~현재

한국어 뉴스 자동자막생성



네이버지도

2016.12~현재
주소검색 / 플레이스 검색
클로바 어시스턴트 탑재

2.지도 음성인식의 어려움

2.1 주소, POI의 방대함

2.1 주소, POI의 방대함

서울시

2.1 주소, POI의 방대함

서울시

자치단체

2.1 주소, POI의 방대함

서울시 강남구

자치단체

2.1 주소, POI의 방대함

서울시 강남구

자치단체 일반구

2.1 주소, POI의 방대함

서울시 강남구 도산대로45길

자치단체 일반구

2.1 주소, POI의 방대함

서울시 강남구 도산대로45길

자치단체 일반구 도로명

2.1 주소, POI의 방대함

서울시 강남구 도산대로45길 20

자치단체 일반구 도로명

2.1 주소, POI의 방대함

서울시 강남구 도산대로45길 20

자치단체 일반구 도로명 건물번호

2.1 주소, POI의 방대함

서울시 강남구 도산대로45길 20 도산공원

자치단체 일반구 도로명 건물번호

2.1 주소, POI의 방대함

서울시	강남구	도산대로45길	20	도산공원
자치단체	일반구	도로명	건물번호	상세주소

2.1 주소, POI의 방대함

DEVIEW 2020

서울시 강남구 도산대로45길 20 도산공원

자치단체 일반구 도로명 건물번호 상세주소

total depth : 5~6

total address number : >10M

도로명 / 지번, 상점 이름 등등... : >1B

2.1 주소, POI의 방대함

도산대로45길 20 도산공원

도로명 건물번호 상세주소

total depth : 5~6

total address number : >10M

도로명 / 지번, 상점 이름 등등... : >1B

2.1 주소, POI의 방대함

DEVIEW 2020

강남구 도산대로45길

일반구 도로명

total depth : 5~6

total address number : >10M

도로명 / 지번, 상점 이름 등등... : >1B

2.1 주소, POI의 방대함

도산대로45길 도산공원

도로명 상세주소

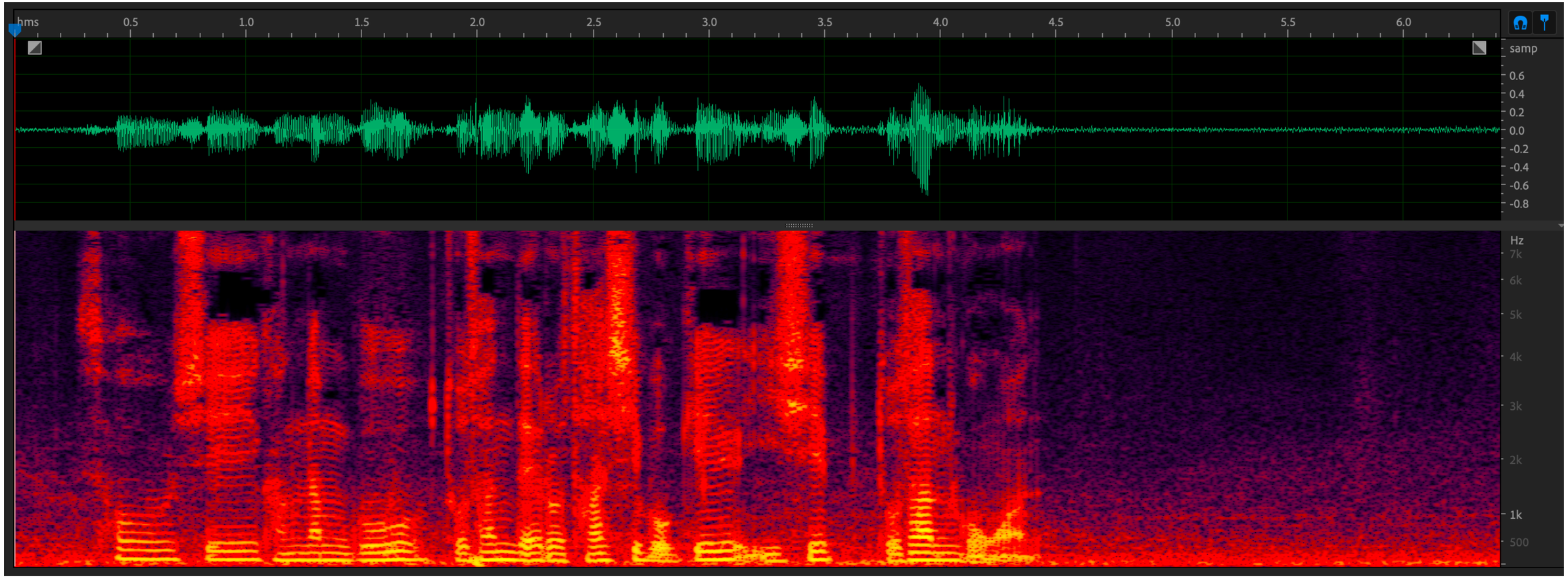
total depth : 5~6

total address number : >10M

도로명 / 지번, 상점 이름 등등... : >1B

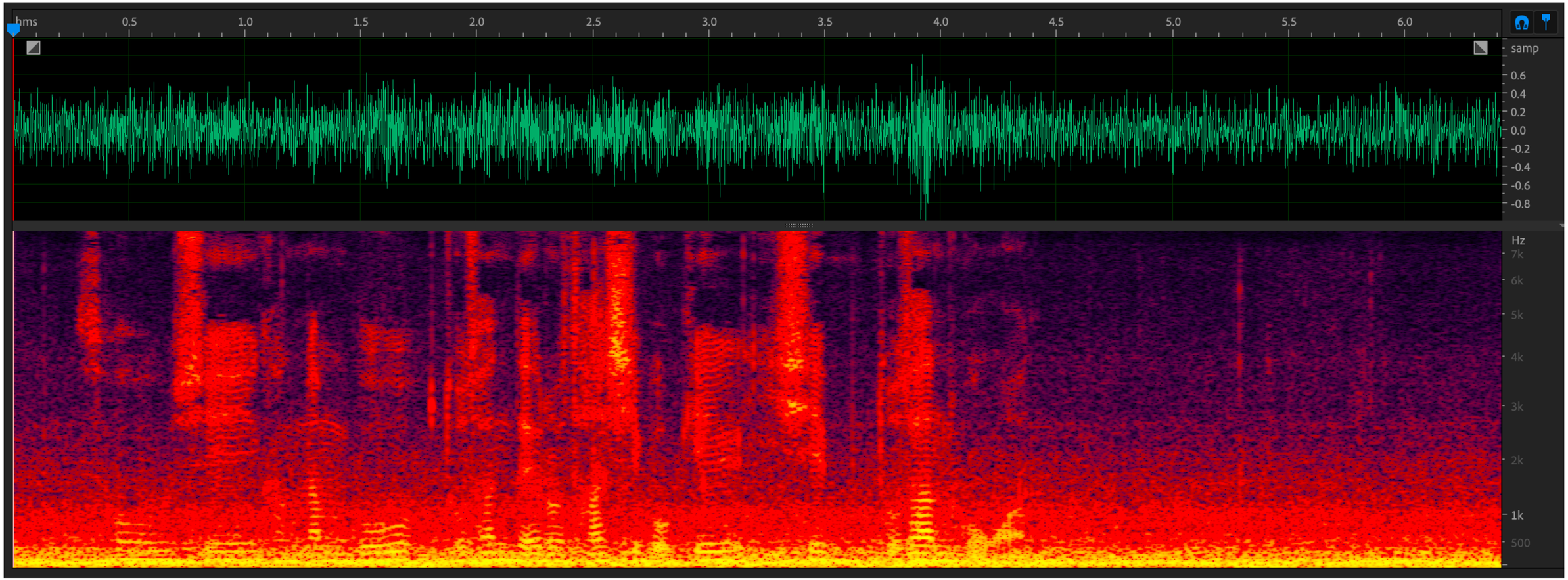
2.2 차량에 의한 잡음

DEVIEW 2020



2.2 차량에 의한 잡음

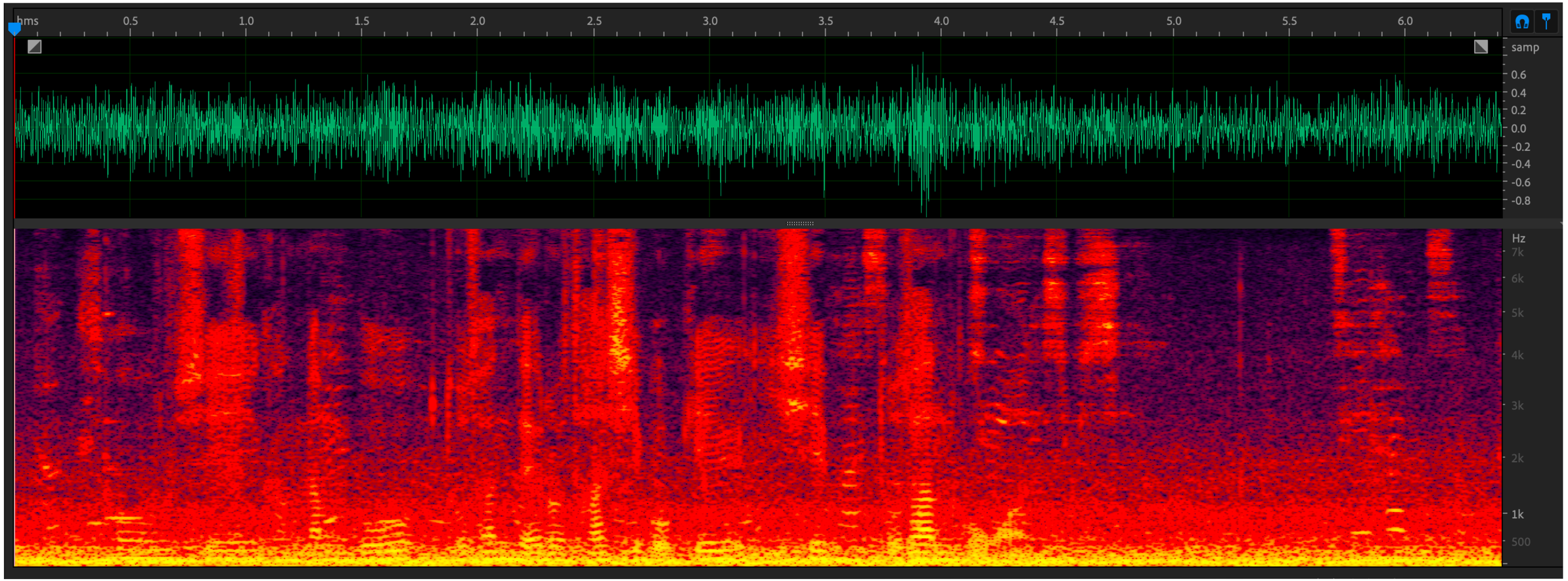
DEVIEW 2020



+ 차량 주행 소음

2.2 차량에 의한 잡음

DEVIEW 2020



+ 차량 주행 소음 + 기타 잡음

2. 결론

DEVIEW 2020

네이버지도 음성인식 서비스의 특징

