DEVIEW 2020

누구나 만드는 내 목소리 합성기 II

커스텀 보이스 파이프라인

유정민 NAVER / NES

CONTENTS

1. 개인화 TTS 서비스

1.1 개인화 TTS란?

papago

1.1 개인화 TTS란?

Speech Synthesis
Text-to-Speech (TTS)

1.1 개인화 TTS란?

DEVIEW 2020

내 애인 목소리로
깨워주는 알람

아이유의 목소리로
읽어주는 나의 일정

엄마, 아빠의 목소리로
읽어주는 동화

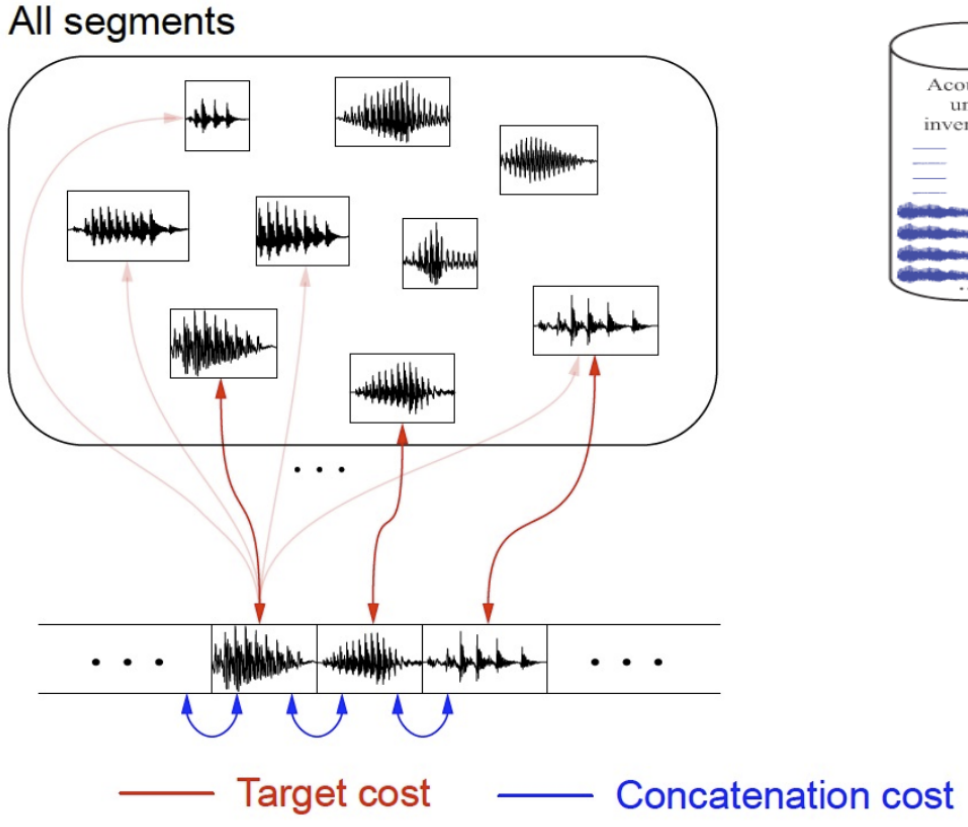
1.1 개인화 TTS란?

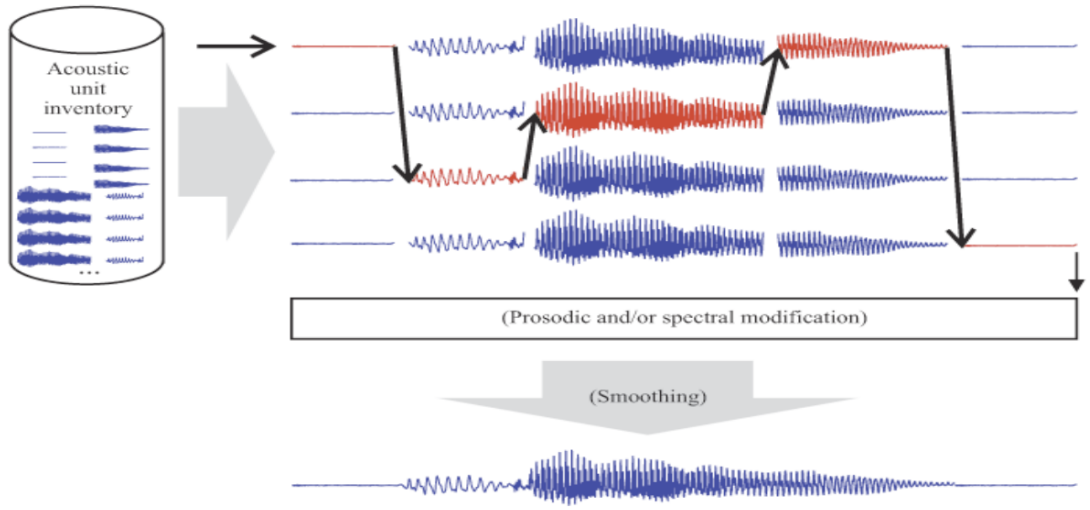
왜 없을까?

1.2 개인화 TTS가 어려운 이유

DEVIEW 2020

nVoice 합성기 - Concatenation Speech Synthesis





$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^{target}(t_i, u_i) + \sum_{i=2}^{n} C^{join}(u_{i-1}, u_i)$$

$$\hat{u}_1^n = \underset{u_1,\ldots,u_n}{\operatorname{argmin}} C(t_1^n, u_1^n)$$

1.2 개인화 TTS가 어려운 이유

nVoice 합성기 - Concatenation Speech Synthesis

정확한 발음

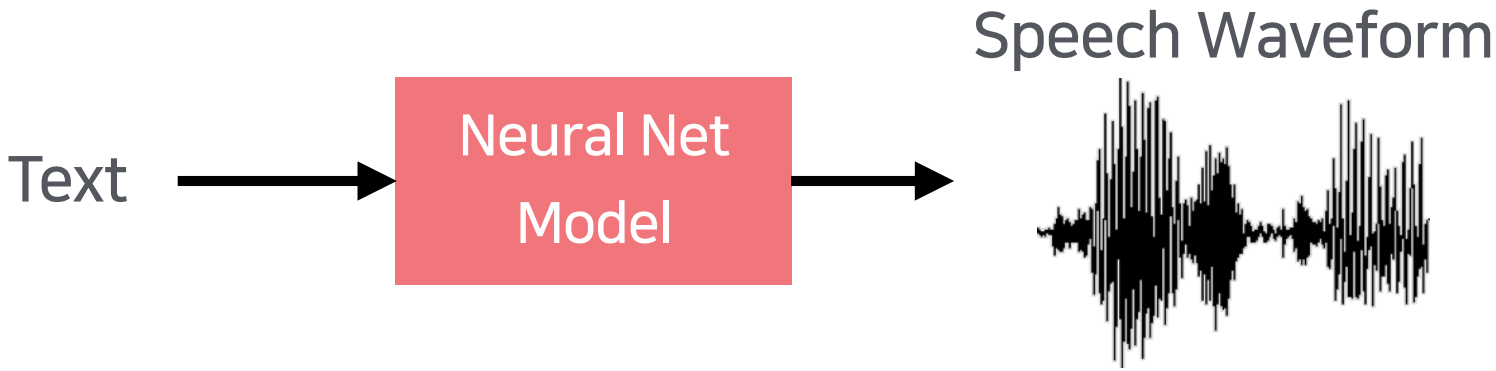
일관된 목소리
깨끗한 음질

정확한 전사

많은 녹음 분량

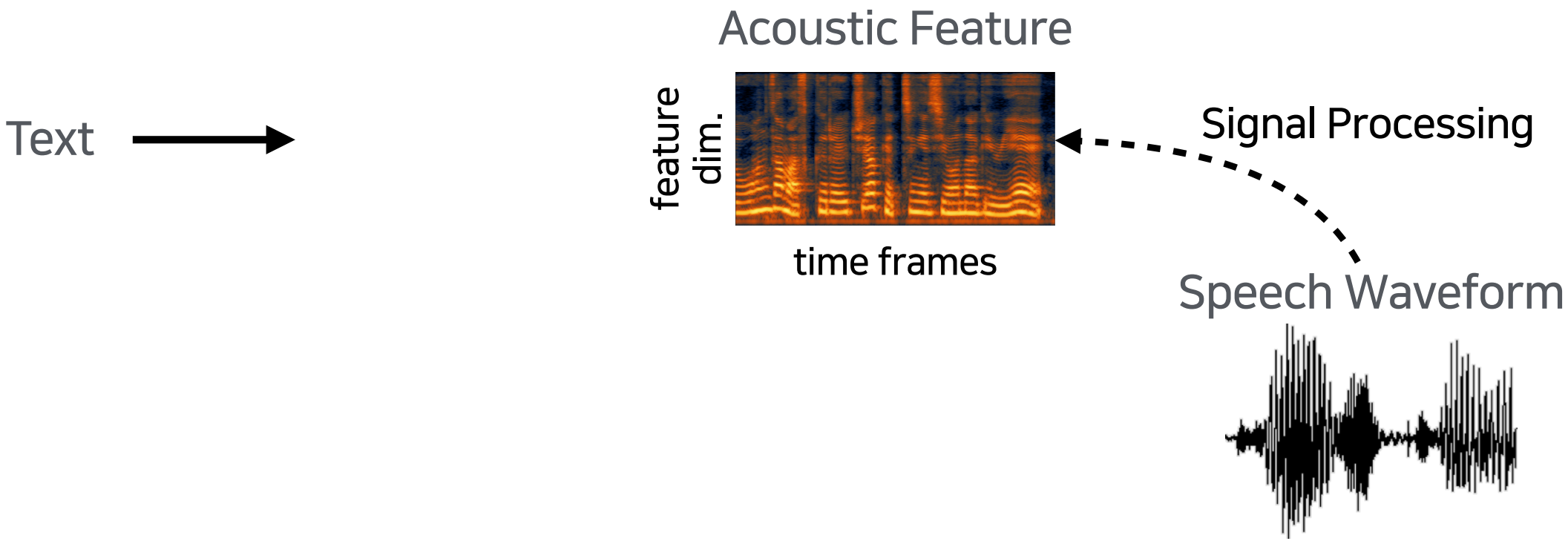
1.3 End-to-end 딥러닝으로 가능할까?

