DEVIEW
2019

감정연기와 외국어가 가능한 인공지능 성우

이영근
네오사피엔스

Deview에서 본 음성합성

2017

"딥러닝"을 활용한 음성합성

Deview에서 본 음성합성

2018

"다양한 목소리"를 가진 음성합성

Deview에서 본 음성합성

2019

1. “감정연기”가 가능한 음성합성
2. “외국어”가 가능한 음성합성

1. 감정연기가 가능한 음성합성

DEVIEW
2019

Source: Netflix 드라마 지정생존자 시즌1 1화

2. 외국어가 가능한 음성합성

DEVIEW
2019

Source: https://www.youtube.com/watch?v=YF1iXrxwcLA

이후 순서

DEVIEW
2019

1. 딥러닝 음성합성 개요
2. 감정연기를 할 수 있는 음성합성
3. 다양한 언어를 할 수 있는 음성합성

1. 딥러닝 음성합성 개요

딥러닝 기반의 End-to-end 음성합성

DEVIEW
2019

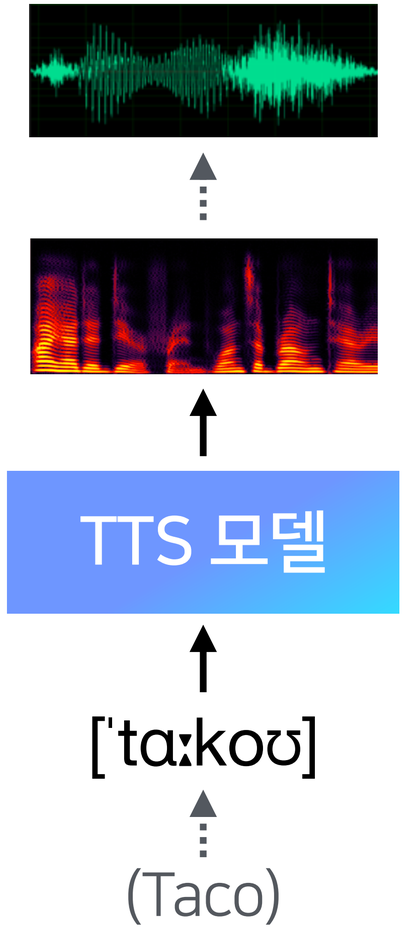
- 입력, 출력 정의

- 모델 정의

음성합성 모델의 입출력

- 출력: Spectrogram
  - 음성파형을 2차원 행렬로 표현한 것

- 입력: 음소 sequence
  - 발음의 기본단위

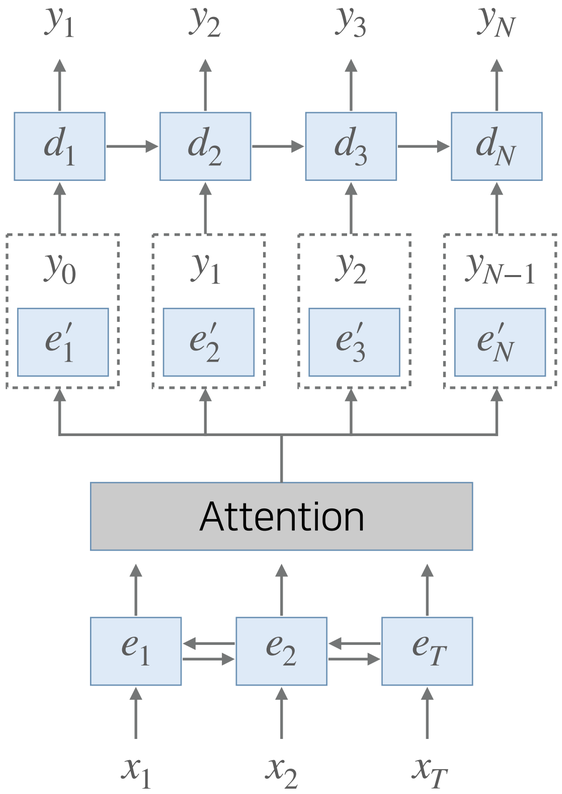


모델: Tacotron[1]

DEVIEW 2019

Decoder

Encoder



x	텍스트 입력
y	음성 출력
T	텍스트 길이
N	음성 길이
⬚	Concatenate

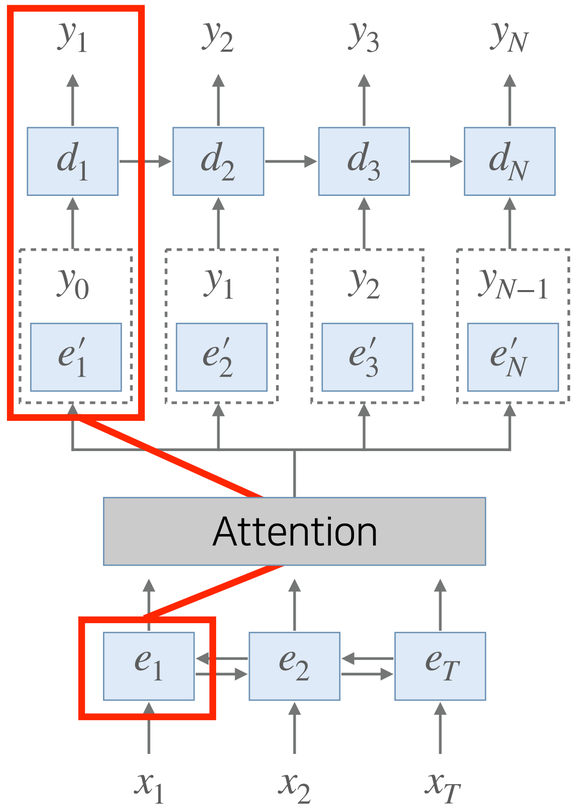
1. Wang, Yuxuan, et al. "Tacotron: Towards End-to-End Speech Synthesis." Interspeech 2017, (2017).

모델: Tacotron[1]

DEVIEW 2019

Decoder

Encoder



x	텍스트 입력
y	음성 출력
T	텍스트 길이
N	음성 길이
⬚	Concatenate

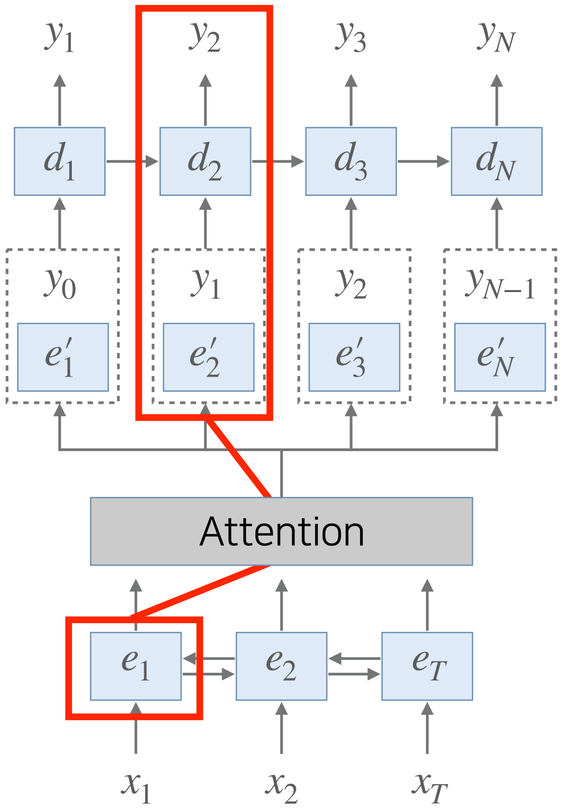
1. Wang, Yuxuan, et al. "Tacotron: Towards End-to-End Speech Synthesis." Interspeech 2017, (2017).