1. 위치 검색에서는 대상이 방대하나, 발화 도메인이 정적이다.
   (주소표기법은 자주 변하지 않음)

2. 잡음이 심하지만, 패턴이 비교적 정형화 되어 있다.
   (차량 주행소음, 차량 내 잡음 패턴, 차량 공간 자체가 정형화 되어 있음.)

DEVIEW 2020

3.네이버지도를 위한 음성인식 개선

3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

1. HMM-DNN based Acoustic Model

A. Forward Phase

Dahl, G. E., Yu, D., Deng, L. & Acero, A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20, 33–42 (2012).

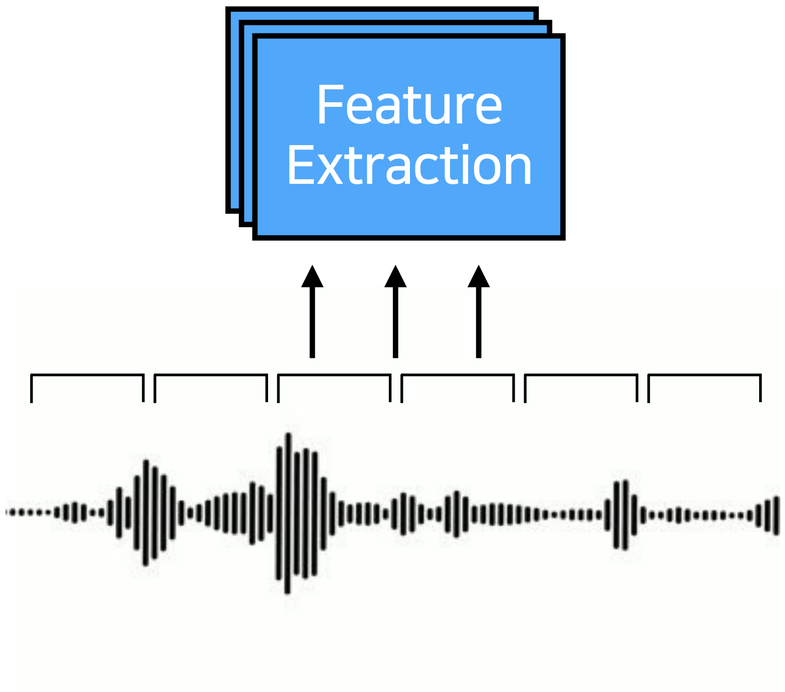
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

1. HMM-DNN based Acoustic Model

A. Forward Phase



Dahl, G. E., Yu, D., Deng, L. & Acero, A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20, 33–42 (2012).

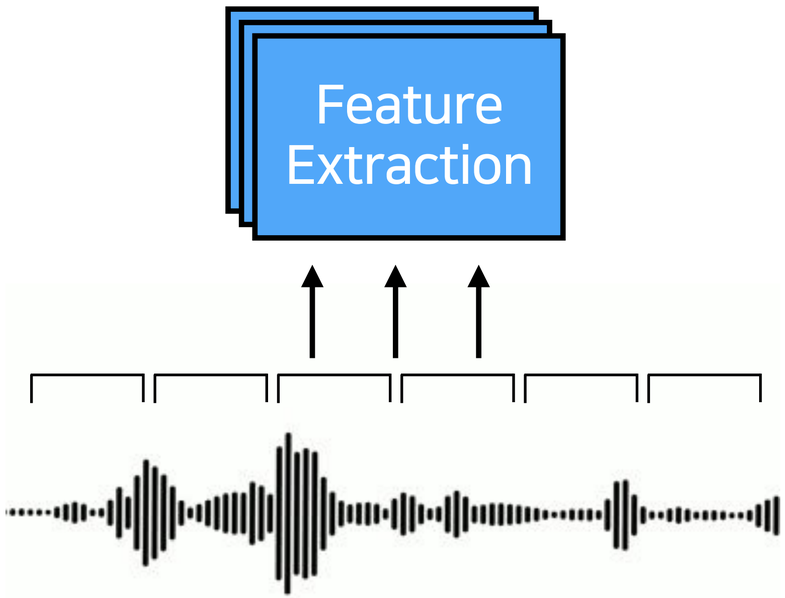
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

1. HMM-DNN based Acoustic Model

A. Forward Phase



1. Spectral analysis
   (spectrogram, log mel-filterbank)
2. Cepstral analysis (MFCC)
3. Neural analysis
   (CNN, pyramidal LSTM, wav2vec, ...)

Time-frame Segmentation (windowing)

Dahl, G. E., Yu, D., Deng, L. & Acero, A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20, 33–42 (2012).

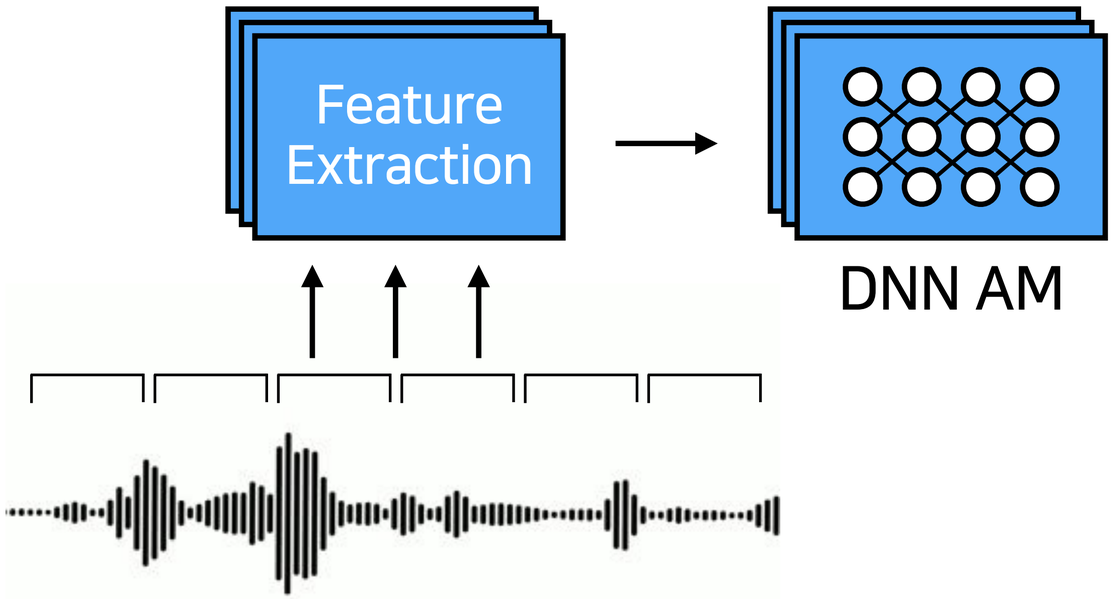
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

1. HMM-DNN based Acoustic Model

A. Forward Phase



Dahl, G. E., Yu, D., Deng, L. & Acero, A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20, 33–42 (2012).