End-to-end 딥러닝



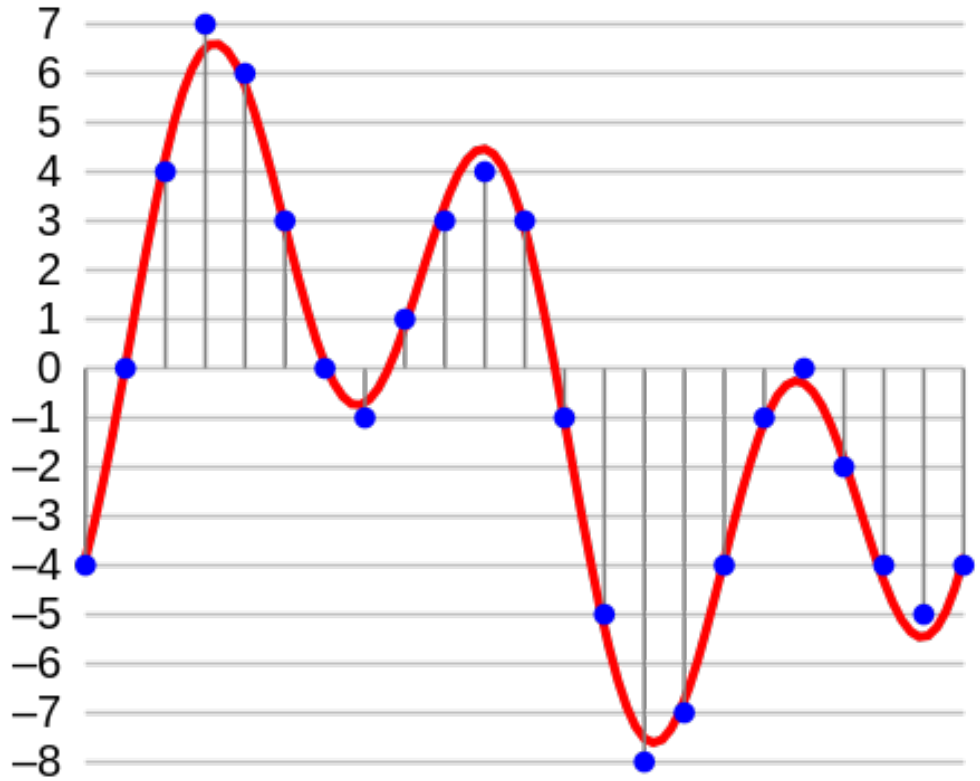
1.3 End-to-end 딥러닝으로 가능할까?

End-to-end 딥러닝



1.3 End-to-end 딥러닝으로 가능할까?

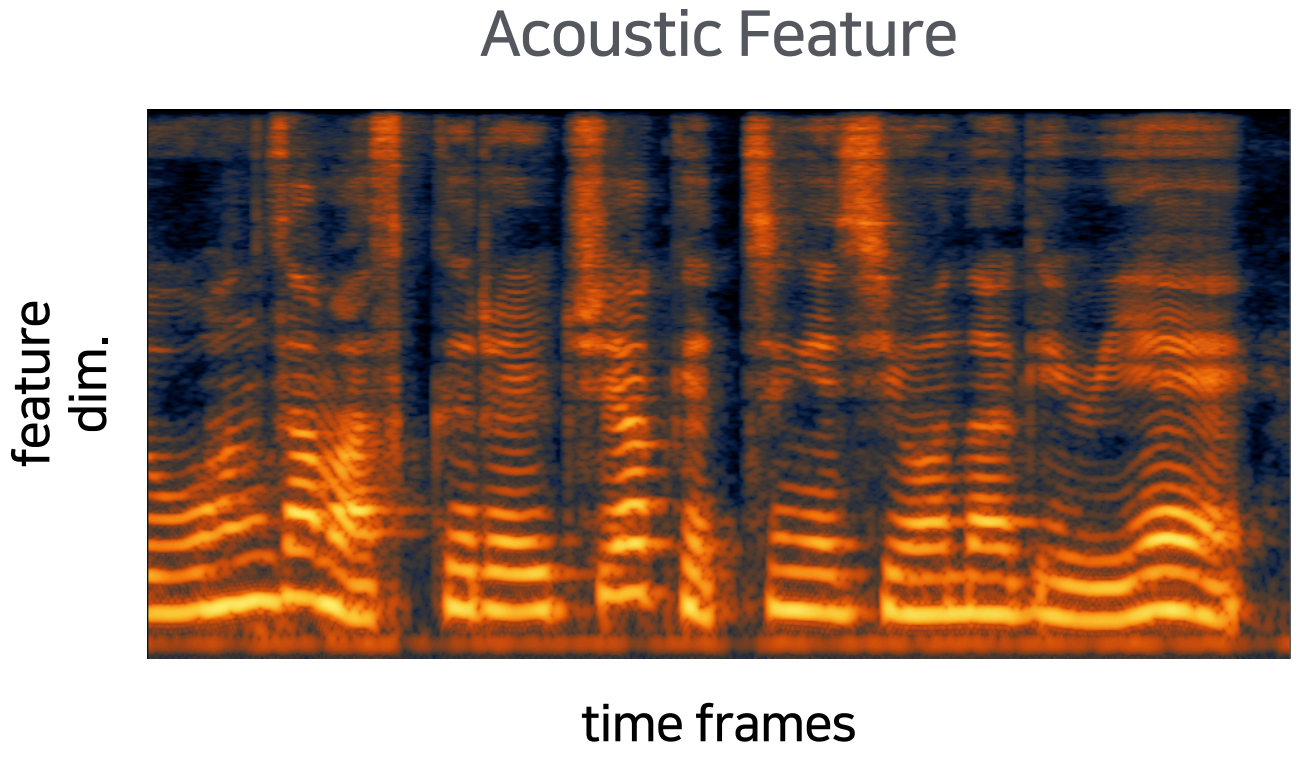
End-to-end 딥러닝



1.3 End-to-end 딥러닝으로 가능할까?

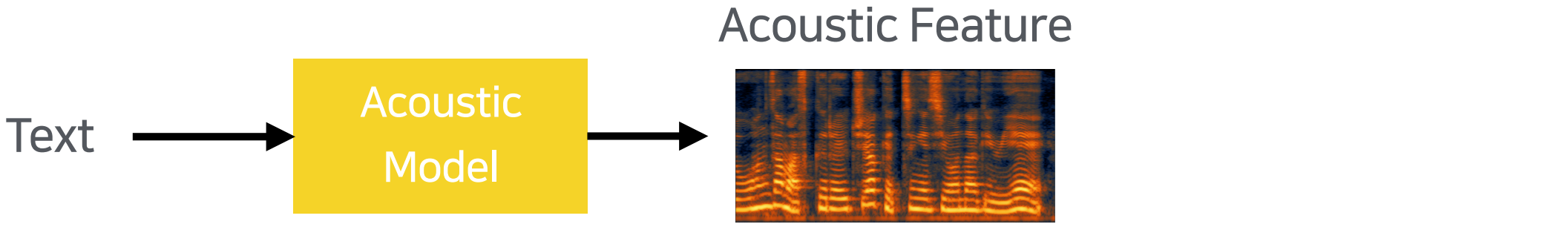
DEVIEW 2020

End-to-end 딥러닝

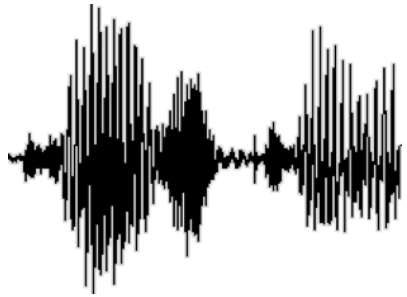


1.3 End-to-end 딥러닝으로 가능할까?

End-to-end 딥러닝

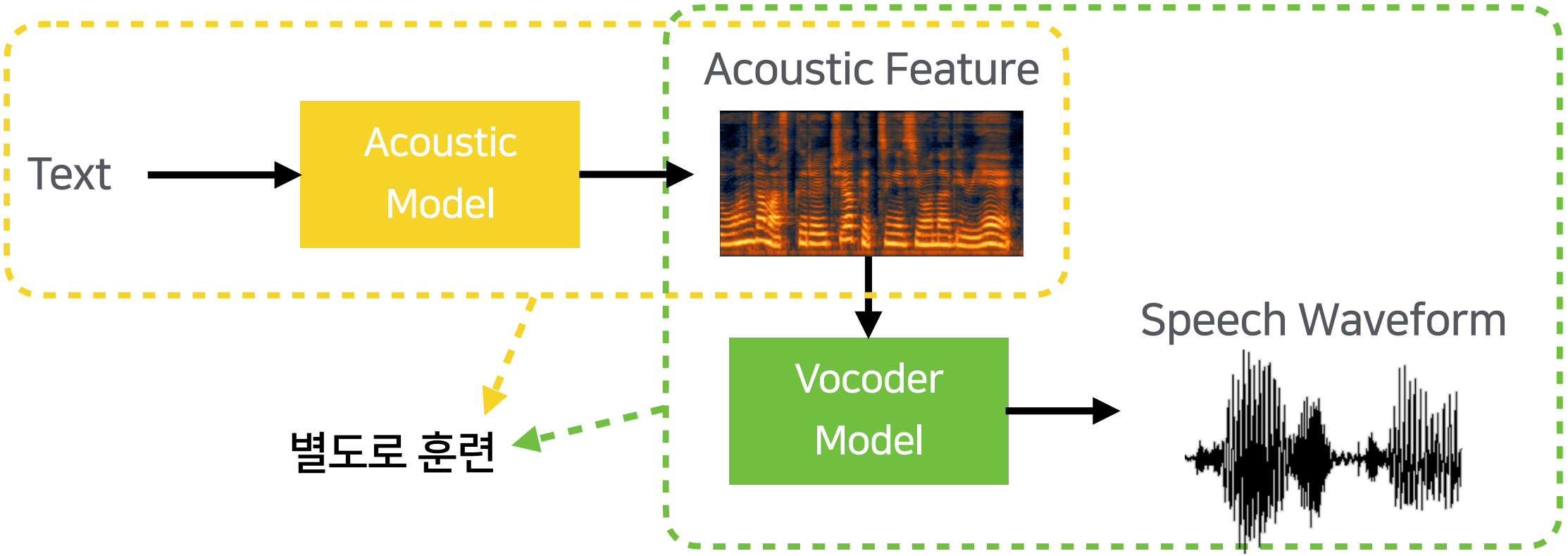


Speech Waveform



1.3 End-to-end 딥러닝으로 가능할까?

End-to-end 딥러닝



1.3 End-to-end 딥러닝으로 가능할까?

End-to-end 딥러닝

1. 전사작업 필요 없음
   - Acoustic Model이 어느 텍스트가 몇 분 몇 초의 음성에 대응되는지 학습함

2. Model Adaptation
   - 대량의 데이터를 가지고 있는 기존 nVoice 화자를 학습한 base model 생성
   - 소량의 개인 화자 데이터에 model adaptation
   - 개인 화자 데이터에는 발음이 조금 틀리거나 일관되지 않은 문장이 있어도 됨