모델: Tacotron[1]

DEVIEW 2019

Decoder

Encoder



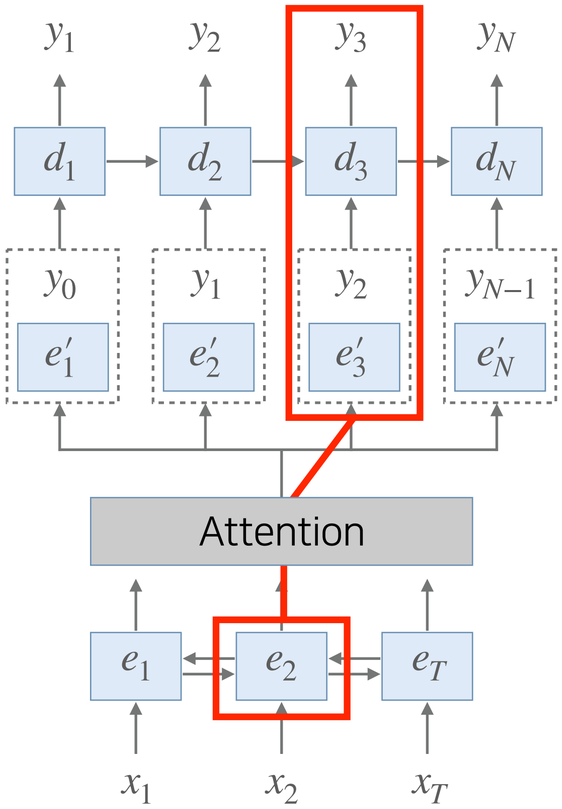
x	텍스트 입력
y	음성 출력
T	텍스트 길이
N	음성 길이
⬚	Concatenate

1. Wang, Yuxuan, et al. "Tacotron: Towards End-to-End Speech Synthesis." Interspeech 2017, (2017).

모델: Tacotron[1]

Decoder

Encoder



x	텍스트 입력
y	음성 출력
T	텍스트 길이
N	음성 길이
⬚	Concatenate

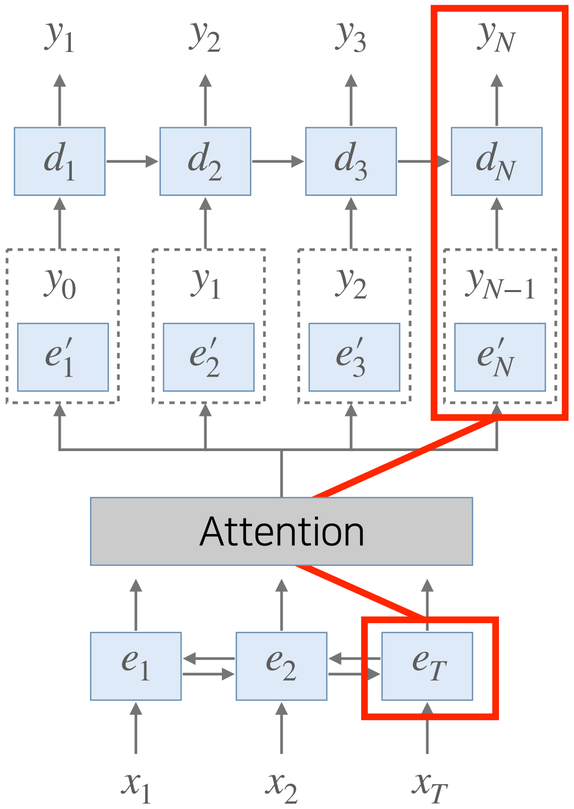
1. Wang, Yuxuan, et al. "Tacotron: Towards End-to-End Speech Synthesis." Interspeech 2017, (2017).

모델: Tacotron[1]

DEVIEW 2019

Decoder

Encoder

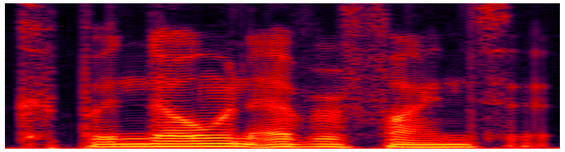


x	텍스트 입력
y	음성 출력
T	텍스트 길이
N	음성 길이
⬚	Concatenate

1. Wang, Yuxuan, et al. "Tacotron: Towards End-to-End Speech Synthesis." Interspeech 2017, (2017).

Tacotron의 학습

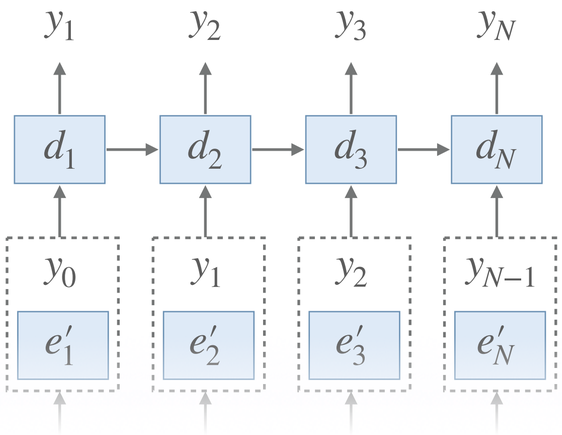
DEVIEW 2019



$$Loss = ||GT - y_{1:N}||_1$$

실제음성과 합성음성 간의 L1 loss

Decoder



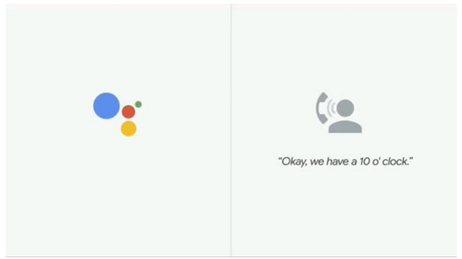
* GT: 원본 음성데이터

2. “감정연기"를 할 수 있는 음성합성

TTS는 이미 충분히 자연스럽다?

DEVIEW 2019

최근 발표된 TTS 모델들의 성능이 상당히 좋은 것으로 알려져 있음.



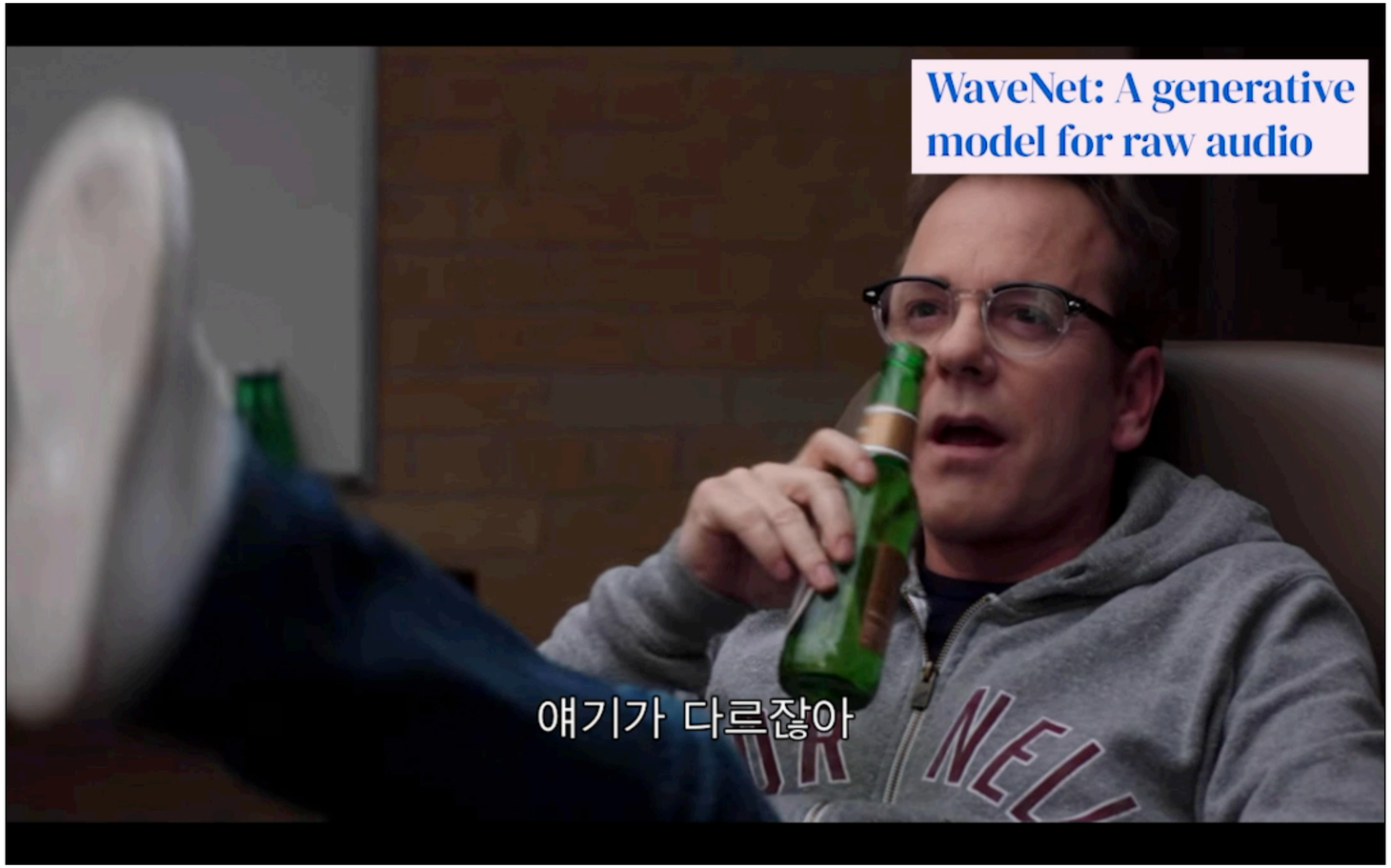
System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

차이가 매우 적음

Table 1. Mean Opinion Score (MOS) evaluations with 95% confidence intervals computed from the t-distribution for various systems.