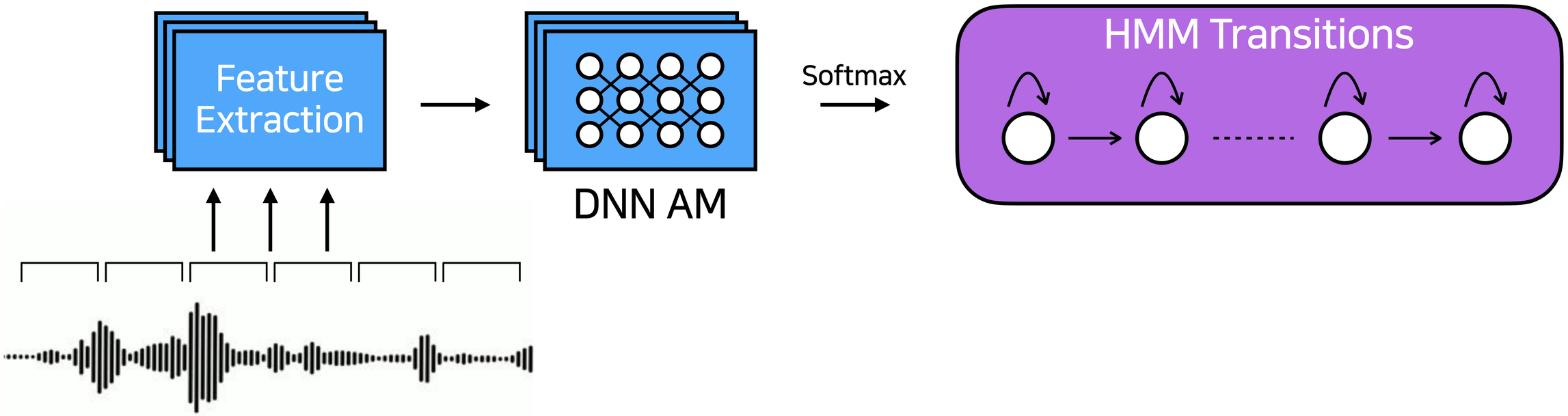
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

1. HMM-DNN based Acoustic Model

A. Forward Phase



Dahl, G. E., Yu, D., Deng, L. & Acero, A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20, 33–42 (2012).

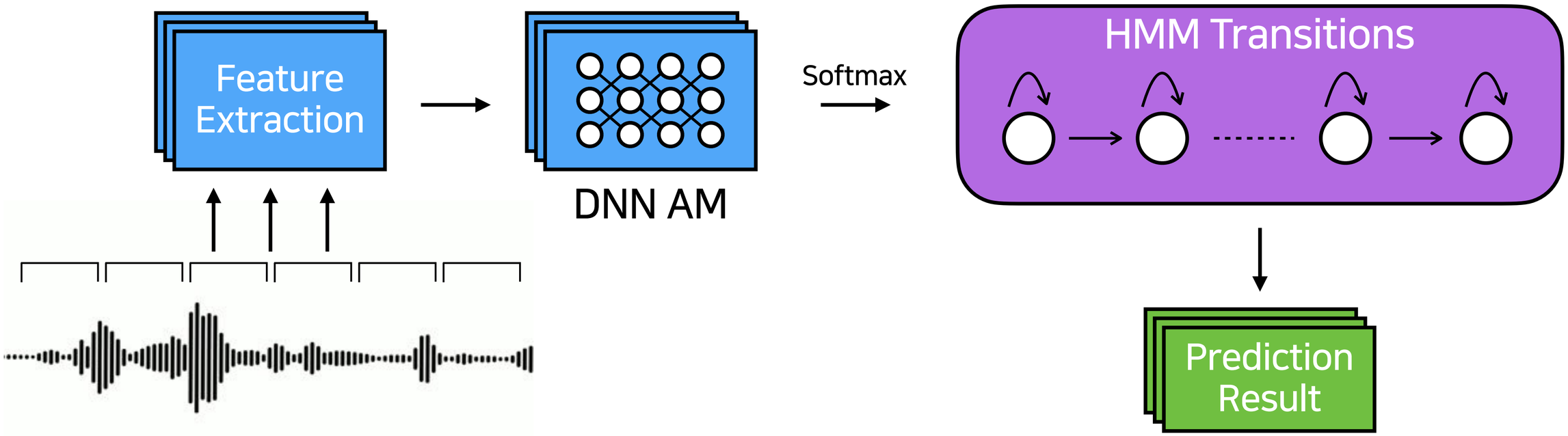
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

1. HMM-DNN based Acoustic Model

A. Forward Phase



Dahl, G. E., Yu, D., Deng, L. & Acero, A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20, 33–42 (2012).

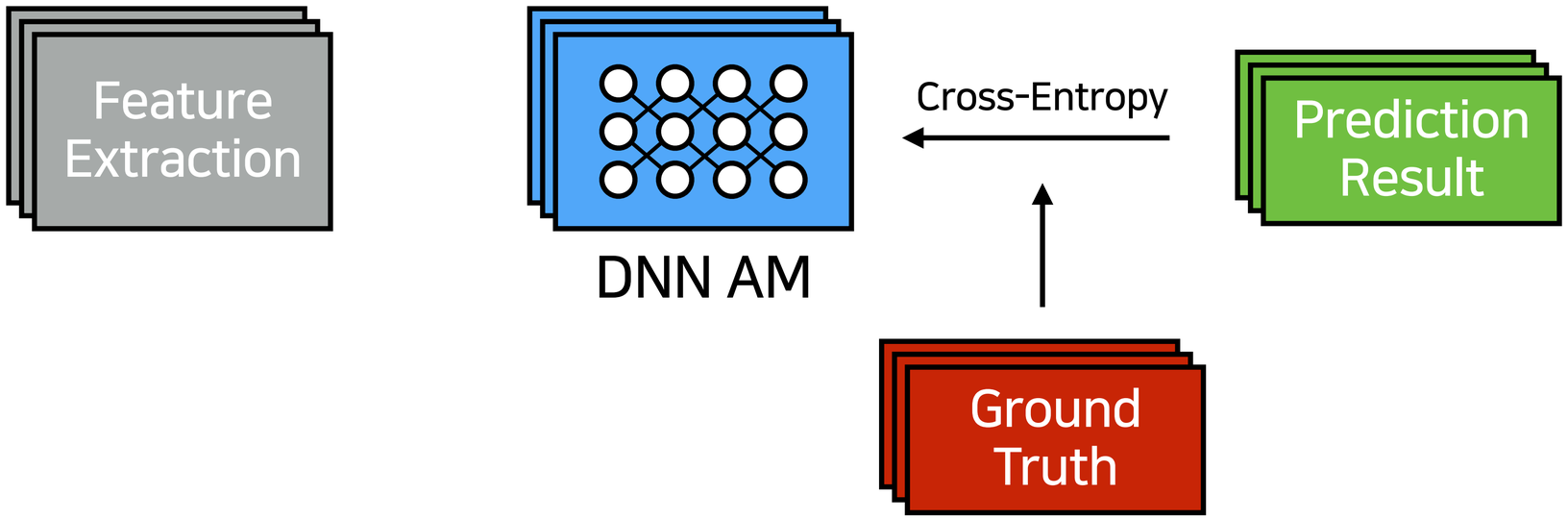
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

1. HMM-DNN based Acoustic Model

B. Backward Phase



Dahl, G. E., Yu, D., Deng, L. & Acero, A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20, 33–42 (2012).

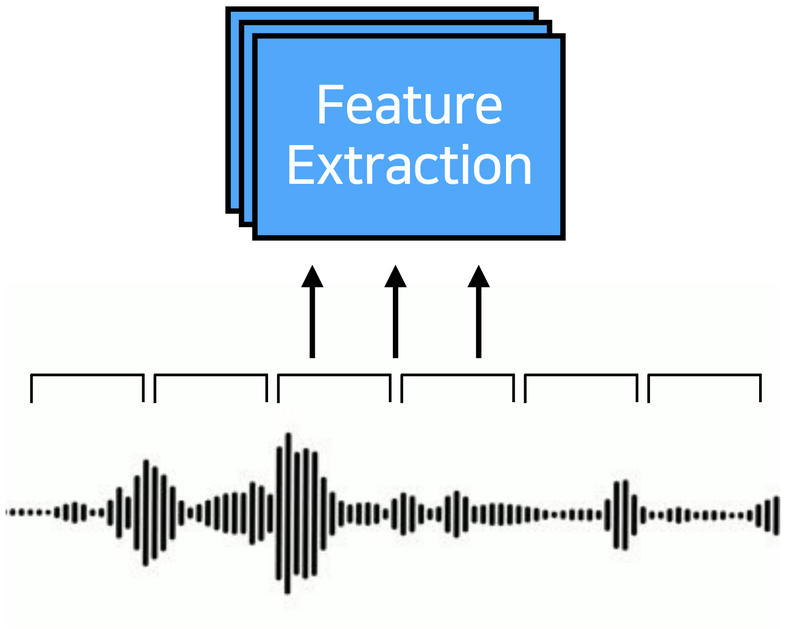
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

2. End-to-end Speech Recognition Model

A. Forward Phase



W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” CoRR, vol. abs/1508.01211, 2015.