1.3 End-to-end 딥러닝으로 가능할까?

DEVIEW 2020

nVoice 합성기 - Concatenation Speech Synthesis

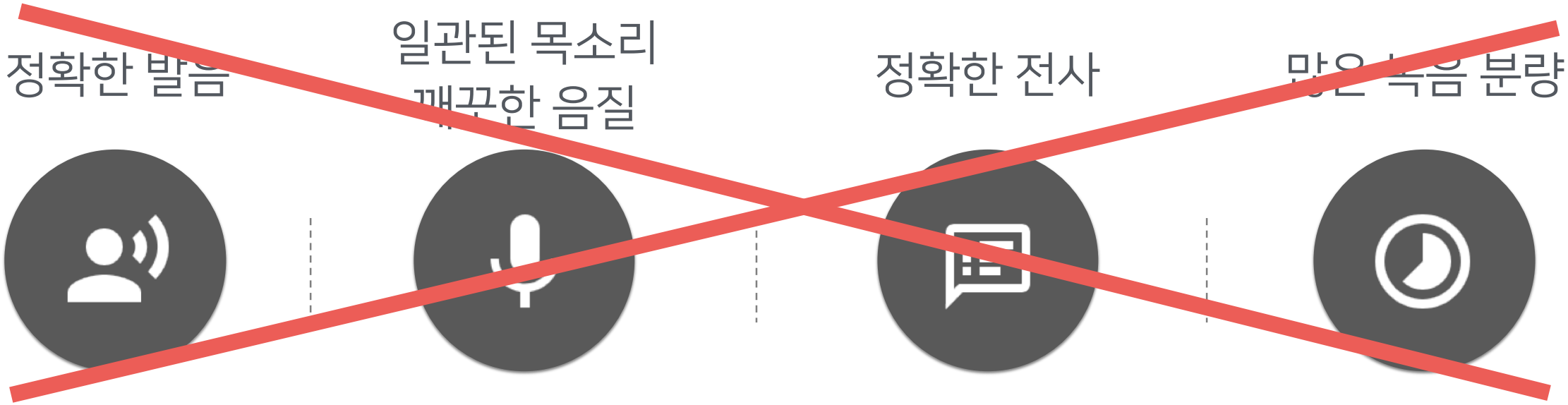
정확한 발음

일관된 목소리
깨끗한 음질

정확한 전사

많은 녹음 분량

1.3 End-to-end 딥러닝으로 가능할까?

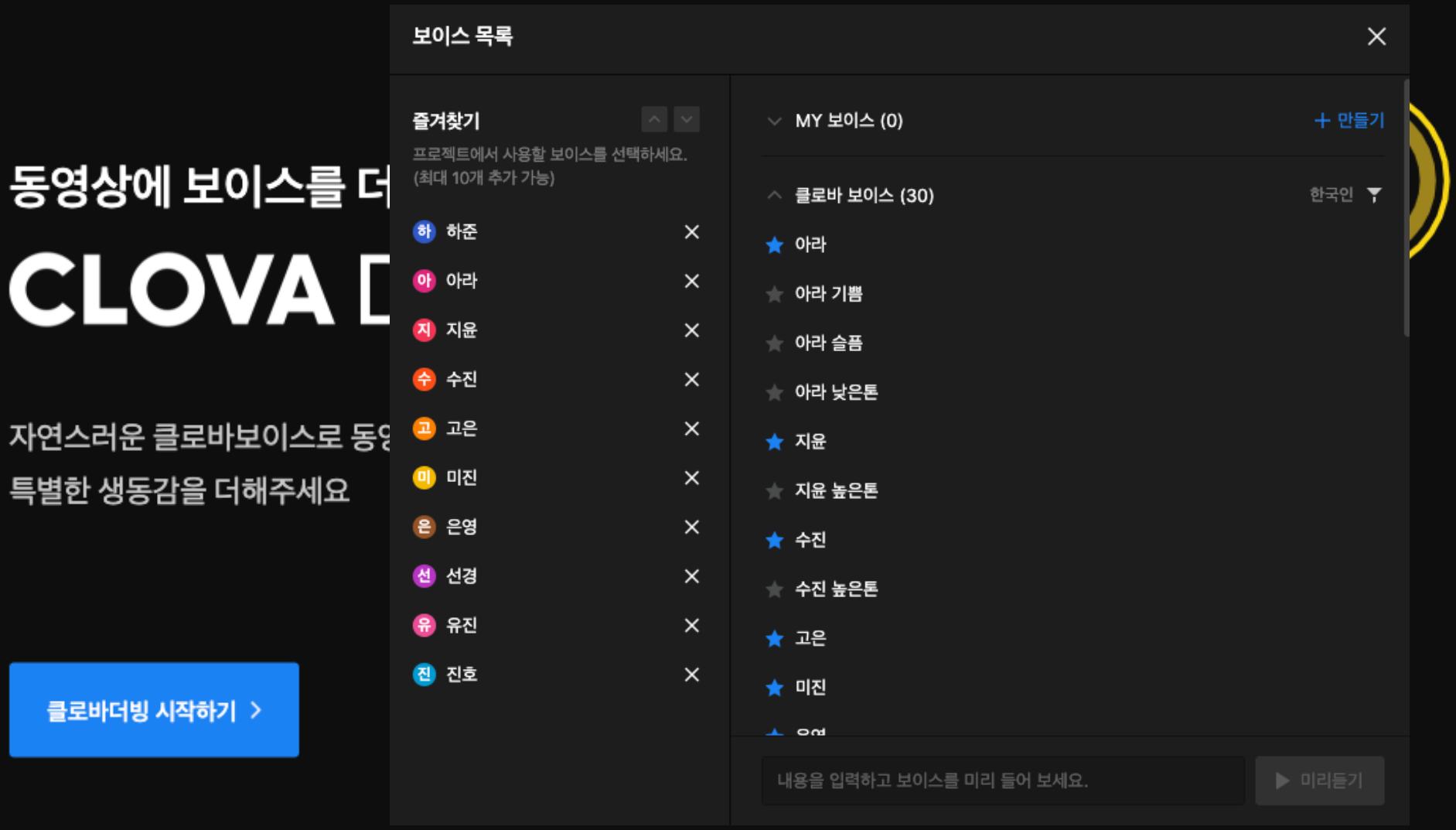


End-to-end 딥러닝으로 해결 가능!

2. End-to-end 음성합성기

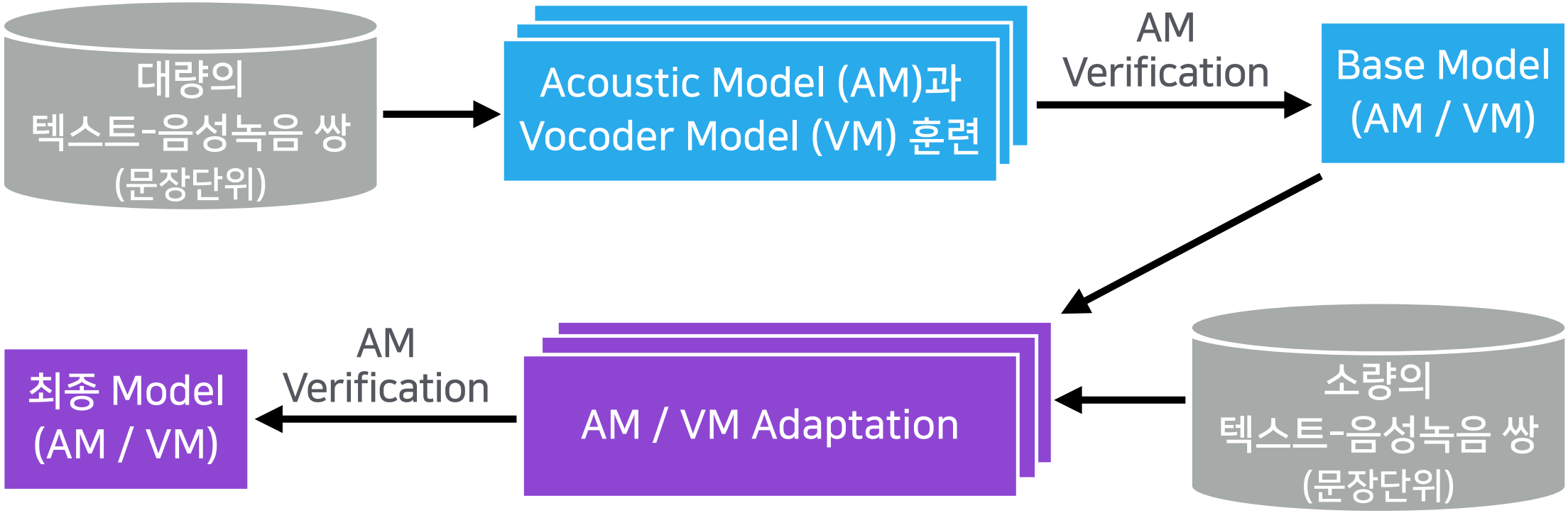
화자 18명 서비스하기

2.1 NES - Natural End-to-end Speech Synthesis System



2.1 NES - Natural End-to-end Speech Synthesis System

모델 제작 과정



2.2 Acoustic Model

Sequence to Sequence 모델

1. Autoregressive 방식
   - 시간 순서대로 Acoustic Feature를 추정하는 방식
   - Encoder (+ Attention) + Autoregressive Decoder
   - 예) Tacotron1, 2, Transformer-TTS