Source: Google duplex, 네이버 AiCall 유튜브 캡처.
Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." 2018 ICASSP. IEEE, 2018.

TTS 더빙이 어려움



TTS가 해결해야 할 문제점

1. 상황에 맞는 “스타일”로 음성을 생성할 수 있어야 함.
2. 문장의 특정한 부분의 스타일도 컨트롤 할 수 있어야 함.

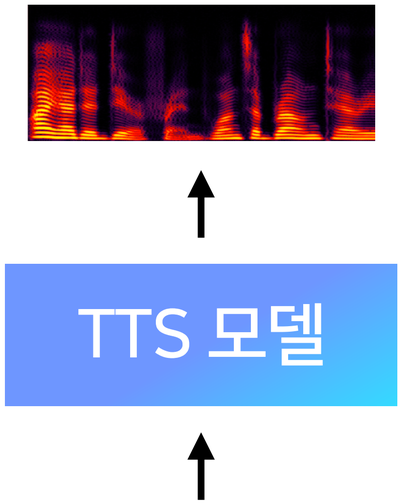
TTS가 다양한 스타일을 표현하려면?

DEVIEW 2019

- 일반적인 TTS의 입출력
  Input: 텍스트
  Output: 음성

- Input의 텍스트만 바꿔선 스타일을 조종하기 어려움.

-> 텍스트 이외의 입력을 받아서 조종이 가능하게 만들자.

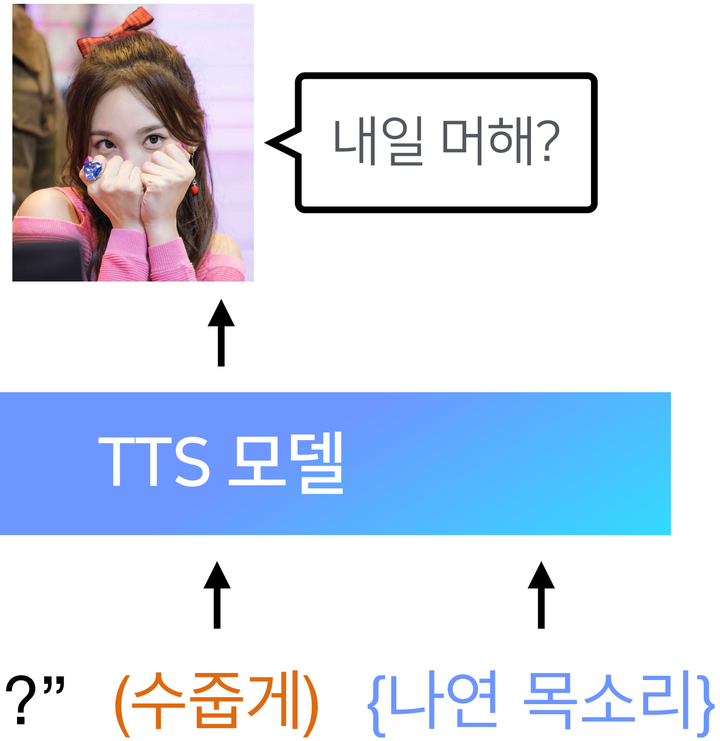


TTS가 다양한 스타일을 표현하려면?

DEVIEW 2019

- 일반적인 TTS의 입출력
  - Input: 텍스트
  - Output: 음성

- 개선된 TTS의 입출력
  - Input: 텍스트, 스타일, 화자(목소리), …
  - Output: 음성



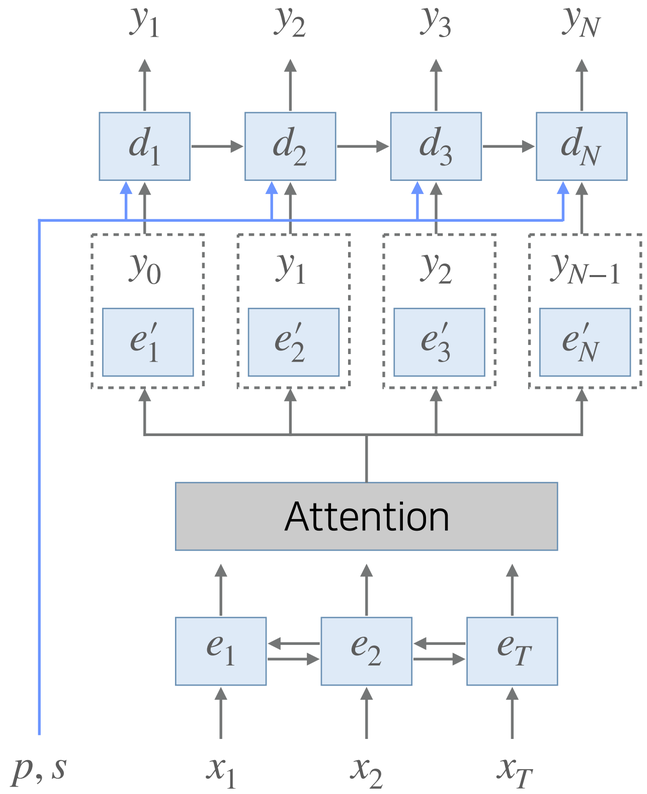
Source: http://mcjoun.tistory.com/92

추가 입력이 반영된 Tacotron

DEVIEW 2019

Decoder

Encoder



x	텍스트 입력
y	음성 출력
p	스타일 정보
s	화자 정보
T	텍스트 길이
N	음성 길이
⬚	Concatenate

스타일 정보는 어떻게 줄까?

- 감정과 같은 스타일 정보는 label을 만들기 까다로움.

- 음성에서 스타일 정보를 추출하는 Neural network를 이용.

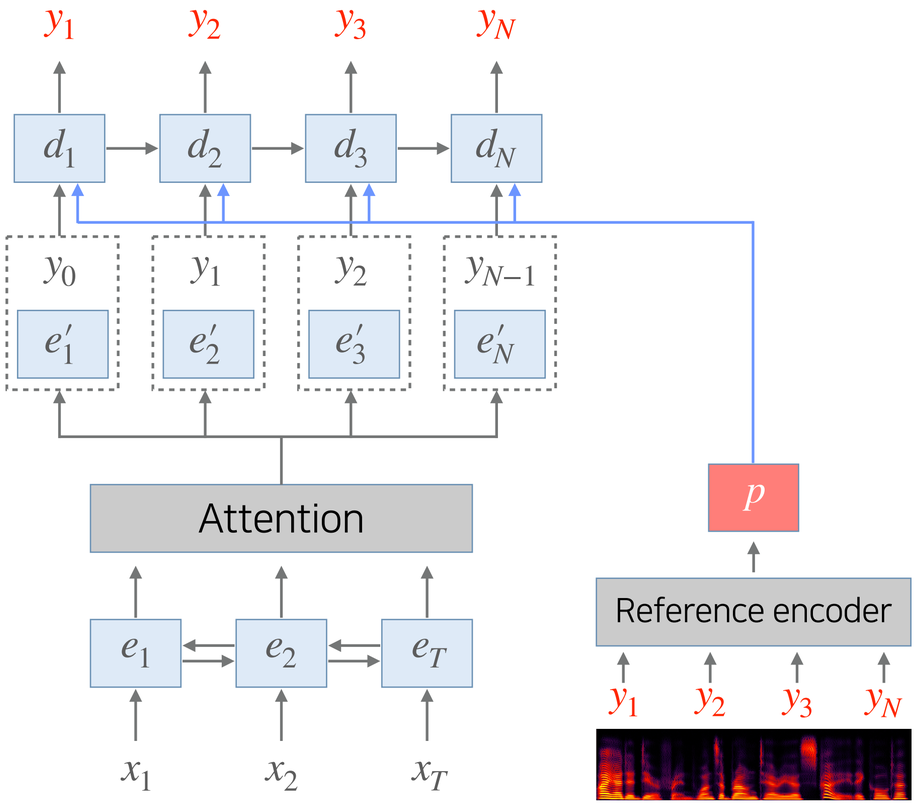
- 추출된 스타일 정보를 담고 있는 style embedding을 추가 입력으로 (p) 사용.