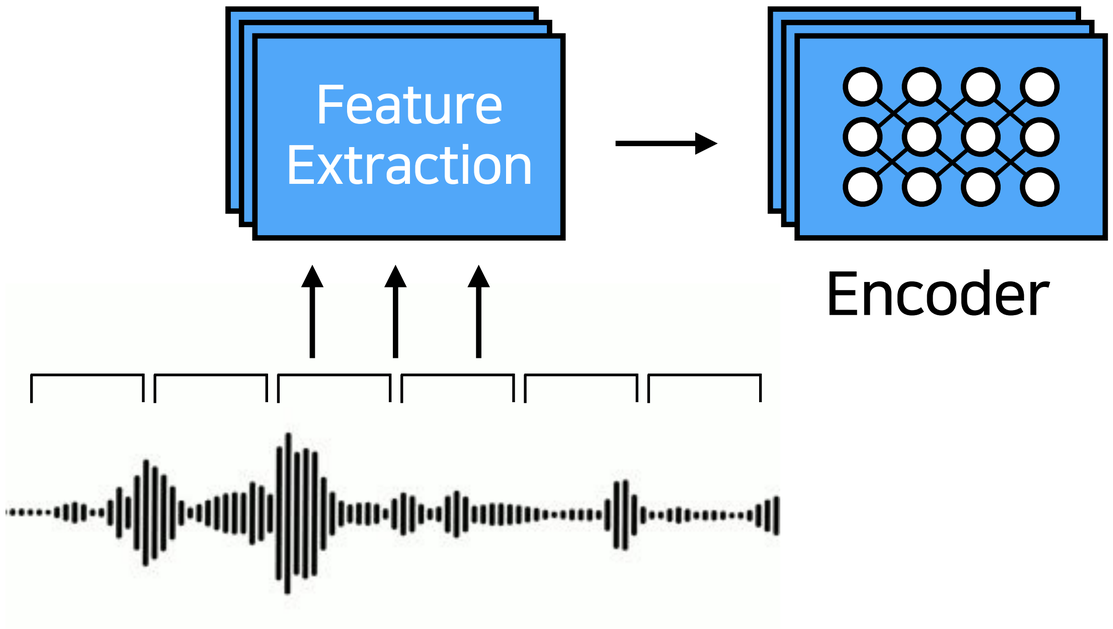
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

2. End-to-end Speech Recognition Model

A. Forward Phase



W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” CoRR, vol. abs/1508.01211, 2015.

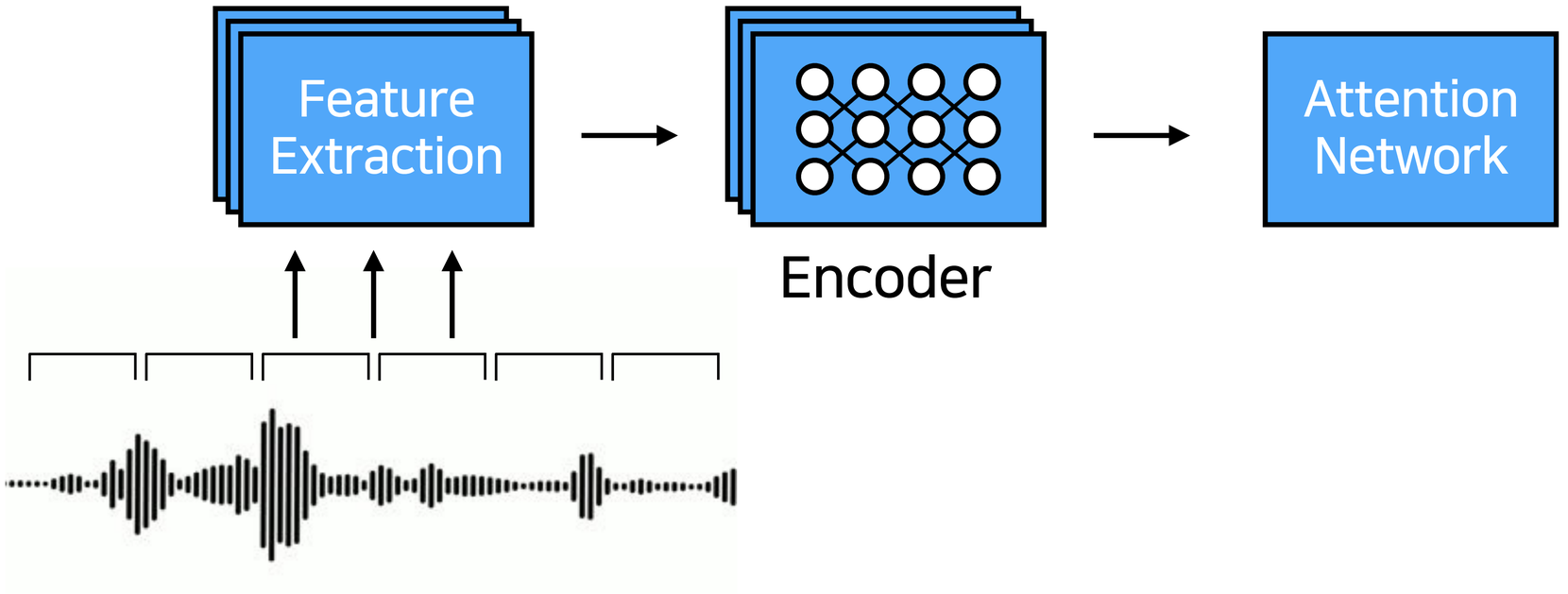
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

2. End-to-end Speech Recognition Model

A. Forward Phase



W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” CoRR, vol. abs/1508.01211, 2015.

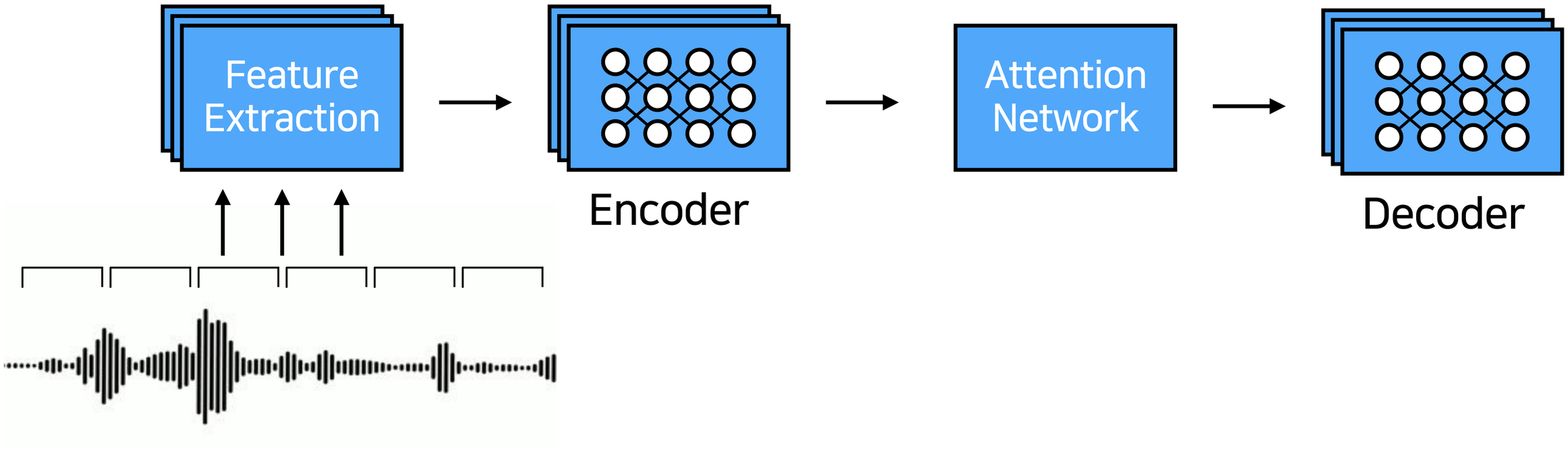
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

2. End-to-end Speech Recognition Model

A. Forward Phase



W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” CoRR, vol. abs/1508.01211, 2015.

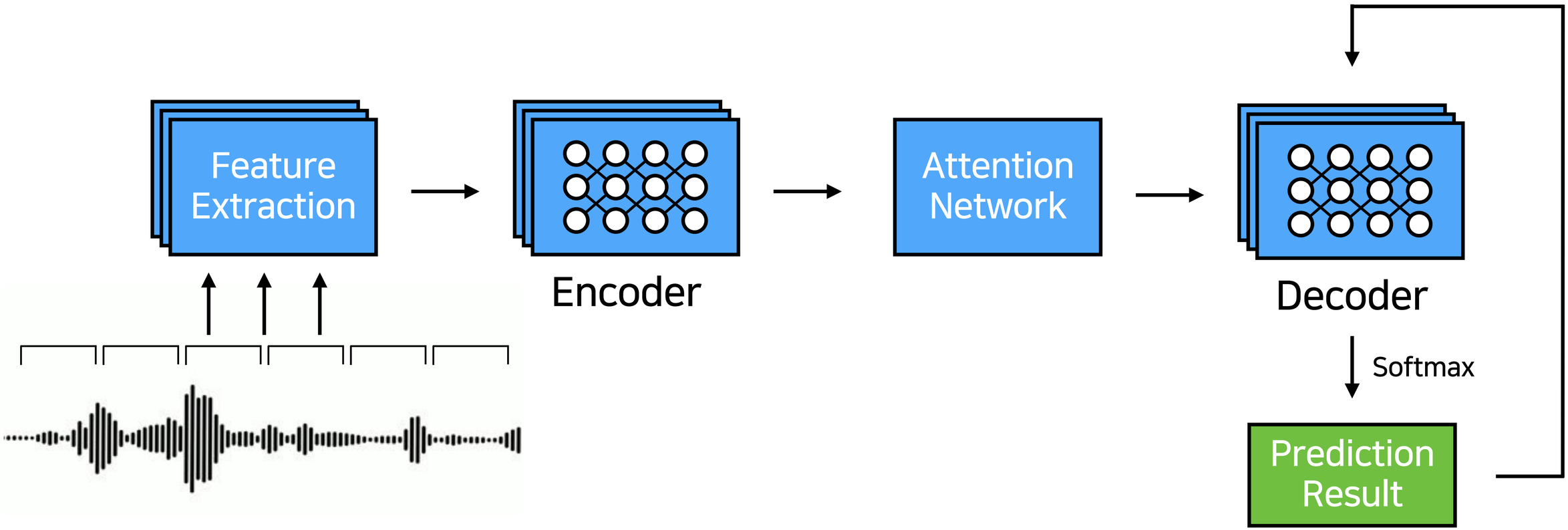
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

2. End-to-end Speech Recognition Model

A. Forward Phase



W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” CoRR, vol. abs/1508.01211, 2015.