2.2 Acoustic Model

Sequence to Sequence 모델

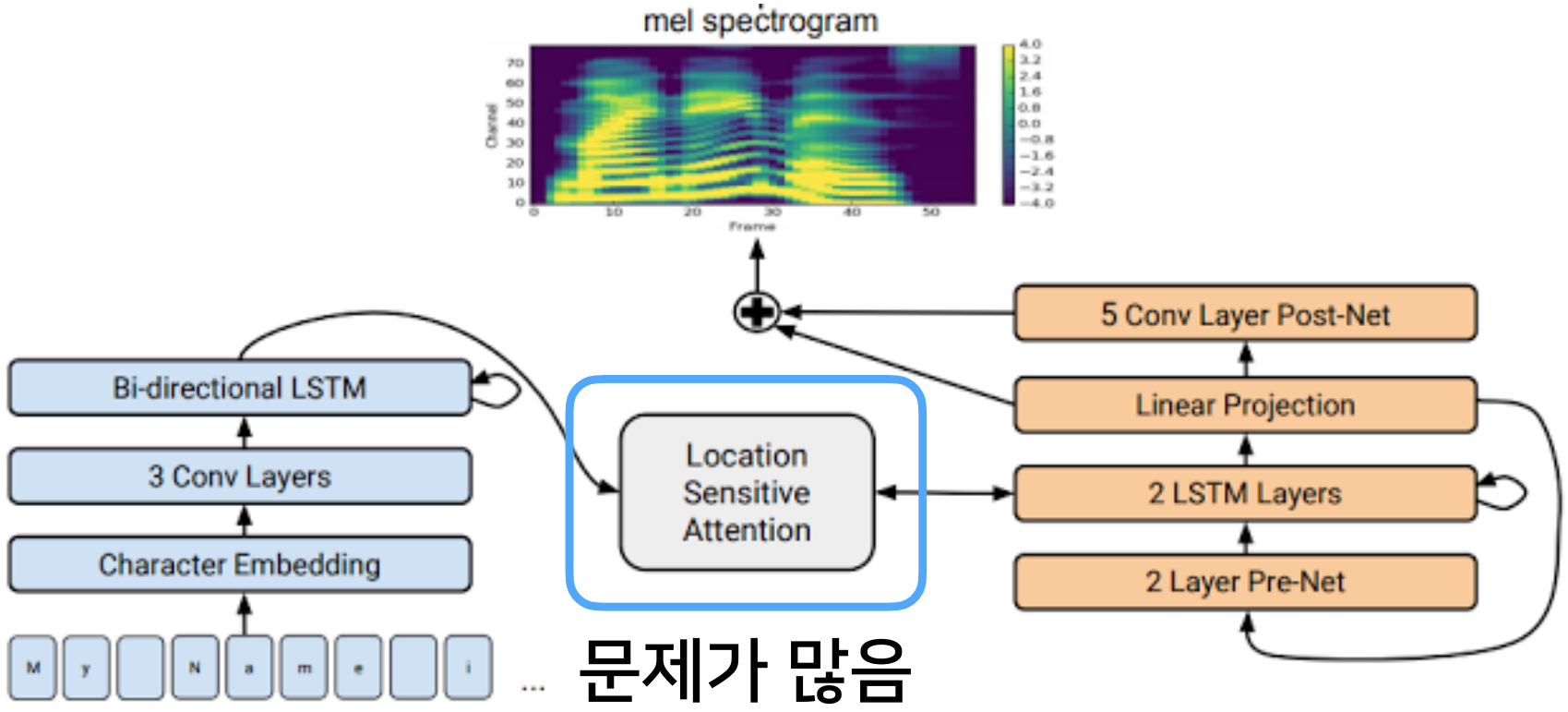
1. Autoregressive 방식
   - 시간 순서대로 Acoustic Feature를 추정하는 방식
   - Encoder (+ Attention) + Autoregressive Decoder
   - 예) Tacotron1, 2, Transformer-TTS

2. Non-autoregressive 방식
   - 시간 순서대로 추정할 필요가 없어 GPU를 활용한 병렬 연산 가능
   - Feed-Forward-Transformer / Flow-based Model
   - 예) FastSpeech, Flowtron, Glow-TTS, ...