Style embedding을 사용한 Tacotron

DEVIEW 2019

Decoder

Encoder



x	텍스트 입력
y	음성 출력
p	스타일 정보
T	텍스트 길이
N	음성 길이
⬚	Concatenate

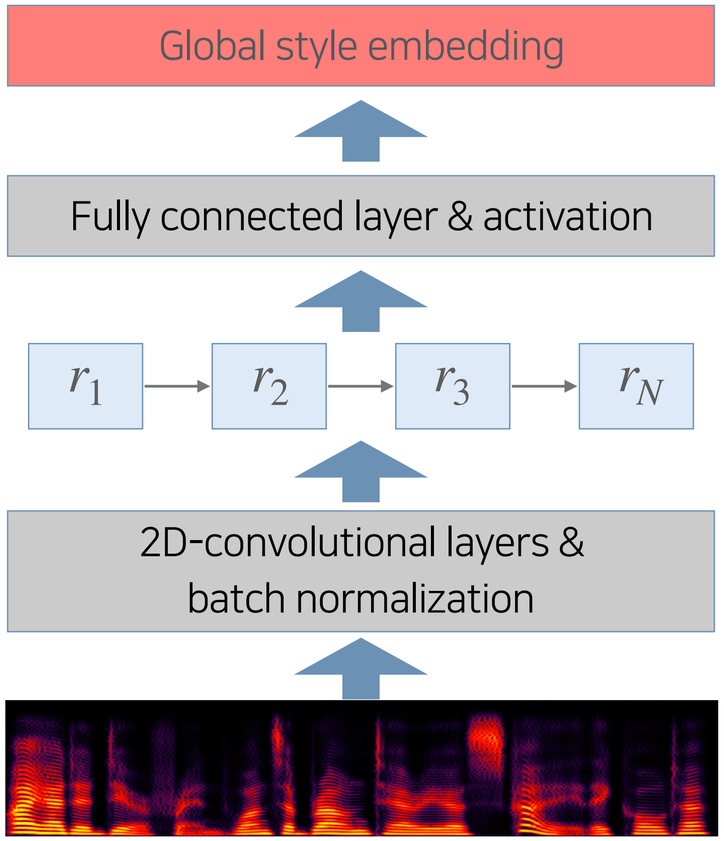
1. Skerry-Ryan, R. J., et al. "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron." International Conference on Machine Learning. 2018.