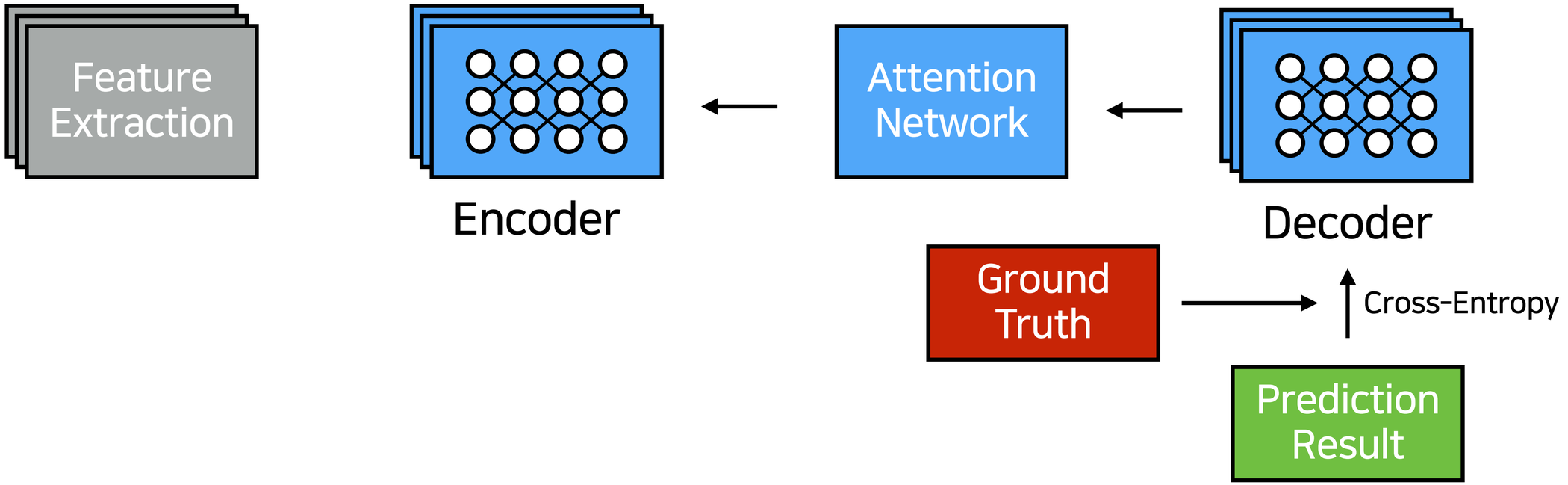
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

2. End-to-end Speech Recognition Model

B. Backward Phase



W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” CoRR, vol. abs/1508.01211, 2015.

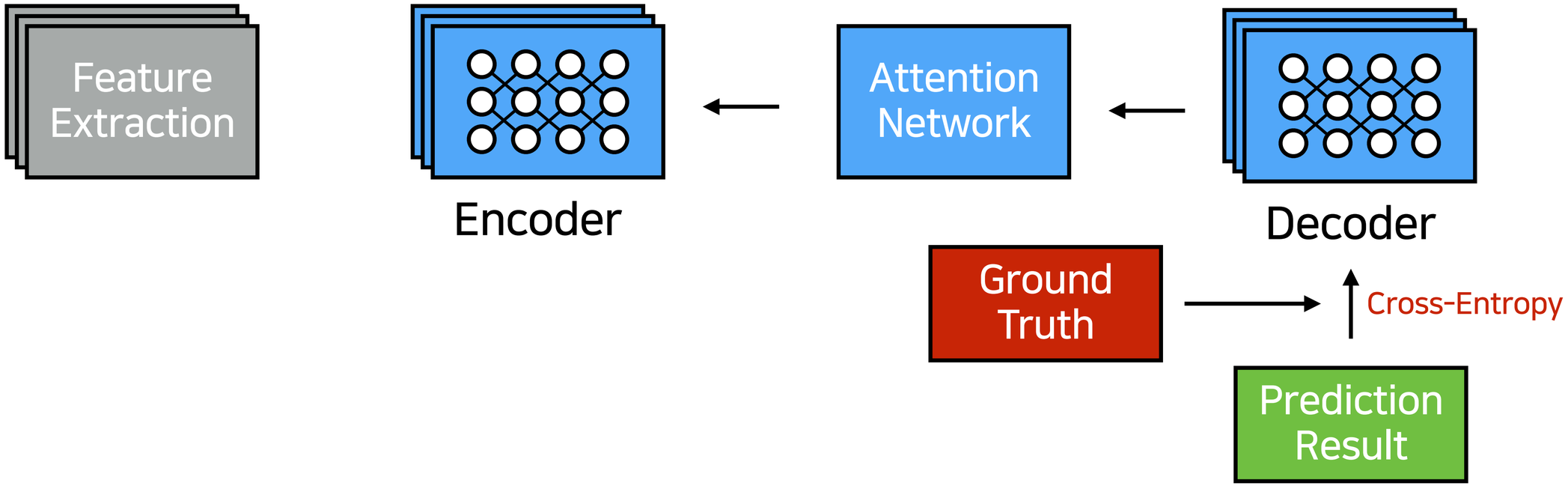
3.1 Sequence-based Acoustic Model

DEVIEW 2020

Conventional Acoustic Model

2. End-to-end Speech Recognition Model

   B. Backward Phase



W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” CoRR, vol. abs/1508.01211, 2015.

3.1 Sequence-based Acoustic Model

General Maximum Likelihood Objective :

$$\lambda = \arg\max_{\lambda} F_{ML}$$

$$F_{ML} = \log p_{\lambda}(X \mid W_r)$$

X : Input Features

W_r : Ground Truth Word Seqeunce

λ : Model Parameters

3.1 Sequence-based Acoustic Model

DEVIEW 2020

Expansion to Maximum Mutual Information :

$$F_{MMI} = \log \frac{p_\lambda(X \mid W_r)}{p_\lambda(X)}$$

Daniel Povey, "Discriminative Training for Large Vocabulary Speech Recognition," PhD thesis, Cambridge University Engineering Dept, 2003

3.1 Sequence-based Acoustic Model

DEVIEW 2020

Expansion to Maximum Mutual Information :

$$F_{MMI} = \log \frac{p_\lambda(X \mid W_r)}{p_\lambda(X)}$$

$p_\lambda(X \mid W_r)$: conditional probability of X given W_r

$p_\lambda(X)$: marginal probability of X

Daniel Povey, "Discriminative Training for Large Vocabulary Speech Recognition," PhD thesis, Cambridge University Engineering Dept, 2003

3.1 Sequence-based Acoustic Model

Expansion to Maximum Mutual Information :

$$F_{MMI} = \log \frac{p_\lambda(X \mid W_r)}{p_\lambda(X)}$$

$p_\lambda(X \mid W_r)$: conditional probability of X given W_r

$p_\lambda(X)$: marginal probability of X

Bayes' Rule

$$= \log \frac{p(X, W_r)}{\sum_{\hat{W}} p(X, \hat{W})} - \log p(W_r)$$

Joint Probability Expansion

$\hat{W}$: Any Possible Word Seqeunces

3.1 Sequence-based Acoustic Model

Expansion to Maximum Mutual Information :

$$F_{MMI} = \log \frac{p_\lambda(X \mid W_r)}{p_\lambda(X)}$$

$p_\lambda(X \mid W_r)$: conditional probability of X given W_r

$p_\lambda(X)$: marginal probability of X

$$= \log \frac{p(X, W_r)}{\sum_{\hat{W}} p(X, \hat{W})} - \log p(W_r)$$

Bayes' Rule

$$\propto \log \frac{p(X \mid W_r)p(W_r)}{\sum_{\hat{W}} p(X \mid \hat{W})p(\hat{W})}$$

$\hat{W}$: Any Possible Word Seqeunces

3.1 Sequence-based Acoustic Model

Expansion to Maximum Mutual Information :

$$F_{MMI} = \log \frac{p(X \mid W_r)p(W_r)}{\sum_{\hat{W}} p(X \mid \hat{W})p(\hat{W})}$$