2.2 Acoustic Model

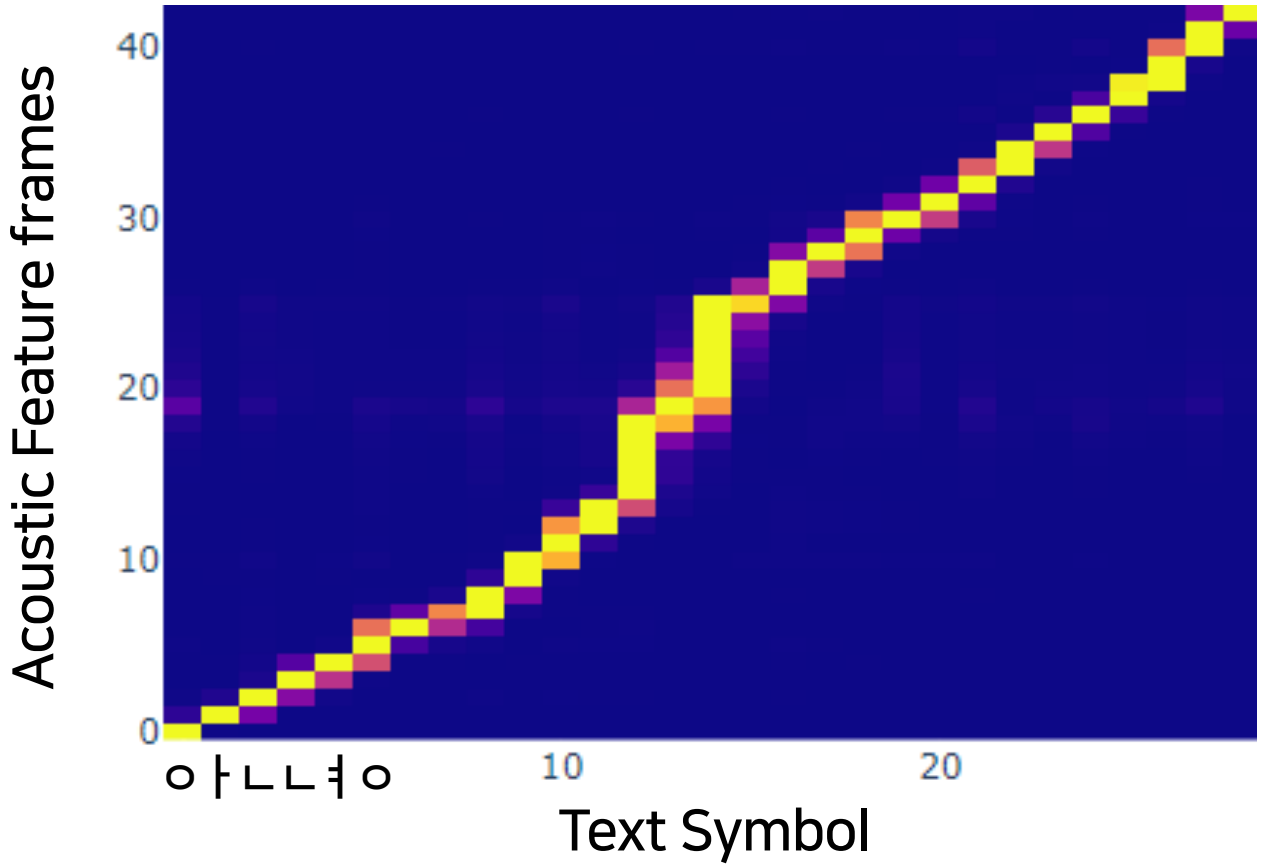
DEVIEW 2020

Tacotron2



2.2 Acoustic Model

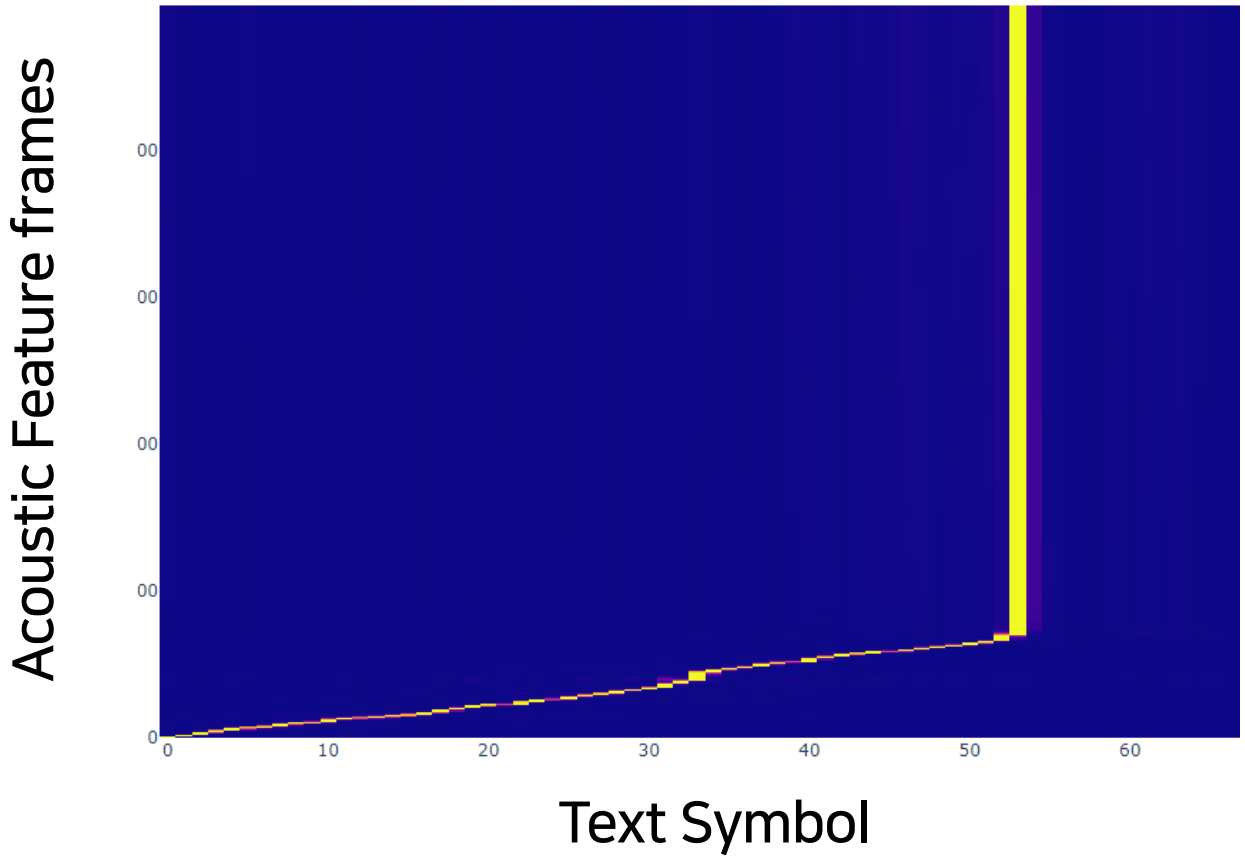
Attention Module - 성공한 경우



2.2 Acoustic Model

DEVIEW 2020

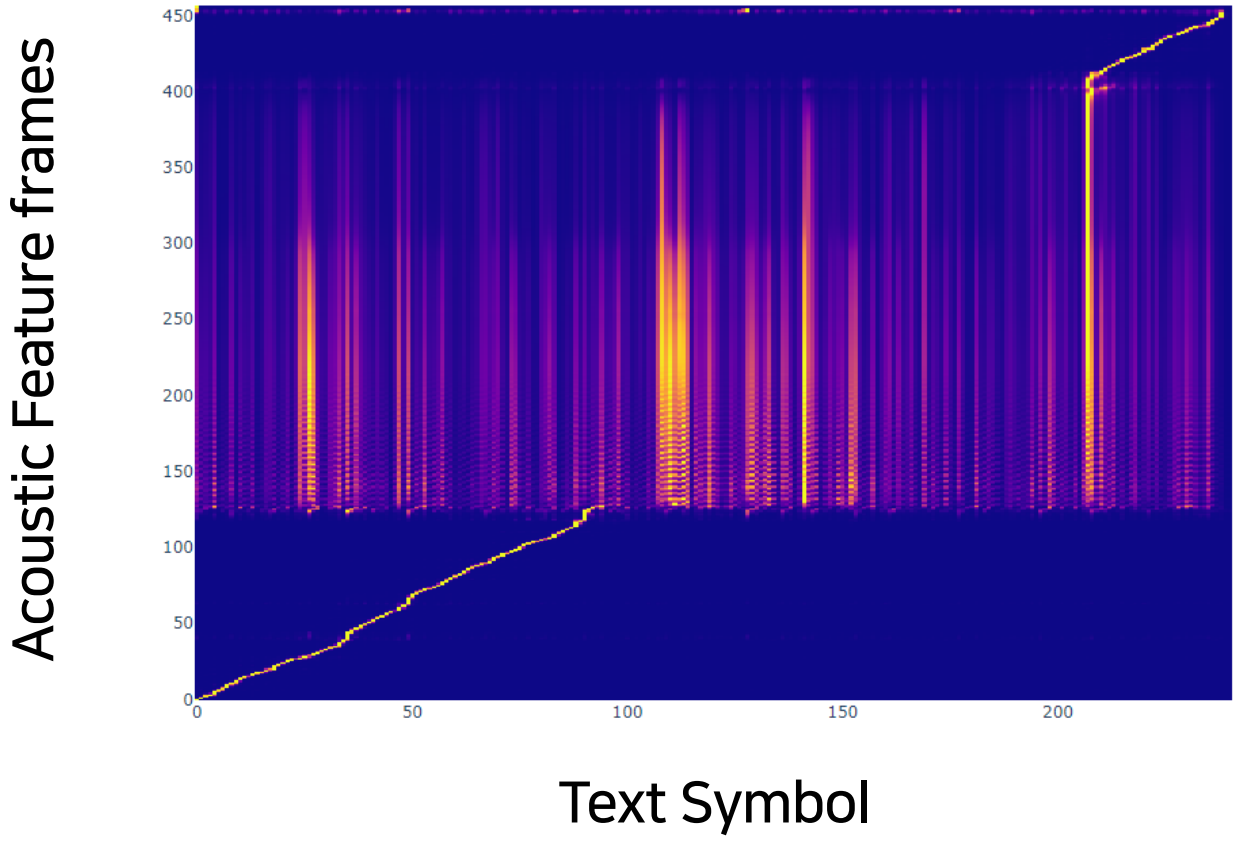
Attention Module - 실패한 경우



2.2 Acoustic Model

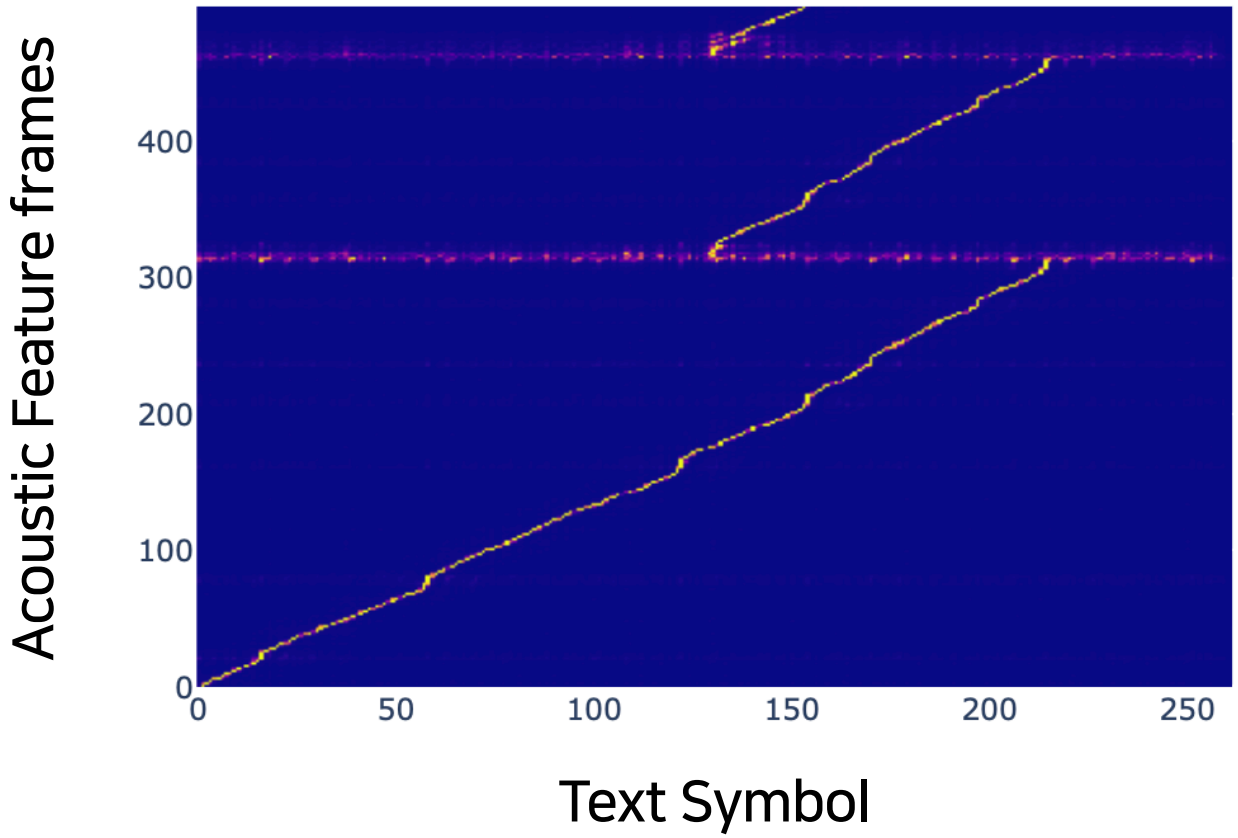
DEVIEW 2020

Attention Module - 실패한 경우



2.2 Acoustic Model

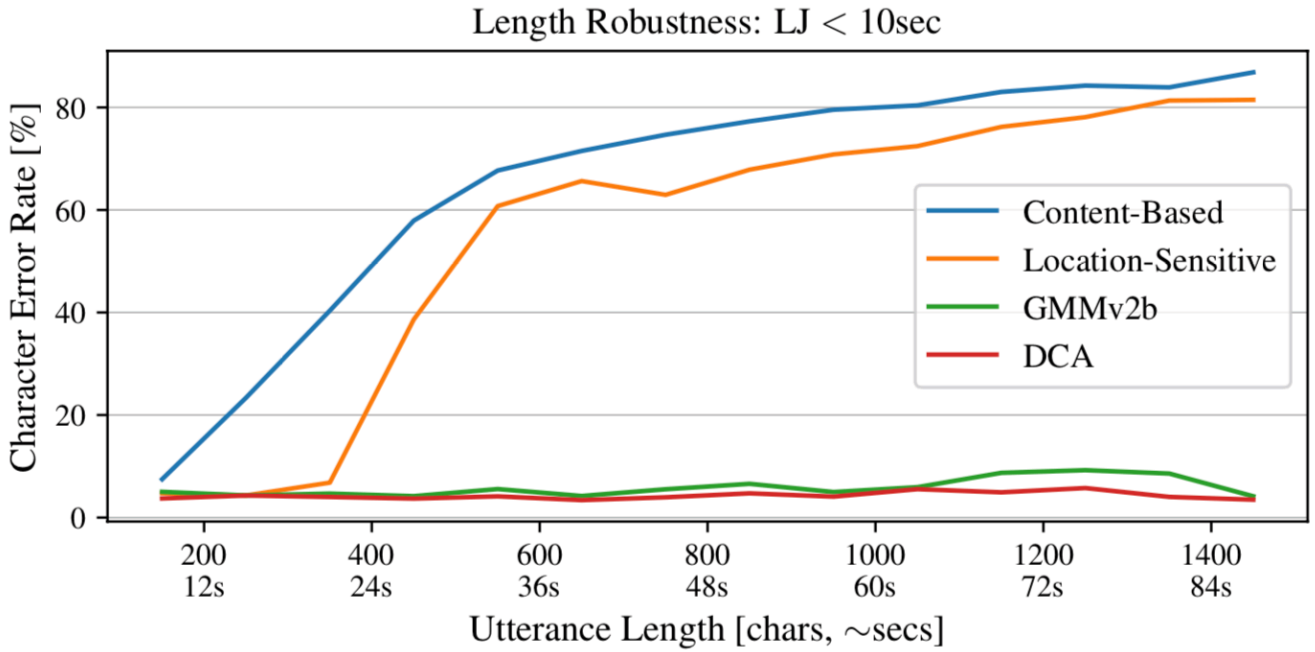
Attention Module - 실패한 경우



2.2 Acoustic Model

DEVIEW 2020

Attention Module



Tacotron1
Tacotron2

Fig. 3. Utterance length robustness for models trained on the Lessac (top) and LJ (bottom) datasets.

Battenberg, Eric, et al. "Location-relative attention mechanisms for robust long-form speech synthesis." ICASSP 2020.

2.2 Acoustic Model

DEVIEW 2020

Attention Module - 실패 방지 방법

1. 새로운 기법: dynamic convolution attention*,
stepwise monotonic attention**, ...