Style reference encoder

Style output

GRU

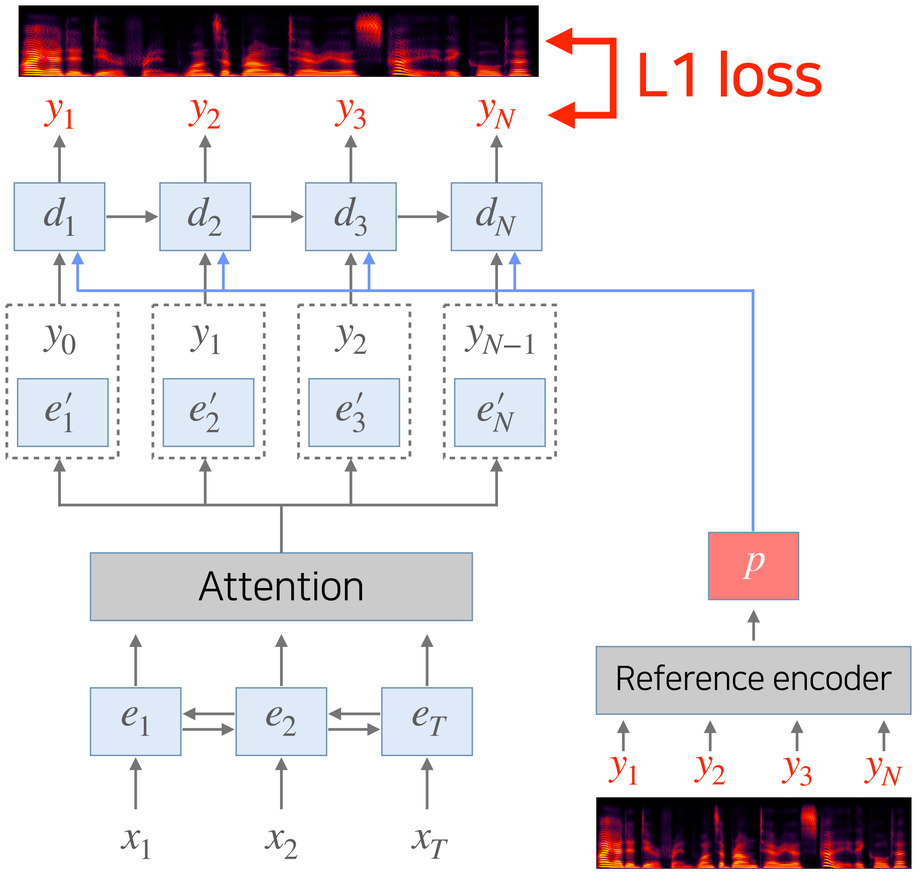
음성 input



Style embedding을 사용한 Tacotron

Decoder

Encoder



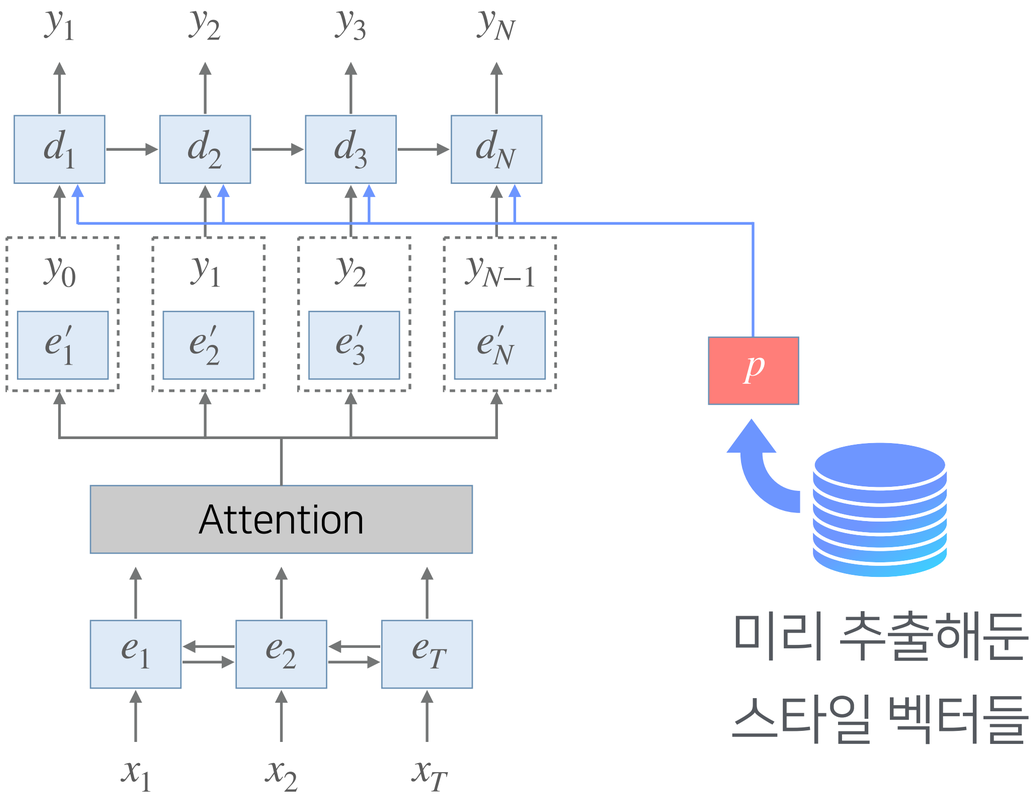
Training

x	텍스트 입력
y	음성 출력
p	스타일 정보
T	텍스트 길이
N	음성 길이
⬚	Concatenate

Style embedding을 사용한 Tacotron

Decoder

Encoder



미리 추출해둔
스타일 벡터들

Inference

x	텍스트 입력
y	음성 출력
p	스타일 정보
T	텍스트 길이
N	음성 길이
⬚	Concatenate

Style embedding의 활용

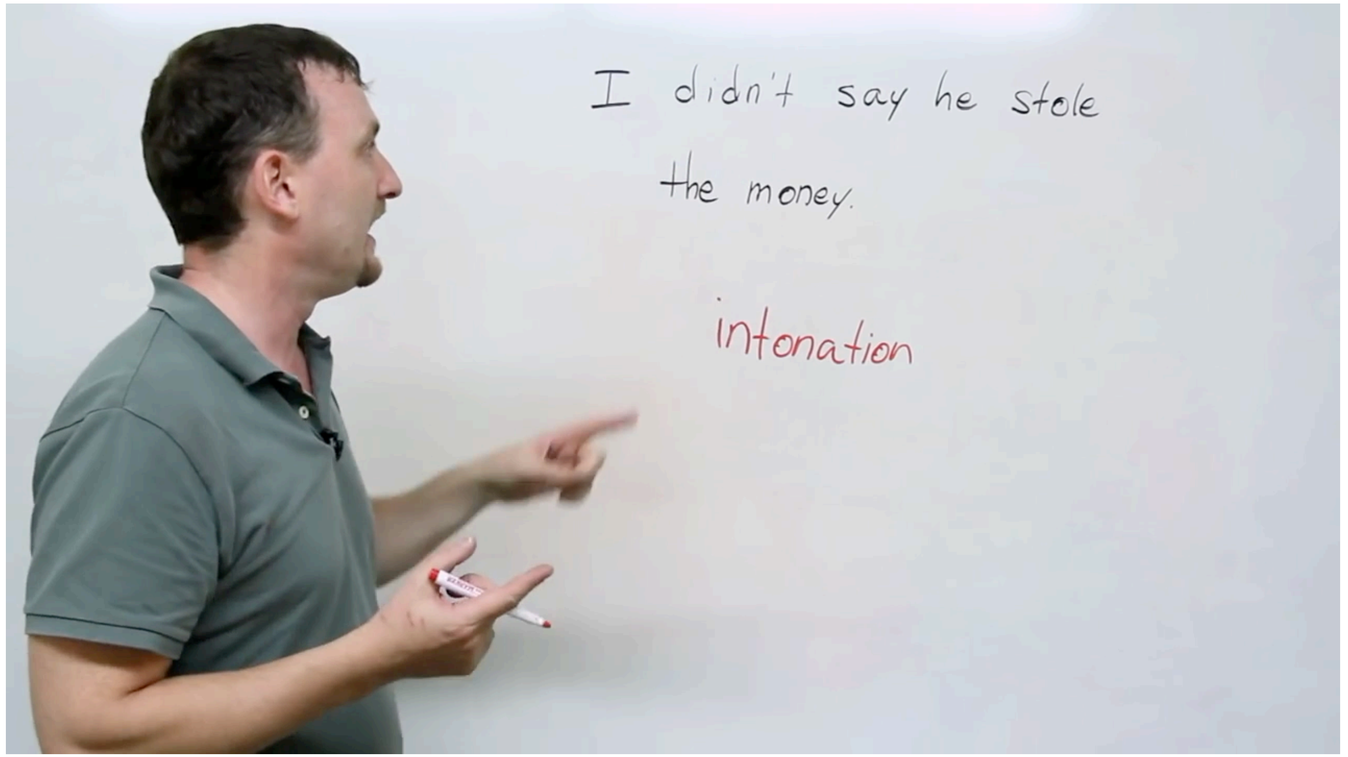
앞서 본 style embedding 의 한계점

- 스타일을 한 문장 단위에서 조종할 수 있었음
  - 전반적인 감정
  - 전반적인 문장의 느낌
  - ...

- 하지만 특정한 음소, 단어 단위의 스타일을 조종하는 것은 여전히 불가능
  - 더 미세한 단위에서 조종할 수 있는 형태의 embedding이 필요

문장의 특정 부분을 강조하여 읽기

DEVIEW
2019

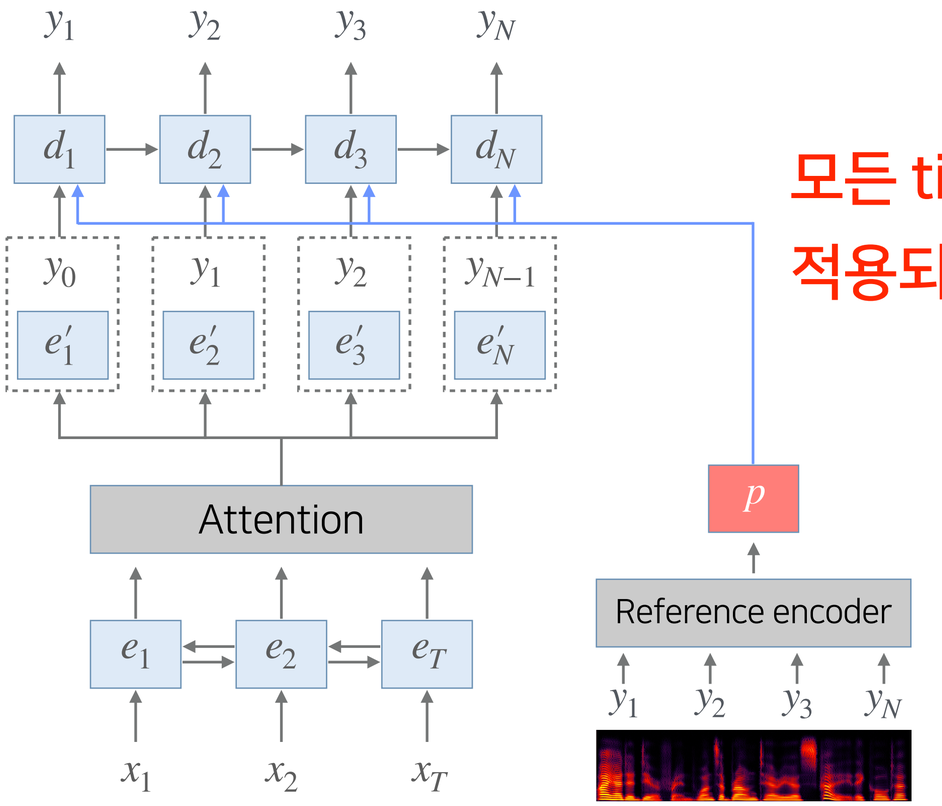


Source: https://www.youtube.com/watch?v=4lqrm82LED4

Style embedding을 사용한 Tacotron

Decoder

Encoder

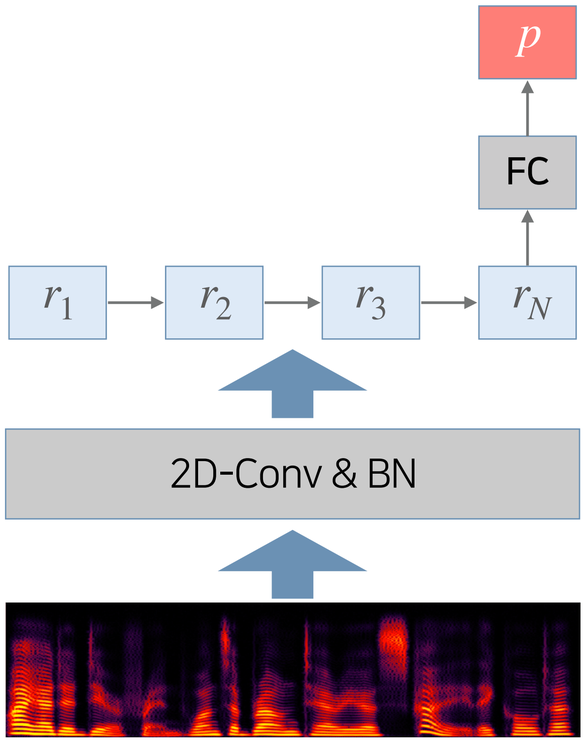


모든 time-step에 동일한 스타일이 적용되기 때문에 일부만 조종 불가.