AM Likelihood

LM Prior

3.1 Sequence-based Acoustic Model

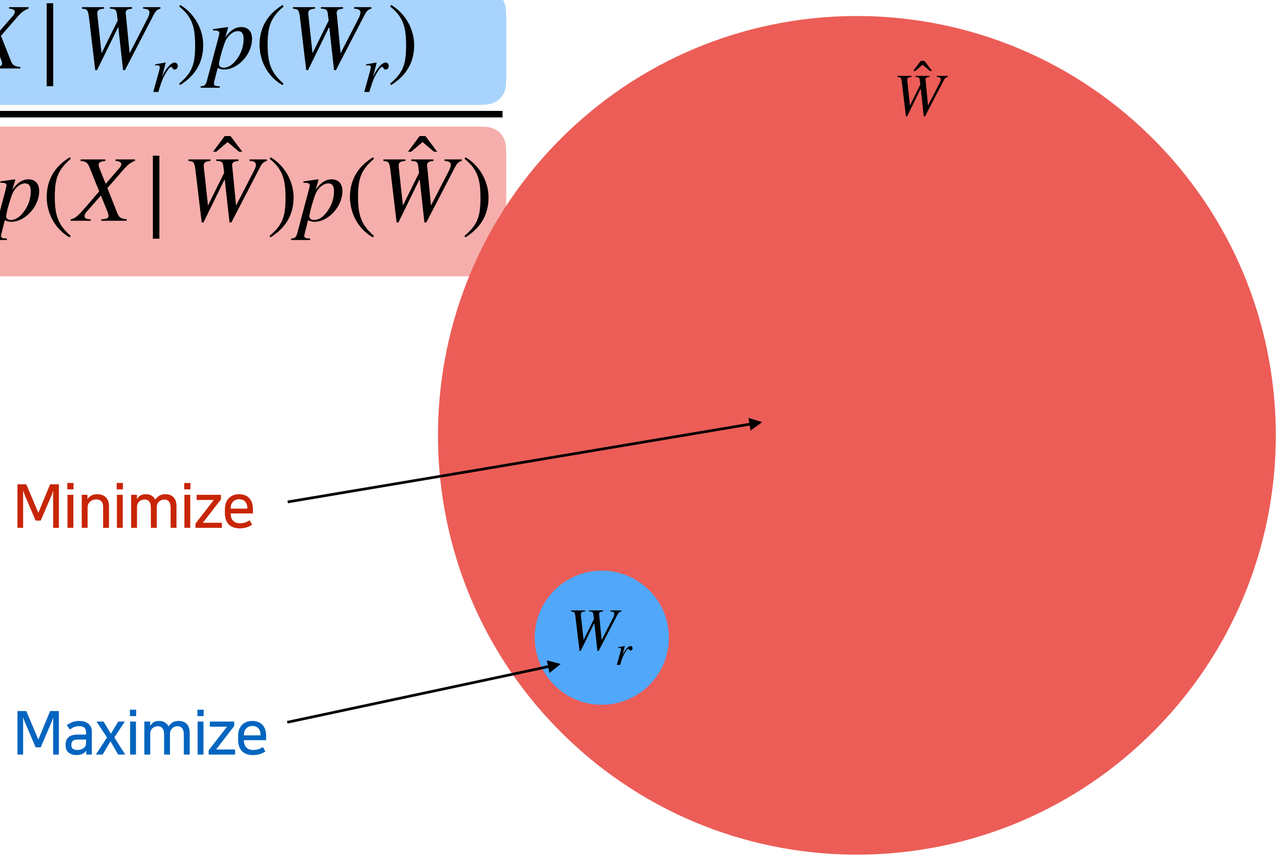
Expansion to Maximum Mutual Information :

$$F_{MMI} = \log \frac{p(X \mid W_r)p(W_r)}{\sum_{\hat{W}} p(X \mid \hat{W})p(\hat{W})}$$

3.1 Sequence-based Acoustic Model

Expansion to Maximum Mutual Information :

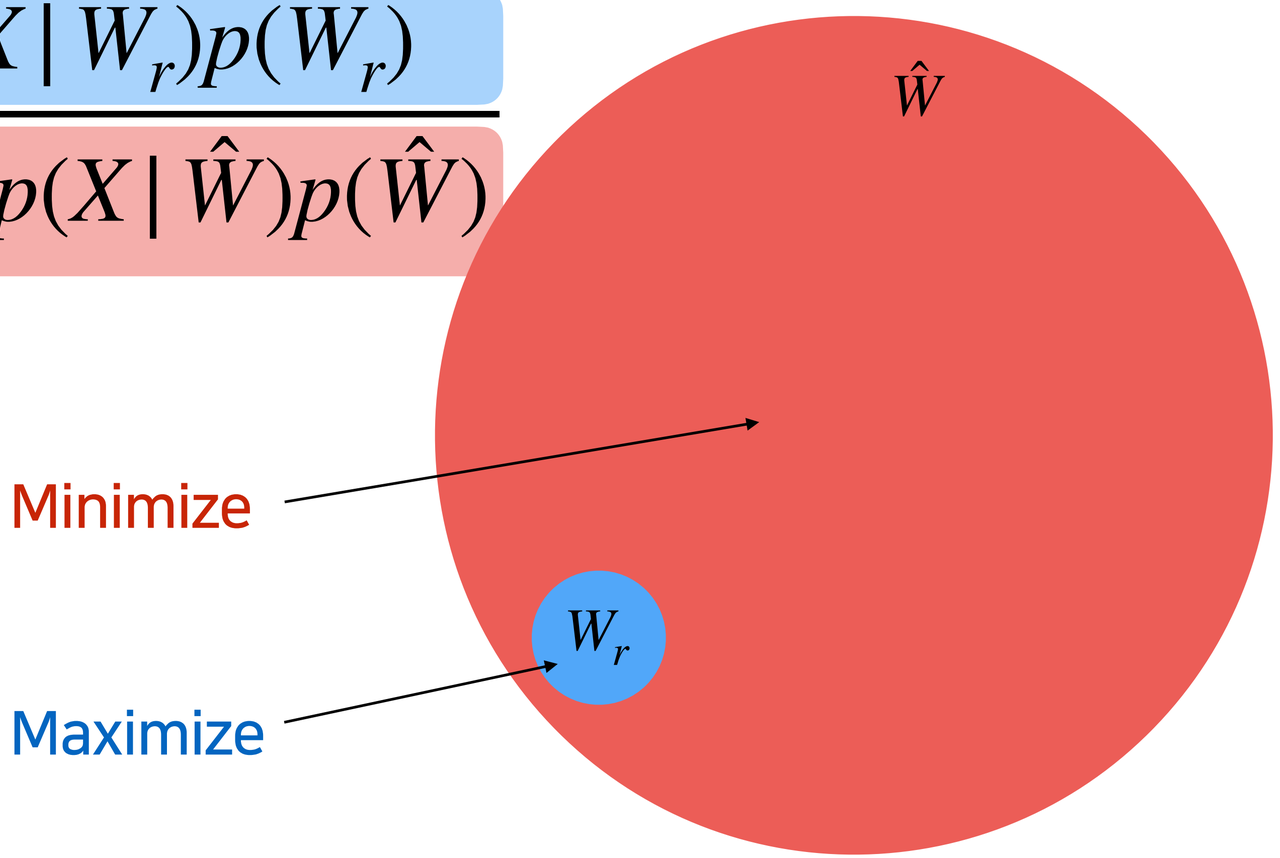
$$F_{MMI} = \log \frac{p(X \mid W_r)p(W_r)}{\sum_{\hat{W}} p(X \mid \hat{W})p(\hat{W})}$$



3.1 Sequence-based Acoustic Model

Expansion to Maximum Mutual Information :

$$F_{MMI} = \log \frac{p(X \mid W_r)p(W_r)}{\sum_{\hat{W}} p(X \mid \hat{W})p(\hat{W})}$$

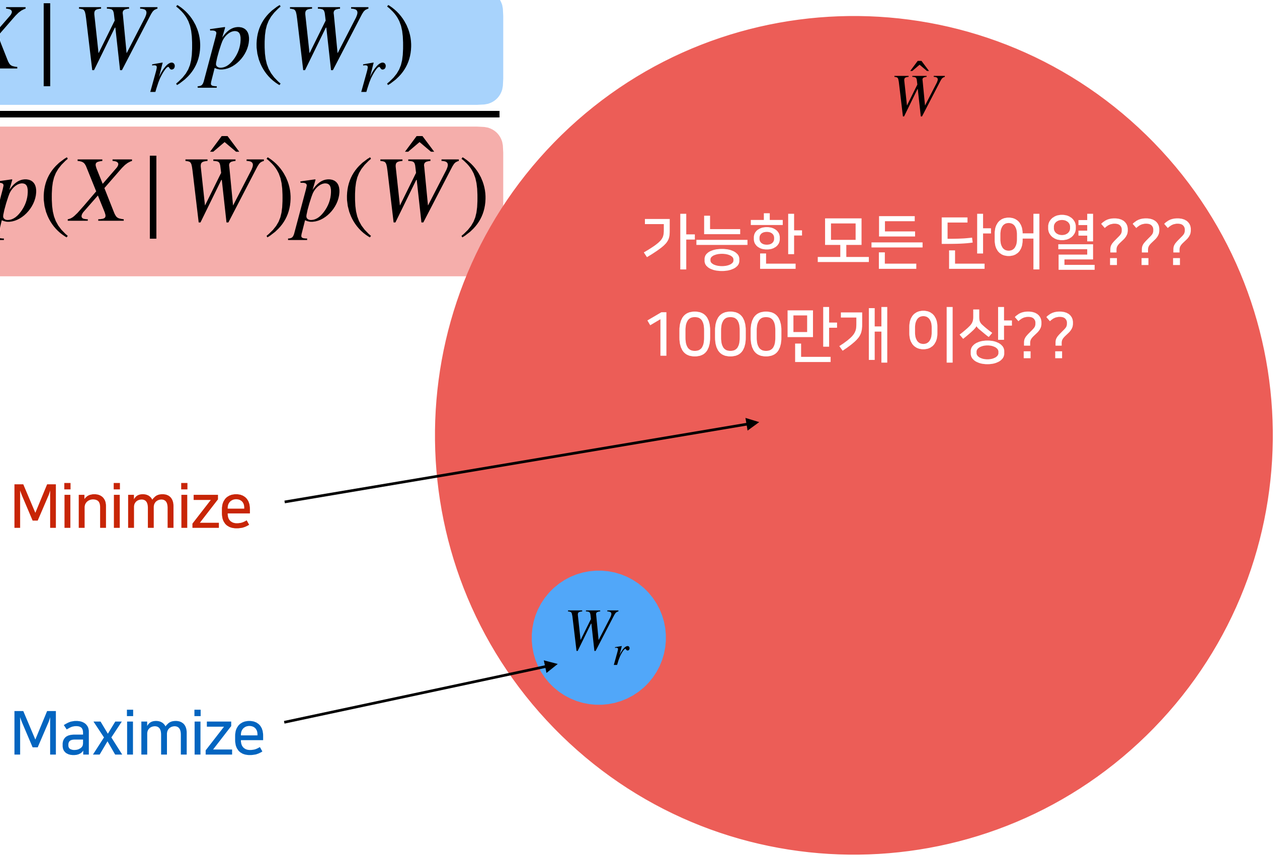


"Sequence Discriminative Training"