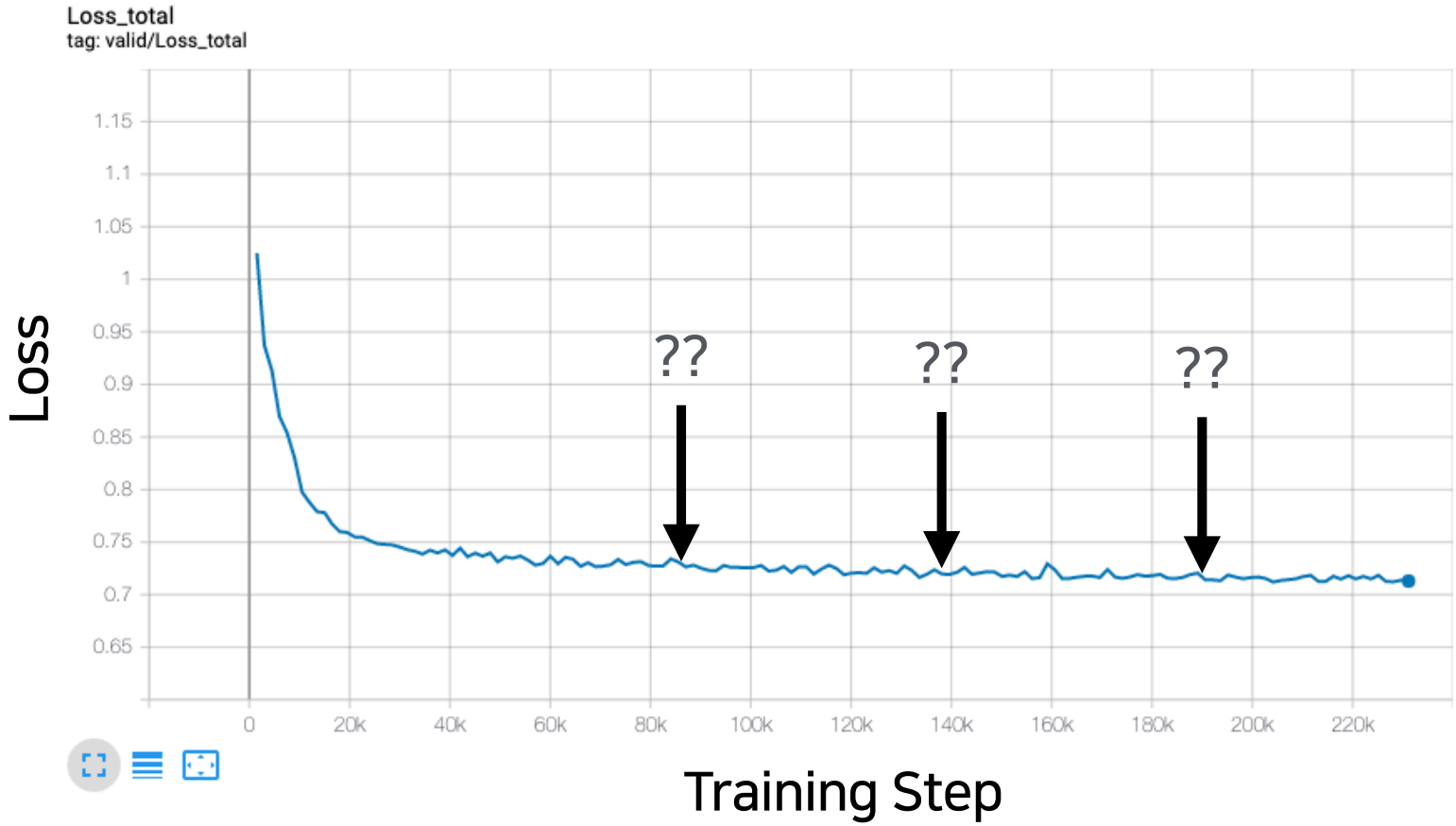
2. Attention 오류 상황 감지하여 정상화

* Battenberg, Eric, et al. "Location-relative attention mechanisms for robust long-form speech synthesis." ICASSP 2020.

** He, Mutian, Yan Deng, and Lei He. "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural TTS." arXiv preprint arXiv:1906.00672 (2019).

2.3 Acoustic Model Verification

음성합성 모델의 성능을 평가하는 기준?



2.3 Acoustic Model Verification

DEVIEW 2020

음성합성 모델의 성능을 평가하는 기준?

1. 수천~수만 문장에 대해 Attention 실패율 체크
   - 어떤 심볼, 어떤 문장에서 실패했는지 판단하는 알고리즘

2. Loss가 작은 모델 중 Attention 실패율이 가장 적은 모델(스텝) 3~10개 선정
   - 클로바더빙에 출시된 화자들의 attention 실패율: **0.03 % 미만**

3. 그 중 사람이 들었을 때 가장 발음이 명확하고 운율이 자연스러운 모델 선정

2.4 Vocoder Model

1. 신호처리를 사용한 전통적 Vocoder
   - WORLD, STRAIGHT, ...
   - 기계음 같은 느낌
   - Acoustic Model 출력의 미세한 오류에 민감하게 반응
   - 학습 과정이 없고 속도가 매우 빠름

2. Neural Vocoder
   - 실제 녹음과 구분하기 힘든 수준의 음질
   - 학습 및 합성 속도는 상대적으로 느림

2.4 Vocoder Model

Neural Vocoder

1. Autoregressive Model
   - WaveNet, WaveRNN, ...
   - WaveNet 계열이 음질은 전체 1위
   - 1초짜리 음성 합성하는 데에 수 분 소요

2. Signal Processing Hybrid
   - LPCNet, FeatherWave, NSF, ...
   - CPU에서도 합성 속도가 빠름

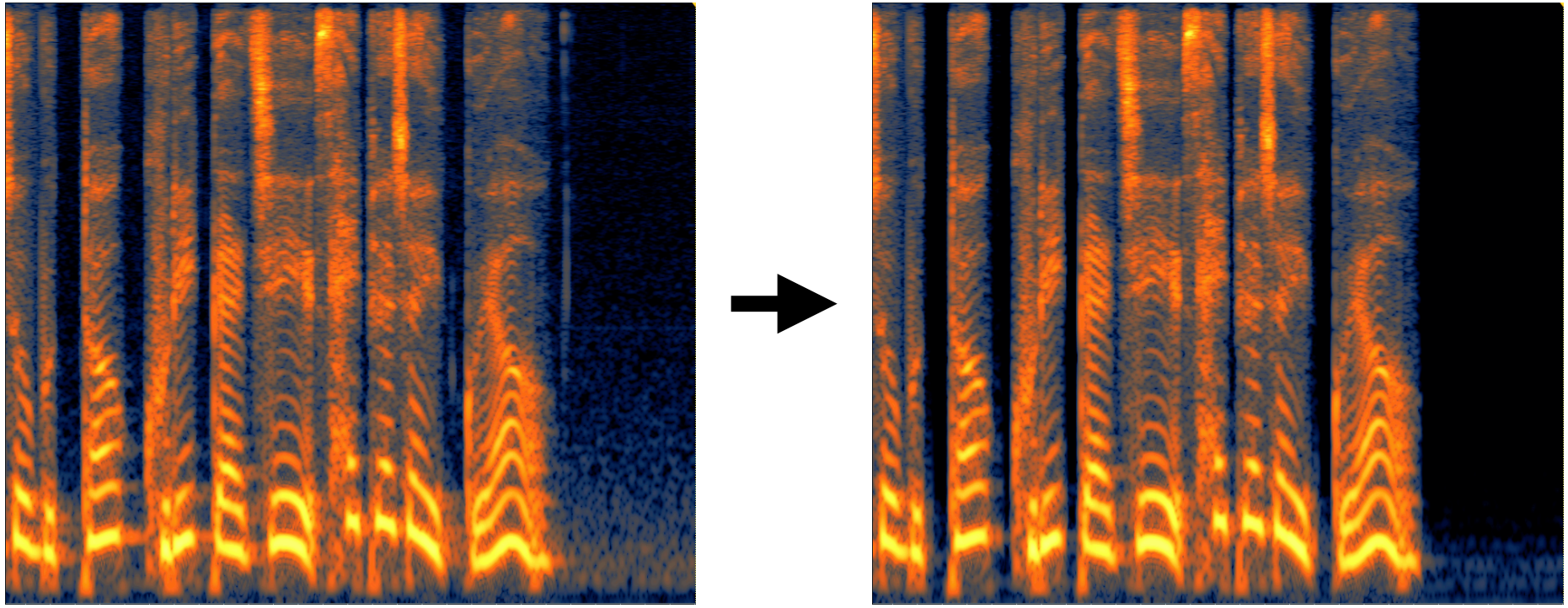
3. GAN-based Model
   - MelGAN, Parallel WaveGAN, ...
   - 품질이 안정적이지 않음

4. Flow-based Model
   - Parallel WaveNet, WaveGlow, ...
   - GPU-friendly

품질과 합성속도 사이의 trade-off!

2.5 합성음 품질 보강

Vocoder 모델에 따라 적절한 신호처리 필요

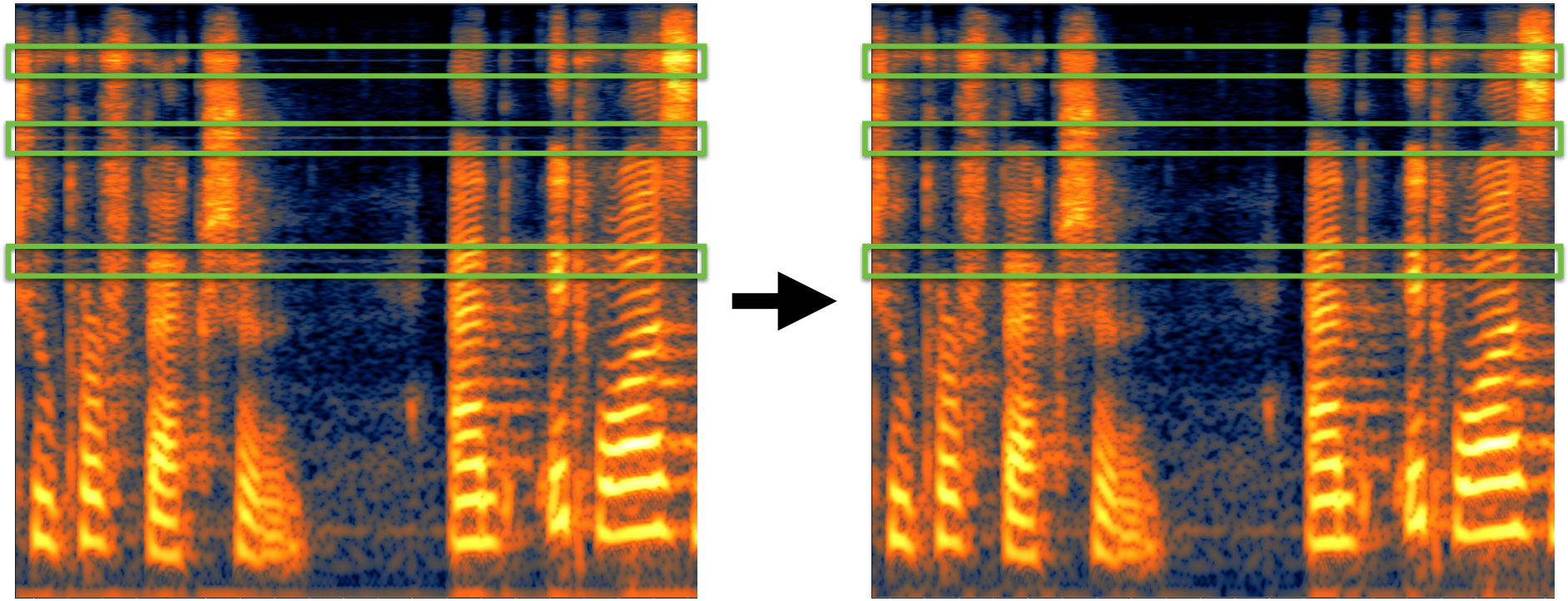


합성음 (quantization noise)