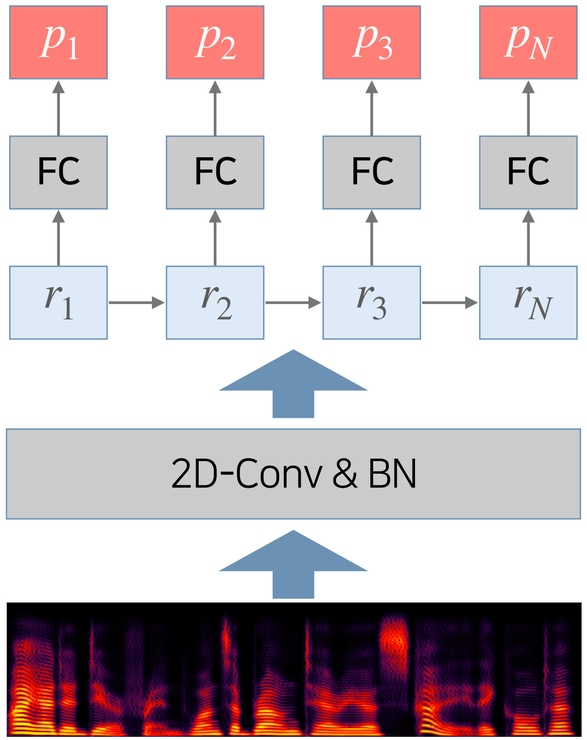
2가지 종류의 style embedding

DEVIEW 2019

Fixed length



Variable length



1. Skerry-Ryan, R. J., et al. "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron." International Conference on Machine Learning. 2018.

문장의 특정부분 스타일을 바꾸는 방법들

DEVIEW
2019

1. Speech-side style control
   - 음성의 특정 시점에서의 스타일을 컨트롤

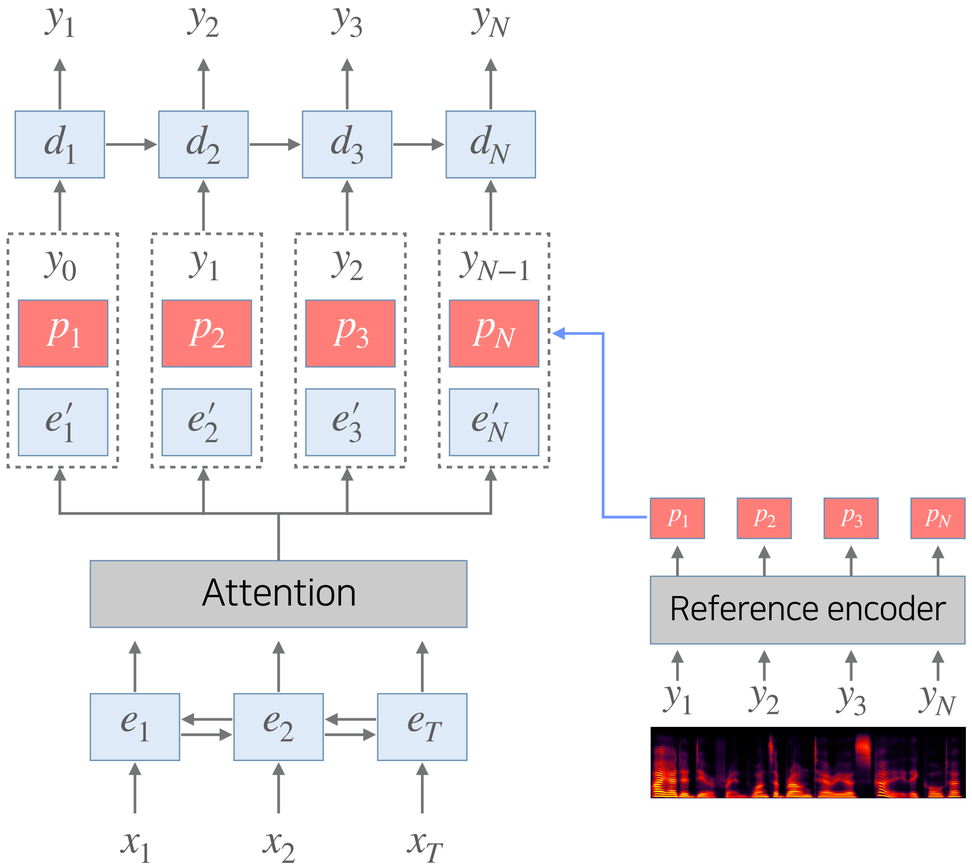
2. Text-side style control
   - 문장의 특정 발음(텍스트)에서의 스타일을 컨트롤

1. Speech-side style control

DEVIEW 2019

Decoder

Encoder



x	텍스트 입력
y	음성 출력
p	스타일 정보
T	텍스트 길이
N	음성 길이
⬚	Concatenate

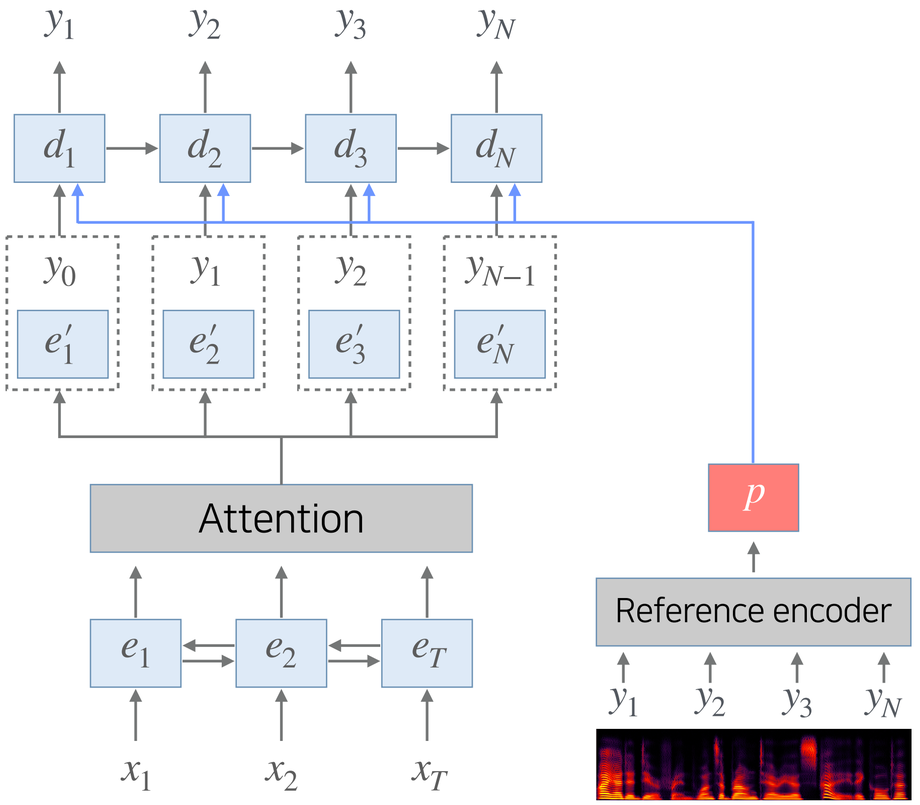
1. Lee, Younggun, and Taesu Kim. "Robust and fine-grained prosody control of end-to-end speech synthesis." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

Style embedding을 사용한 Tacotron

DEVIEW 2019

Decoder

Encoder



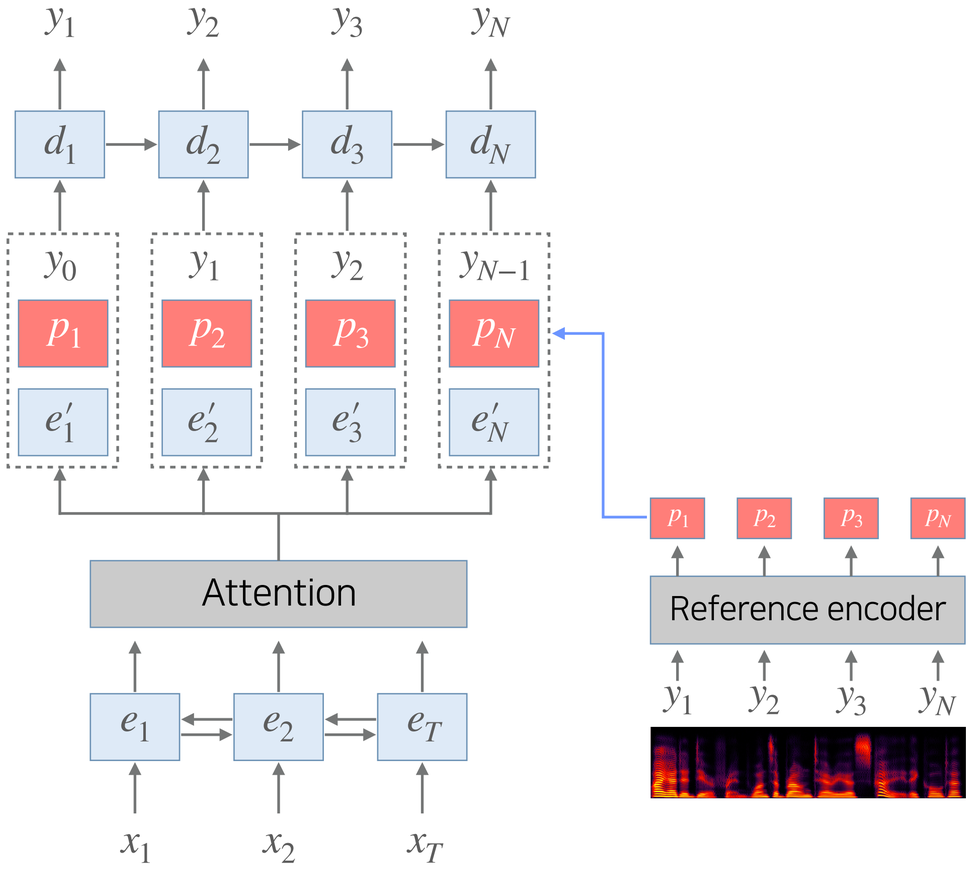
x	텍스트 입력
y	음성 출력
p	스타일 정보
T	텍스트 길이
N	음성 길이
⬚	Concatenate