3.1 Sequence-based Acoustic Model

Expansion to Maximum Mutual Information :

$$F_{MMI} = \log \frac{p(X \mid W_r)p(W_r)}{\sum_{\hat{W}} p(X \mid \hat{W})p(\hat{W})}$$



"Sequence Discriminative Training"

3.1 Sequence-based Acoustic Model

DEVIEW 2020

Word Sequences to Phone Sequences

$$F_{MMI} = \log \frac{p(X \mid W_r)p(W_r)}{\sum_{\hat{W}} p(X \mid \hat{W})p(\hat{W})}$$

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahrmani, Vimal Manohar, Xingyu, Na, Yiming Wang and Sanjeev Khudanpur, “Purely sequence-trained neural networks for ASR based on lattice-free MMI” in Proc. Interspeech, 2016.

3.1 Sequence-based Acoustic Model

DEVIEW 2020

Word Sequences to Phone Sequences

$$F_{MMI} = \log \frac{p(X \mid W_r)p(W_r)}{\sum_{\hat{W}} p(X \mid \hat{W})p(\hat{W})}$$

$$\propto \log \frac{p(X \mid P_r)p(P_r)}{\sum_{\hat{P}} p(X \mid \hat{P})p(\hat{P})} \leftarrow$$

using FST graph ,
split to small fixed-size chunk,
& GPU parallelize

$\hat{P}$: Any Possible Phone Seqeunces

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahrmani, Vimal Manohar, Xingyu, Na, Yiming Wang and Sanjeev Khudanpur, “Purely sequence-trained neural networks for ASR based on lattice-free MMI” in Proc. Interspeech, 2016.

3.1 Sequence-based Acoustic Model

DEVIEW 2020

Word Sequences to Phone Sequences & Using Phone FST

$$F_{MMI} = \log \frac{p(X \mid W_r)p(W_r)}{\sum_{\hat{W}} p(X \mid \hat{W})p(\hat{W})}$$

$$F_{LF-MMI} = \log \frac{p(X \mid P_r)p(P_r)}{\sum_{\hat{P}} p(X \mid \hat{P})p(\hat{P})}$$

without using word lattices -> lattice-free MMI

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahrmani, Vimal Manohar, Xingyu, Na, Yiming Wang and Sanjeev Khudanpur, “Purely sequence-trained neural networks for ASR based on lattice-free MMI” in Proc. Interspeech, 2016.

3.1 Sequence-based Acoustic Model

Table 3: Performance of LF-MMI with different models on the Hub5 '00 eval set, using SWBD-300 Hr data

Model	WER	
	Total	SWBD
TDNN-C + CE	18.2	12.5
TDNN-C + LF-MMI	15.5	10.2
LSTM + CE	16.5	11.6
LSTM + LF-MMI	15.6	10.3
BLSTM + CE	14.9	10.3
BLSTM + LF-MMI	14.5	9.6

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahrmani, Vimal Manohar, Xingyu, Na, Yiming Wang and Sanjeev Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI" in Proc. Interspeech, 2016.

3.2 Class-based Language Model

3.2 Class-based Language Model

서울시 강남구 도산대로45길 20 도산공원

3.2 Class-based Language Model

DEVIEW 2020

서울시	강남구	도산대로45길	20	도산공원
부산시	마포구	44길	19	
인천시	종로구	43길	18	
대구시	용산구	42길	17	
경기도	서초구	41길	16	

3.2 Class-based Language Model

DEVIEW 2020

서울시	강남구	도산대로45길	20	도산공원
부산시	마포구	44길	19	
인천시	종로구	43길	18	
대구시	용산구	42길	17	
경기도	서초구	41길	16	

3.2 Class-based Language Model

같은 depth에 있는 entry는 같은 확률로 tying!

서울시	강남구	도산대로45길	20	도산공원
부산시	마포구	44길	19	
인천시	종로구	43길	18	
대구시	용산구	42길	17	
경기도	서초구	41길	16	

3.2 Class-based Language Model

DEVIEW 2020

같은 depth에 있는 entry는 같은 확률로 tying!

서울시	강남구	도산대로45길	20	도산공원
부산시	마포구	44길	19	
인천시	종로구	43길	18	
대구시	용산구	42길	17	
경기도	서초구	41길	16	

entry 간의 bias를 해소 -> 음성인식 성능 개선!

3.3 운영 모니터링 & 취약 데이터 확보

DEVIEW 2020

오성면 양교리 [illegible]

반포 자이

과속방지턱 구간입니다

풍호동 [illegible]

도로 교통 상황 보여줘

문촌 마을 6단지

경기도 부천시 신흥로 [illegible]

의정부역 센트럴자이

고려대로 13 길 [illegible]

광주시 도척면

우리은행

송도동 [illegible]

* 음성 인식 기술의 향상을 위해 사용자 음성의 일부가 학습 데이터로 활용되고 있으나,
이를 사용자가 '옵트 아웃' 기능을 통해 음성데이터 활용에 선택권을 보장하고 있으며,
수집된 음성 데이터를 누구의 음성인지 알 수 없도록 처리하여 다루고 있습니다.

3.3 운영 모니터링 & 취약 데이터 확보

DEVIEW 2020

오성면 양교리 [illegible]

반포 자이

?

과속방지턱 구간입니다

풍호동 [illegible]

도로 교통 상황 보여줘

문촌 마을 6단지

경기도 부천시 신흥로 [illegible]

의정부역 센트럴자이

고려대로 13 길 [illegible]

광주시 도척면

우리은행

송도동 [illegible]

* 음성 인식 기술의 향상을 위해 사용자 음성의 일부가 학습 데이터로 활용되고 있으나,
이를 사용자가 '옵트 아웃' 기능을 통해 음성데이터 활용에 선택권을 보장하고 있으며,
수집된 음성 데이터를 누구의 음성인지 알 수 없도록 처리하여 다루고 있습니다.

3.3 운영 모니터링 & 취약 데이터 확보

DEVIEW 2020

과속방지턱입니다

과속에 주의하십시오

시속 80킬로미터 이동식 카메라 구간입니다

경로 안내를 시작합니다

잠시후 일정 도시 고속도로 입구로 주행하십시오

시속 60킬로미터 구간입니다

잠시 후 오른쪽 방향입니다

비보호좌회전입니다

260미터 앞 우회전입니다

약 300미터 육십킬로미터 신호 위반 과속단속 구간입니다

사고 다발 지역이 있습니다

전방에 사고 다발 구간입니다

1킬로미터 앞 우회전입니다

목적지에 도착했습니다

시속 100킬로미터로 가는 거잖아요

약 400미터 60킬로미터 구간입니다

안전운행 하십시요

* 음성 인식 기술의 향상을 위해 사용자 음성의 일부가 학습 데이터로 활용되고 있으나,
이를 사용자가 '옵트 아웃' 기능을 통해 음성데이터 활용에 선택권을 보장하고 있으며,
수집된 음성 데이터를 누구의 음성인지 알 수 없도록 처리하여 다루고 있습니다.

3.3 운영 모니터링 & 취약 데이터 확보

DEVIEW 2020

과속방지턱입니다

과속에 주의하십시오

시속 80킬로미터 이동식 카메라 구간입니다

경로 안내를 시작합니다

잠시후 일정 도시 고속도로 입구로 주행하십시오

시속 60킬로미터 구간입니다

잠시 후 오른쪽 방향입니다

비보호좌회전입니다

260미터 앞 우회전입니다

약 300미터 육십킬로미터 신호 위반 과속단속 구간입니다

사고 다발 지역이 있습니다

전방에 사고 다발 구간입니다

1킬로미터 앞 우회전입니다

목적지에 도착했습니다

시속 100킬로미터로 가는 거잖아요

약 400미터 60킬로미터 구간입니다

안전운행 하십시요

사용자 발화가 아닐 가능성이
매우 높기 때문에
사용자 경험 측면에서는
silence로 학습시키는게 좋다!