denoising filter 적용

2.5 합성음 품질 보강

Vocoder 모델에 따라 적절한 신호처리 필요



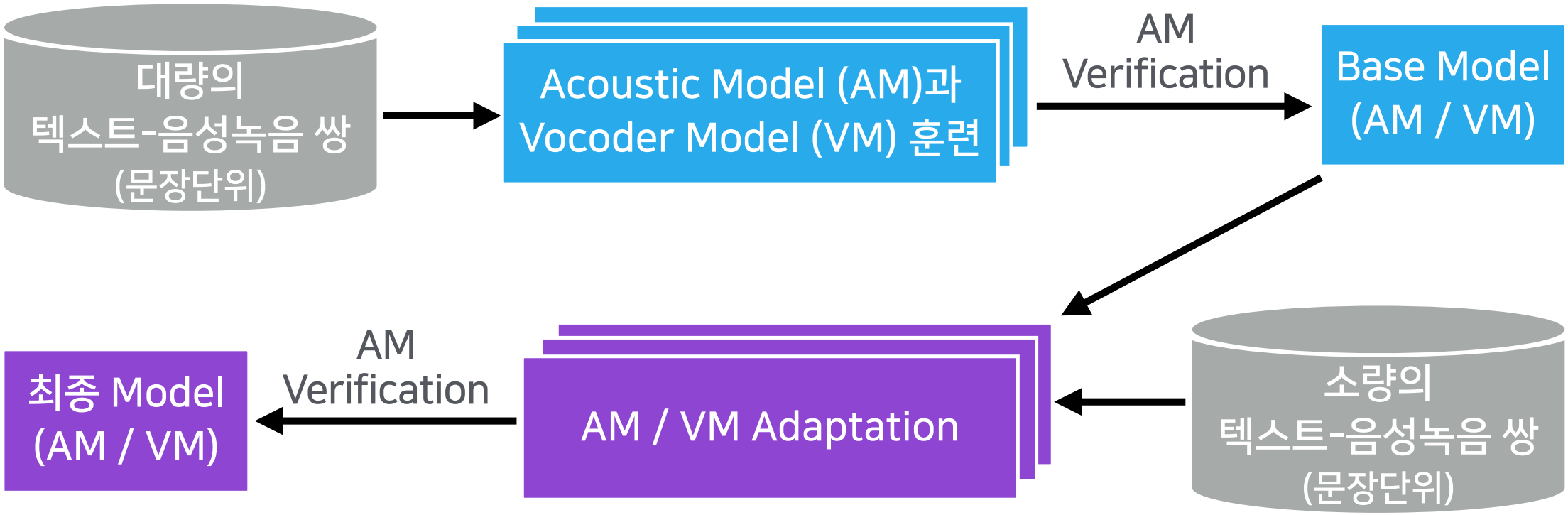
합성음 (harmonic noise)

harmonic noise 제거

3. 개인화 서비스로 나아가기

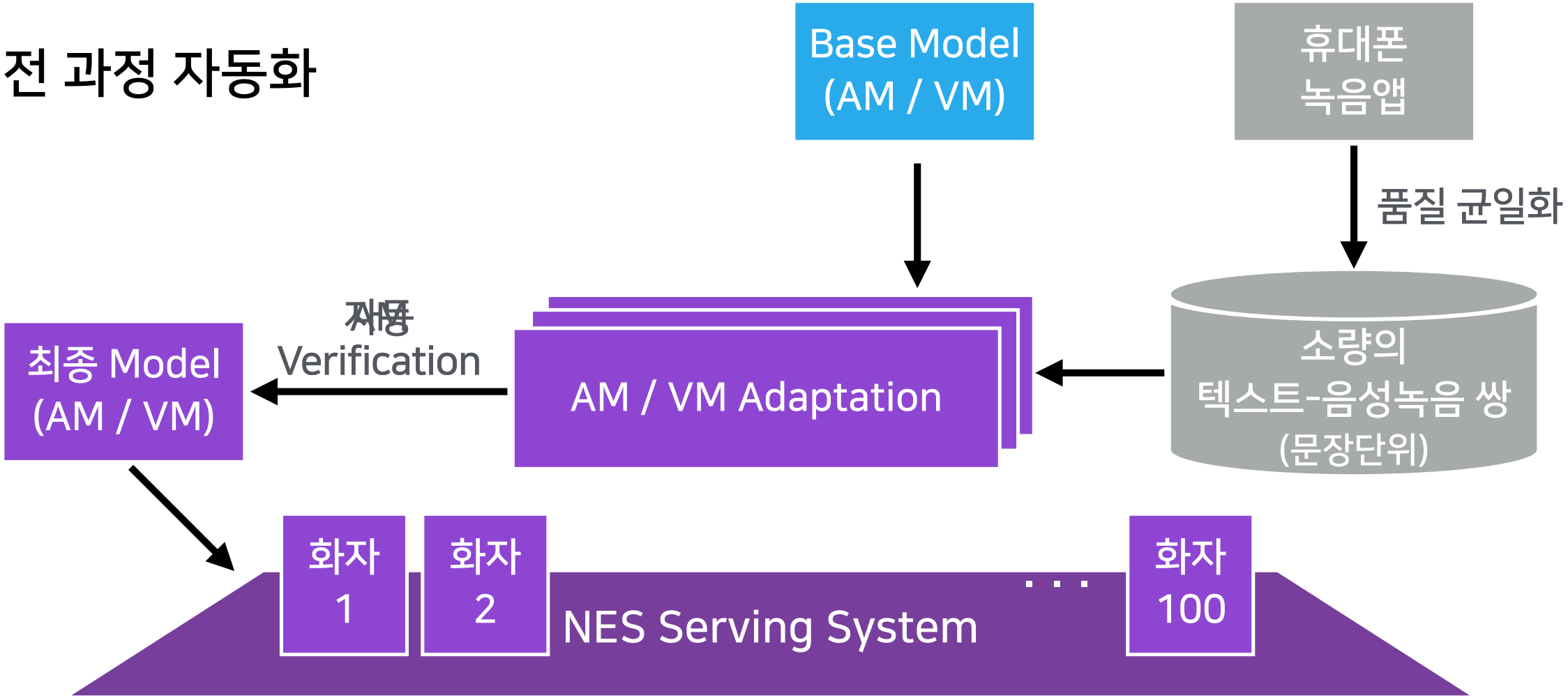
커스텀 보이스 파이프라인

2.1 NES - Natural End-to-end Speech Synthesis System



3.1 커스텀 보이스 파이프라인

전 과정 자동화



3.2 개인이 녹음할 수 있는 분량과 품질

DEVIEW 2020

녹음 스크립트 선별

- 베이스 모델용 화자의 문장 수: 수 만 문장

- 현재 클로바더빙 화자들의 문장 수: 3000 문장

- 개인화 서비스용으로 몇 문장까지 줄일 수 있을까?

200 ~ 400 문장 (약 20~40분)!

3.2 개인이 녹음할 수 있는 분량과 품질

DEVIEW 2020

녹음본 품질

베이스 모델용 화자
클로바더빙용 화자

입-마이크 거리와 각도 계속 바뀜
냉장고, 에어컨 소리
의자 움직이는 소리
기타 생활 소음
...