1. Speech-side style control

DEVIEW 2019

Decoder

Encoder



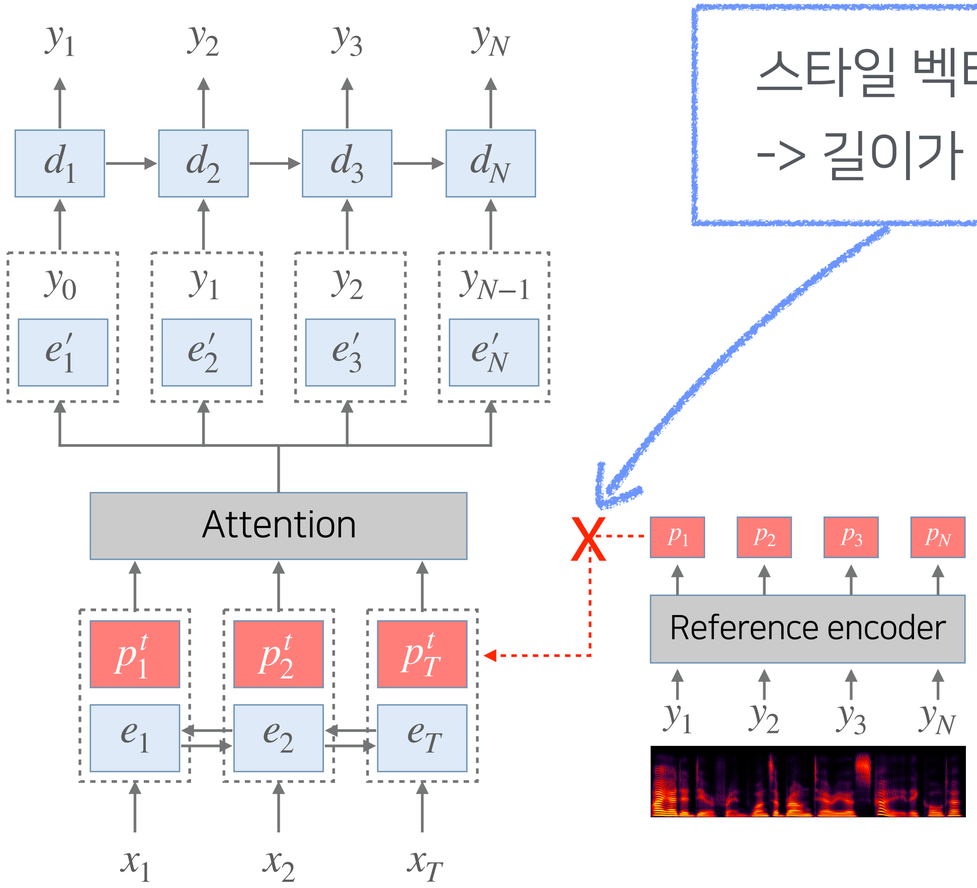
x	텍스트 입력
y	음성 출력
p	스타일 정보
T	텍스트 길이
N	음성 길이
⬚	Concatenate

2. Text-side style control

DEVIEW 2019

Decoder

Encoder



스타일 벡터를 텍스트 쪽에 넣어보자
-> 길이가 다르다는 문제점...

x	텍스트 입력
y	음성 출력
p	스타일 정보
T	텍스트 길이
N	음성 길이
⬚	Concatenate

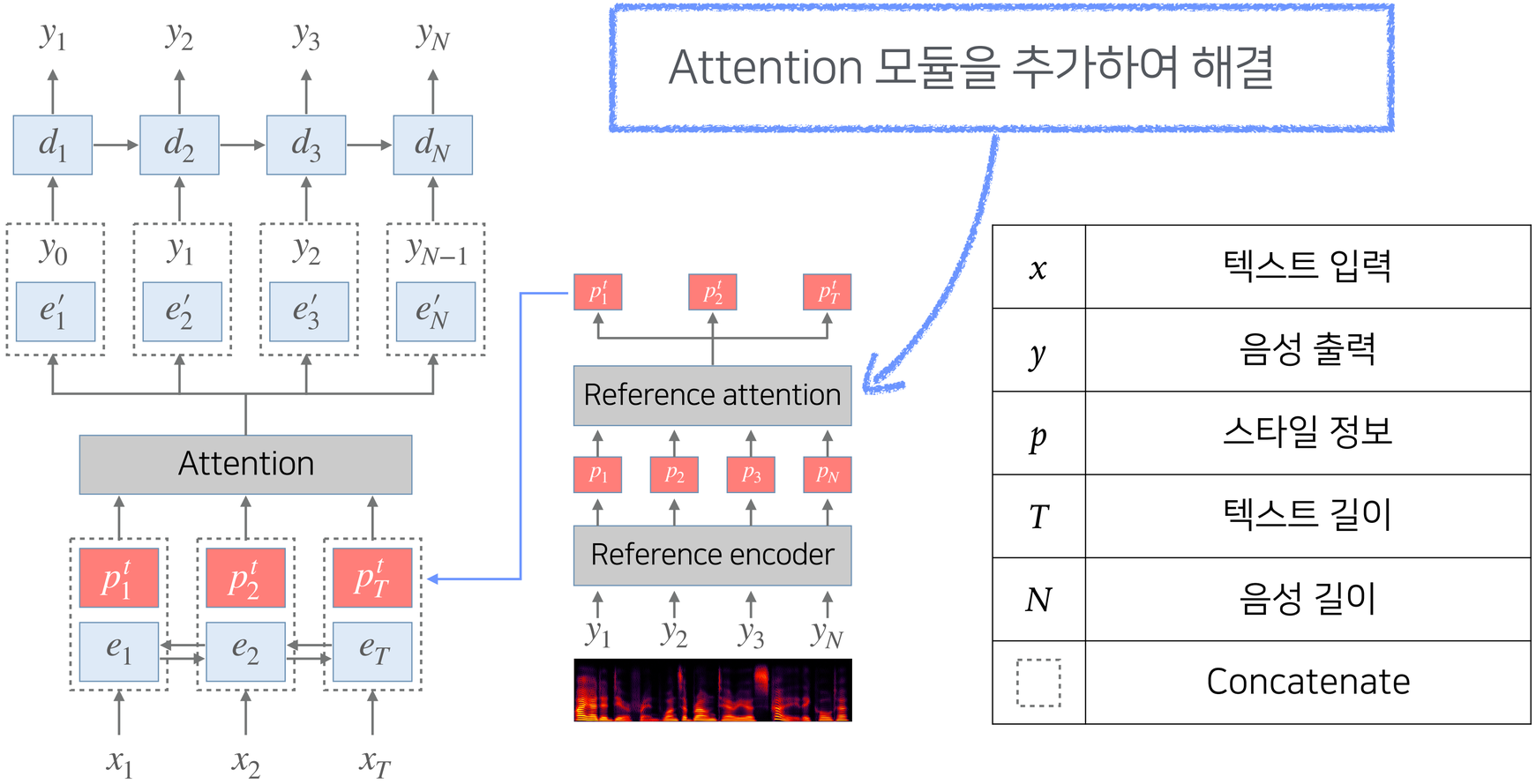
1. Lee, Younggun, and Taesu Kim. "Robust and fine-grained prosody control of end-to-end speech synthesis." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