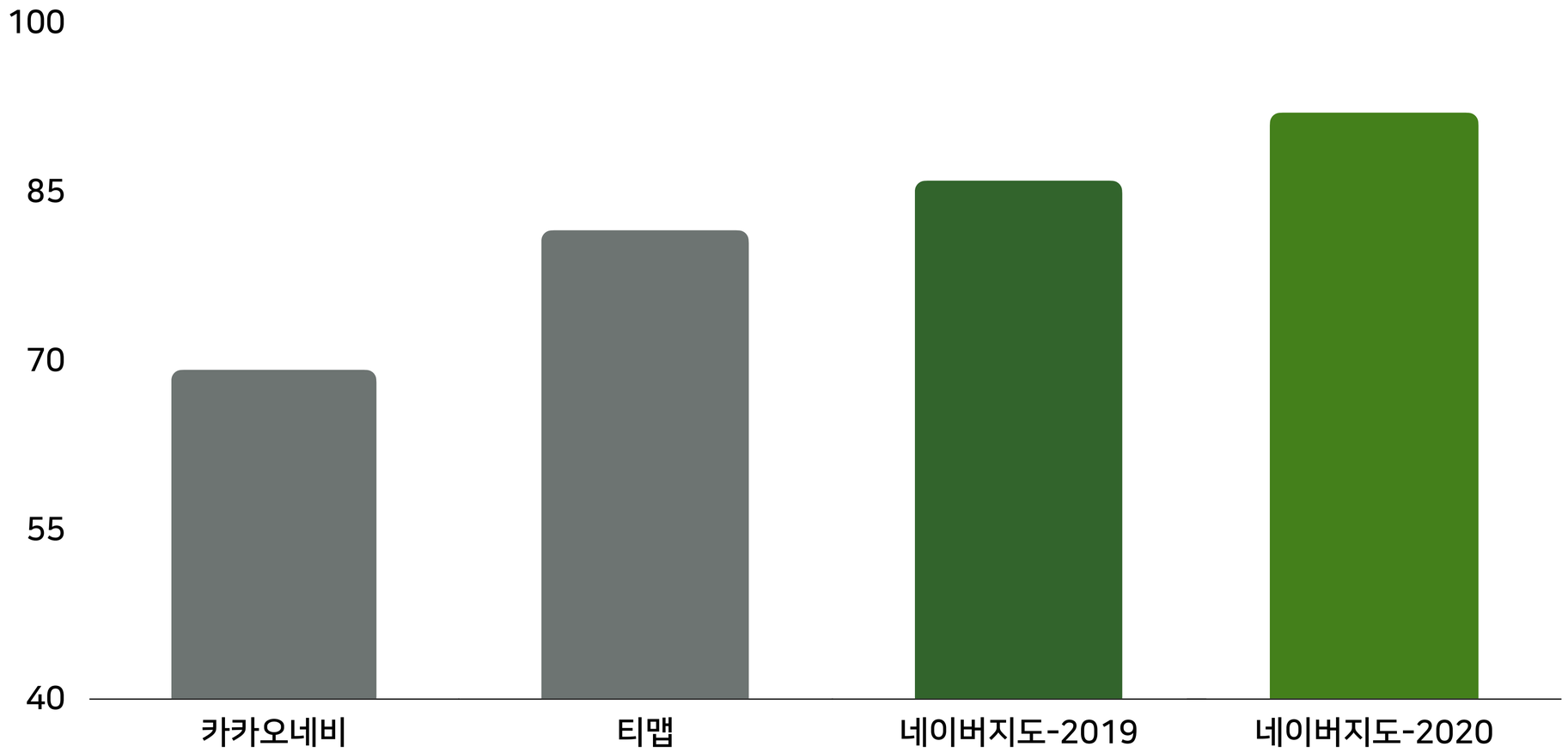
* 음성 인식 기술의 향상을 위해 사용자 음성의 일부가 학습 데이터로 활용되고 있으나,
이를 사용자가 '옵트 아웃' 기능을 통해 음성데이터 활용에 선택권을 보장하고 있으며,
수집된 음성 데이터를 누구의 음성인지 알 수 없도록 처리하여 다루고 있습니다.

4. 성능 확인

4. 실차 테스트를 통한 성능 확인

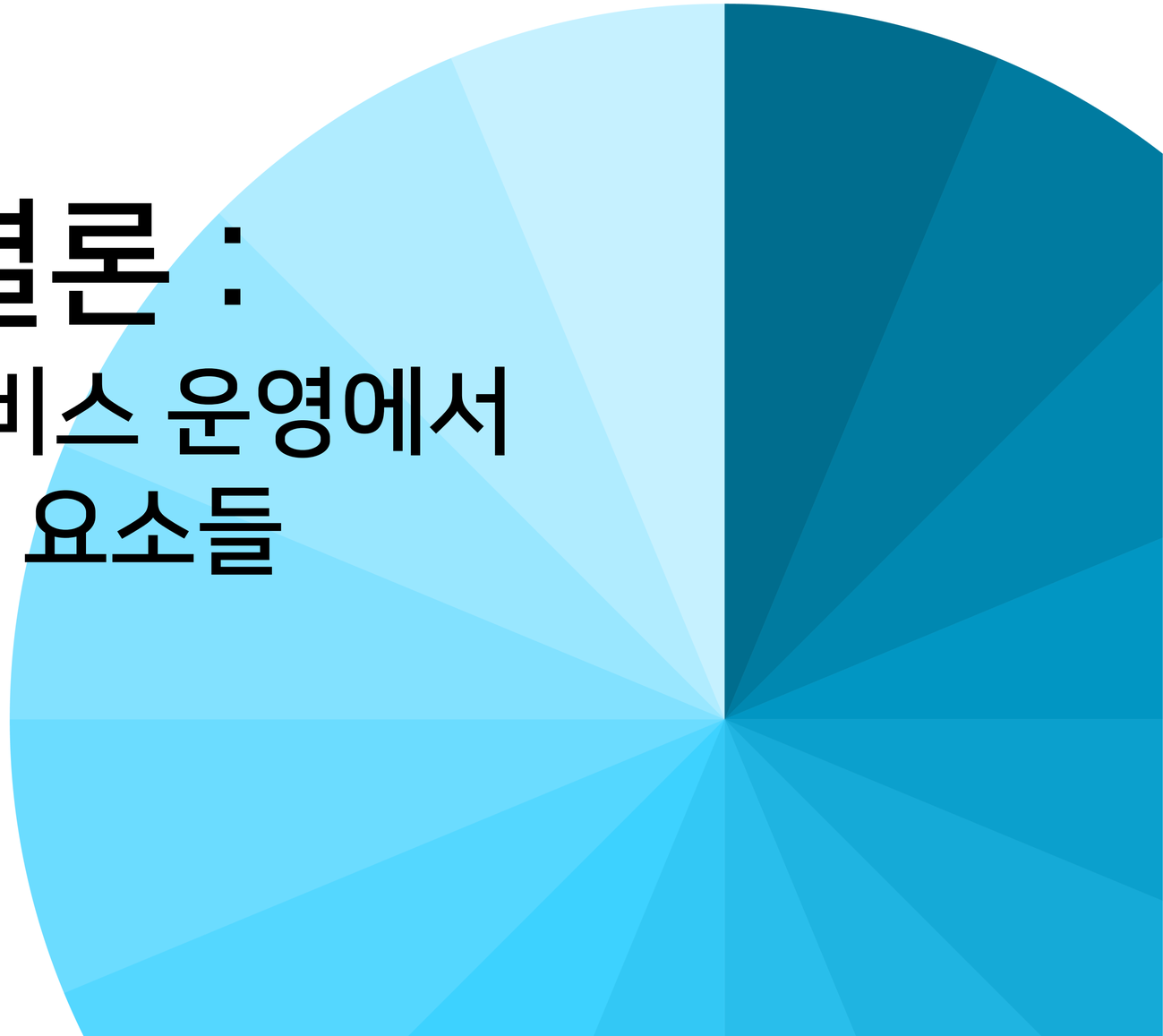
평가 일시 : 2020년 7월 17일 금요일 오전 10시 ~ 오후 3시
평가 차량 : 카니발 디젤 (초반 시내주행, 대부분은 고속도로 주행 상태에서 평가, 터널환경도 포함)
발화 대상 : 남 2명, 여 2명 (1인당 50발화씩),
발화 대본 : 네이버지도 서비스의 출현 POI 중 저빈도 음성인식 결과에서 랜덤샘플
발화 위치 : 차량 전면 중앙에 스마트폰 거치 후, 발화자는 보조석에서 발화 진행
평가 방법 : 발화자가 3개 앱의 마이크 버튼을 연달아 누른 후 발화를 진행
평가 단말 : iPhone X 3대

4. 실차 테스트를 통한 성능 확인



평가 일시 : 2020년 7월 17일 금요일 오전 10시 ~ 오후 3시
평가 차량 : 카니발 디젤 (초반 시내주행, 대부분은 고속도로 주행 상태에서 평가, 터널환경도 포함)
발화 대상 : 남 2명, 여 2명 (1인당 50발화씩),
발화 대본 : 네이버지도 서비스의 출현 POI 중 저빈도 음성인식 결과에서 랜덤샘플
발화 위치 : 차량 전면 중앙에 스마트폰 거치 후, 발화자는 보조석에서 발화 진행
평가 방법 : 발화자가 3개 앱의 마이크 버튼을 연달아 누른 후 발화를 진행
평가 단말 : iPhone X 3대

5. 결론 :
음성인식 서비스 운영에서 중요한 요소들



아무말 대잔치를 경계하라!

신기한 것 말고,
편리한 것을 추구하라

대상 도메인을 정확하게 정의하고 서비스를 기획하라

서비스 초기에 유입되는
운영 데이터를 금같이!

Thank You