개인화 서비스용

3.2 개인이 녹음할 수 있는 분량과 품질

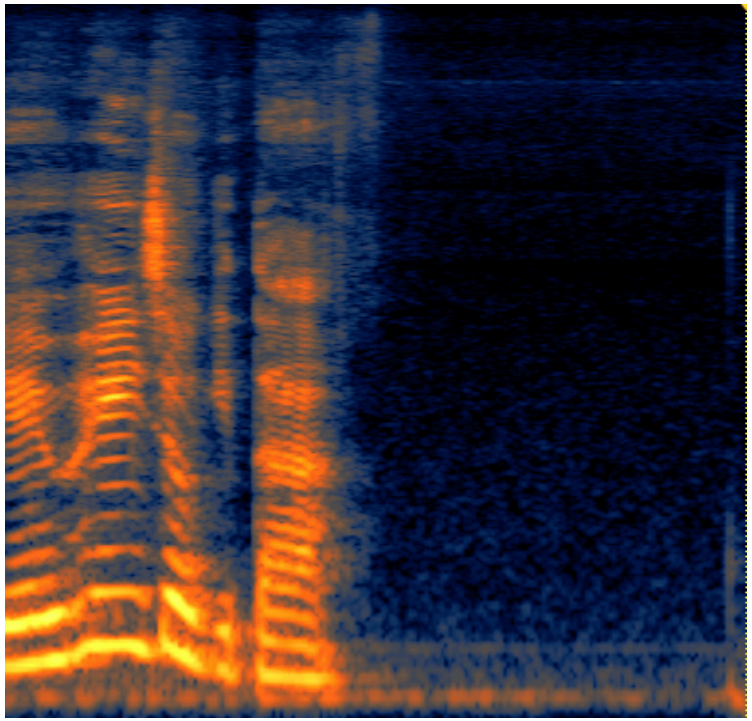
녹음본 품질 균일화

1. 노이즈 제거

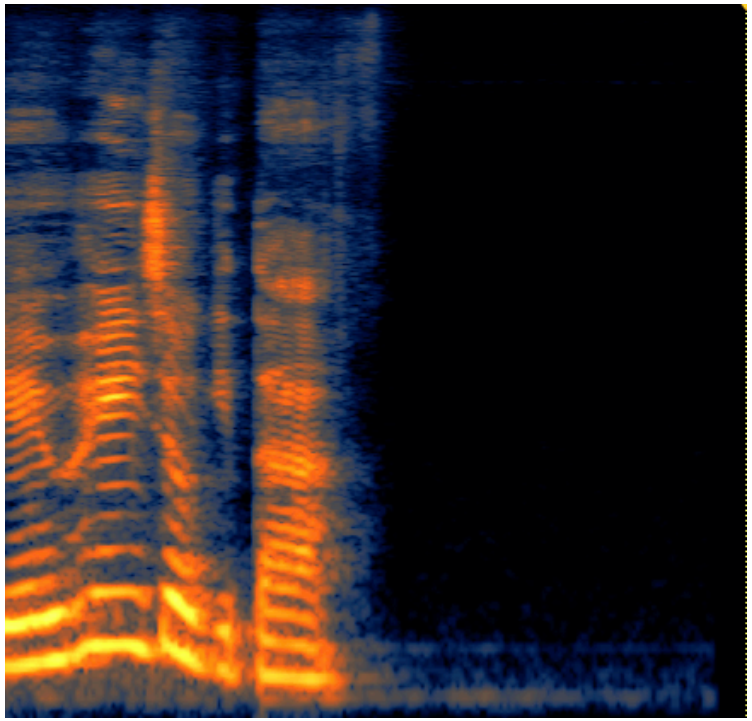
3.2 개인이 녹음할 수 있는 분량과 품질

DEVIEW 2020

녹음본 품질 균일화



생활 소음이 있는 핸드폰 녹음본



노이즈 제거

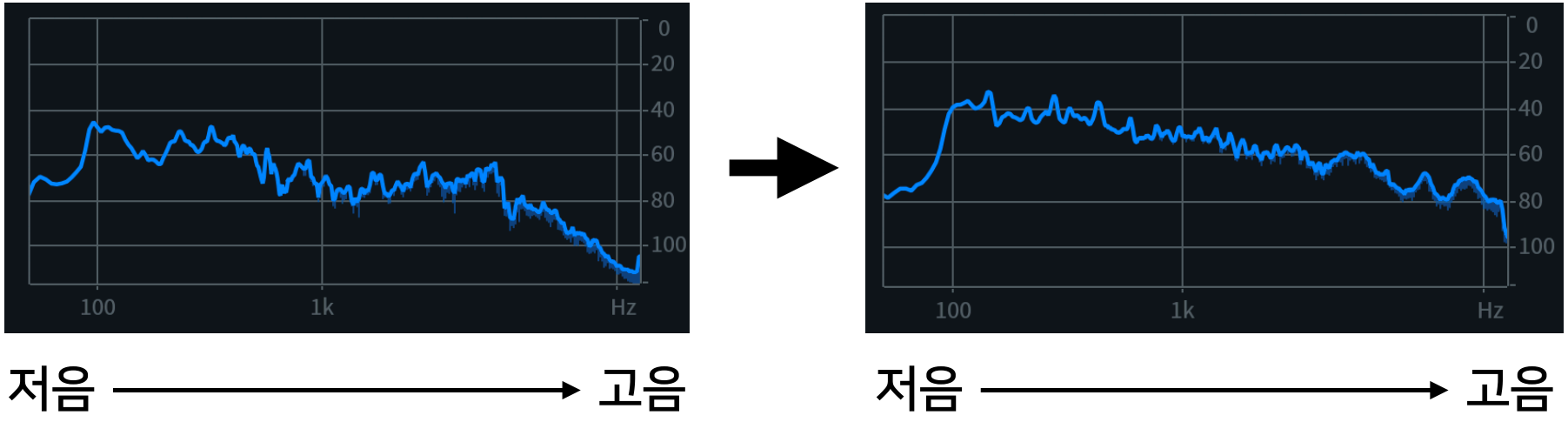
3.2 개인이 녹음할 수 있는 분량과 품질

DEVIEW 2020

녹음본 품질 균일화

1. 노이즈 제거

2. 문장마다 음색 맞추기 (match EQ)



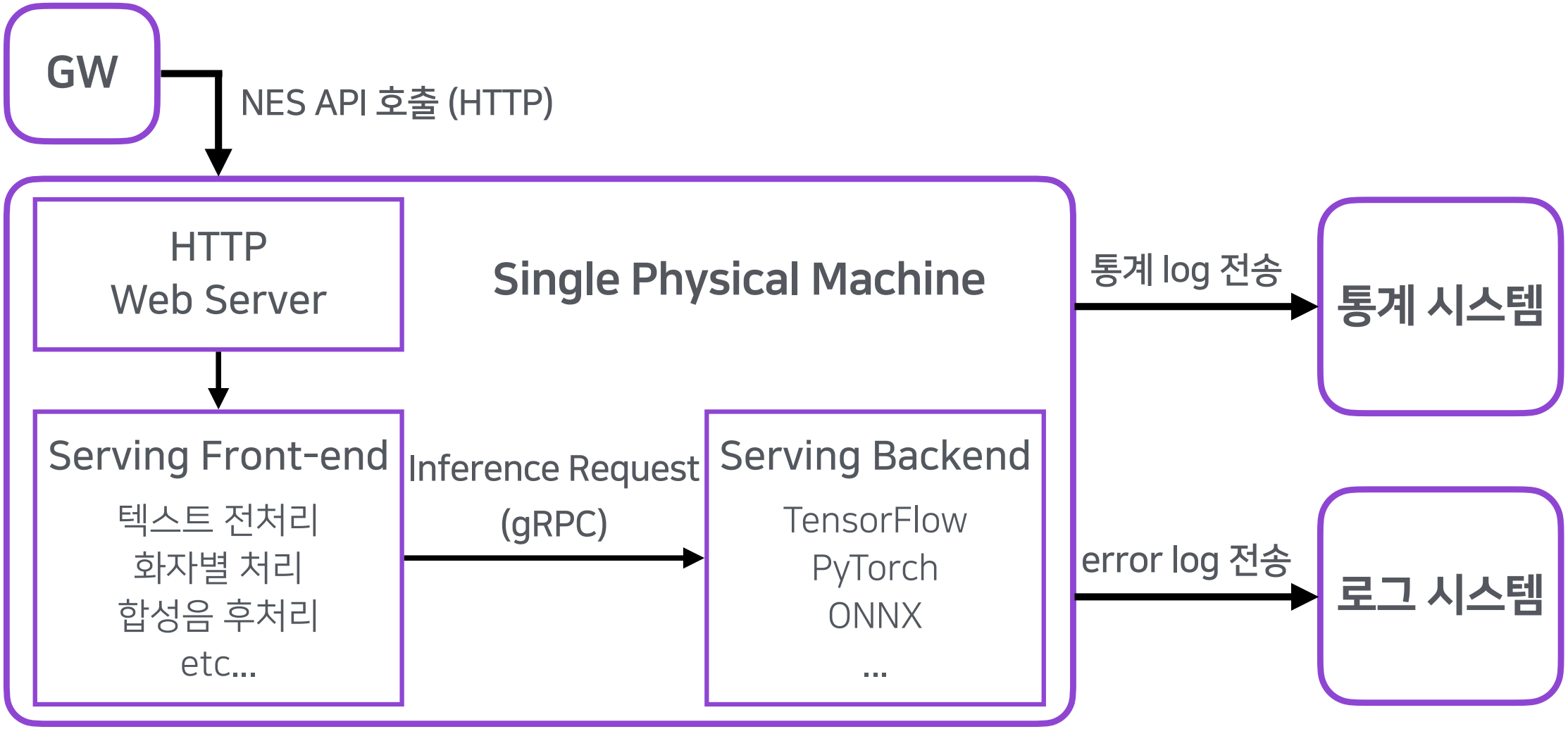
3.2 Acoustic Model Verification 자동화

Loss와 Attention 실패율 모두 작은 모델 자동 선정

- 기존 화자들보다 문장 수가 적어 금방 overfitting이 시작됨
  -> 품질 좋은 스텝 구간이 굉장히 짧음
  -> 그 중 하나 고르는 것은 loss와 attention 실패율만 보고 자동화 가능

- 기존 화자들 학습했을 때의 경험을 기반으로 적절한 criterion 만들기

3.2 딥러닝 모델 서빙


GW
NES API 호출 (HTTP)
HTTP
Web Server
Single Physical Machine
Serving Front-end
텍스트 전처리
화자별 처리
합성음 후처리
etc...
Inference Request
(gRPC)
Serving Backend
TensorFlow
PyTorch
ONNX
...
통계 log 전송
통계 시스템
error log 전송
로그 시스템

3.2 딥러닝 모델 서빙

NES Serving System (NSS)

- 한 서버에서 모델 100개(**화자 100명**)까지 서빙 가능

- 다양한 Machine Learning Framework 지원

- 화자별 설정 관리

#세상에하나뿐인
#내이야기가담긴
#AI보이스

당신만의 이야기가 담긴 목소리를 찾습니다.

당신의 목소리가 클로바의 기술을 만나 세상에 하나밖에 없는 AI 보이스로 탄생합니다.
당신만의 이야기가 담긴 목소리, 네이버 클로바에게 보내주세요.

4. 개인화 TTS 시연

개인화 TTS 시연

DEVIEW 2020

2018년 DEVIEW 버전 합성기
(이봉준님 130문장-10분 데이터)

몰텐과 돈핀도 부부가 되어 새끼를 낳았습니다.

최신 버전 합성기
(이봉준님 130문장-10분 데이터)

몰텐과 돈핀도 부부가 되어 새끼를 낳았습니다.

핸드폰 녹음
(유정민)

몰텐과 돈핀도 부부가 되어 새끼를 낳았습니다.

최신 버전 합성기
(유정민 270문장-20분 데이터)

몰텐과 돈핀도 부부가 되어 새끼를 낳았습니다.

개인화 TTS 시연

DEVIEW 2020

2018년 DEVIEW 버전 합성기
(이봉준님 130문장-10분 데이터)

아무리 중요하다고 해도 한 가지만 보고 채용을 결정할 수 있을까요?

최신 버전 합성기
(이봉준님 130문장-10분 데이터)

아무리 중요하다고 해도 한 가지만 보고 채용을 결정할 수 있을까요?

핸드폰 녹음
(유정민)

아무리 중요하다고 해도 한 가지만 보고 채용을 결정할 수 있을까요?

최신 버전 합성기
(유정민 270문장-20분 데이터)

아무리 중요하다고 해도 한 가지만 보고 채용을 결정할 수 있을까요?

5. 앞으로 할 일

앞으로 할 일

1. 감정, 속도 조절
   - 현재는 각 화자마다 감정을 잘 표현하여 녹음한 데이터가 필요
     - 사람의 감정표현 자체를 학습하는 모델?
   - 현재는 합성음에 후처리로 속도 조절 -> 음질 저하
     - 음성합성 모델 단에서 속도 조절 가능하도록 하려면?

2. 더 쉽고 빠른 개인화
   - 목표: 100 문장 이하 (10분) -> 녹음 하는 데에는 1시간 정도 걸리도록
   - 클라이언트에서 작동하는 가벼운 모델?

3. 품질 향상, 속도 향상
   - 새로운 모델
   - 전처리, 후처리 기술 고도화

We are hiring!

DEVIEW 2020

https://recruit.navercorp.com/naver/job/list/developer?searchSysComCd=&entTypeCd=&searchTxt=clova

인턴 / 신입 / 경력

- ML / DL 개발
  - 음성합성, 음성인식, NLP, 컴퓨터 비전 등
- Data 엔지니어링
- Front-end / Back-end 개발
- Mobile App 개발

Thank You

Q & A