2. Text-side style control

DEVIEW 2019

Decoder

Encoder



x	텍스트 입력
y	음성 출력
p	스타일 정보
T	텍스트 길이
N	음성 길이
⬚	Concatenate

1. Lee, Younggun, and Taesu Kim. "Robust and fine-grained prosody control of end-to-end speech synthesis." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

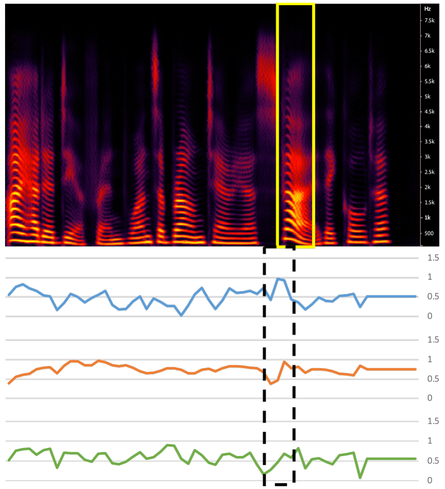
Text-side style control 데모

DEVIEW 2019

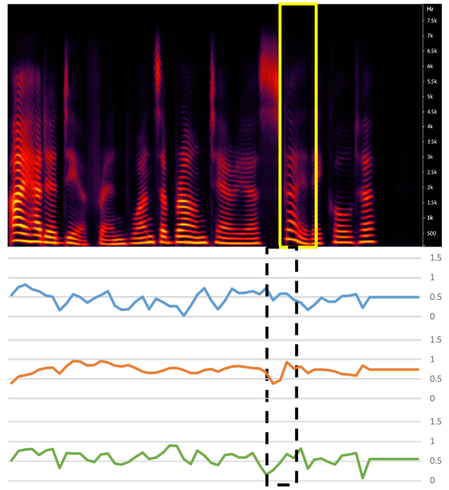
입력: “I had a dear friend once, a brown terrier, Skye they called her.”

음성

스타일 벡터 (p)



음의 높낮이 변화



원본

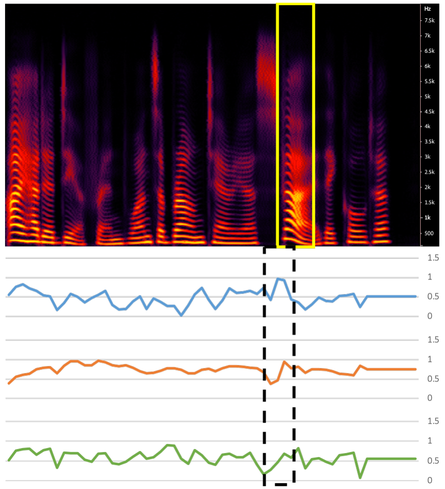
1st dim 수정

Text-side style control 데모

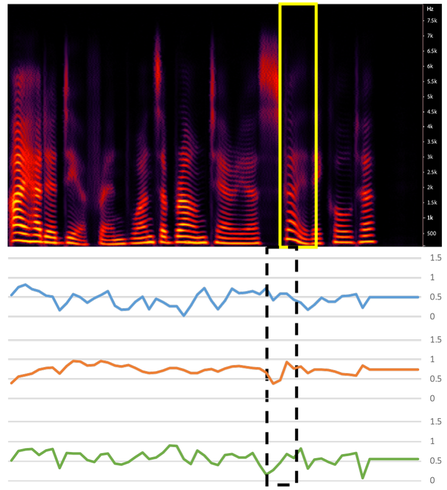
입력: “I had a dear friend once, a brown terrier, Skye they called her.”

음성

스타일
벡터 (p)



음의 높낮이
변화



원본

1st dim
수정

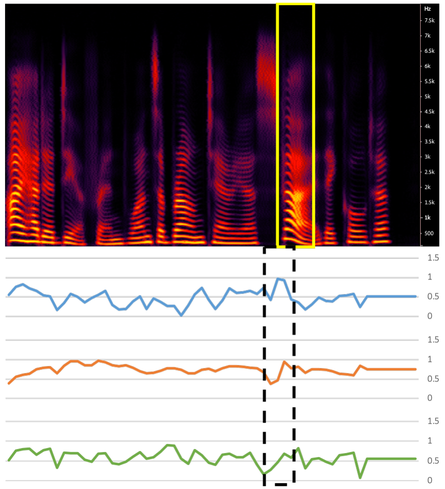
Text-side style control 데모

DEVIEW 2019

입력: “I had a dear friend once, a brown terrier, Skye they called her.”

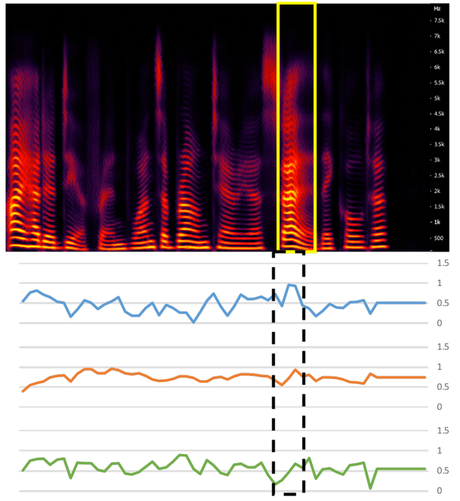
음성

스타일 벡터 (p)



원본

소리크기와 속도 변화



2nd dim 수정

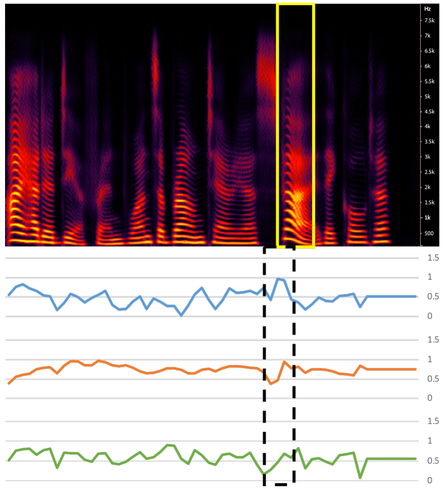
Text-side style control 데모

DEVIEW 2019

입력: “I had a dear friend once, a brown terrier, Skye they called her.”

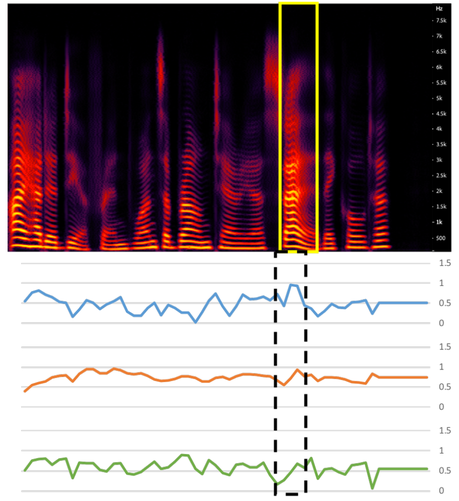
음성

스타일 벡터 (p)



원본

소리크기와 속도 변화



2nd dim 수정

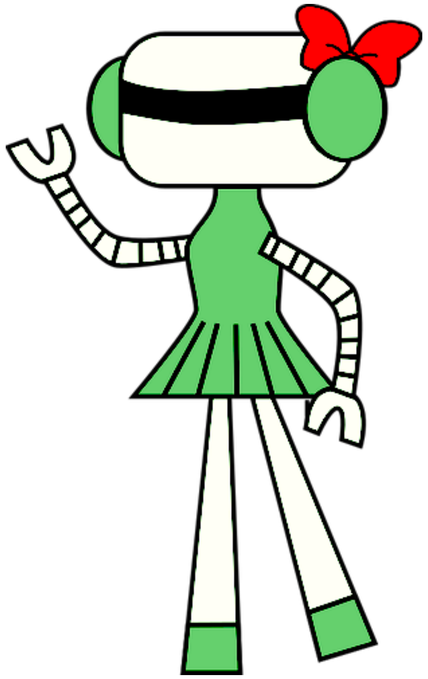
목표: 절대음감 게임

DEVIEW
2019

Source: https://www.youtube.com/watch?v=qEFJpSeTdd4&t=2s

목표: 절대음감 게임

DEVIEW
2019



정리

1. 상황에 맞는 “스타일”로 음성을 생성할 수 있어야 함.
   -> Style embedding의 도입으로 해결

2. 문장의 특정한 부분의 스타일도 컨트롤 할 수 있어야 함.
   -> Speech/Text side style control의 도입으로 해결

3. 다양한 “언어”를 할수 있는 음성합성

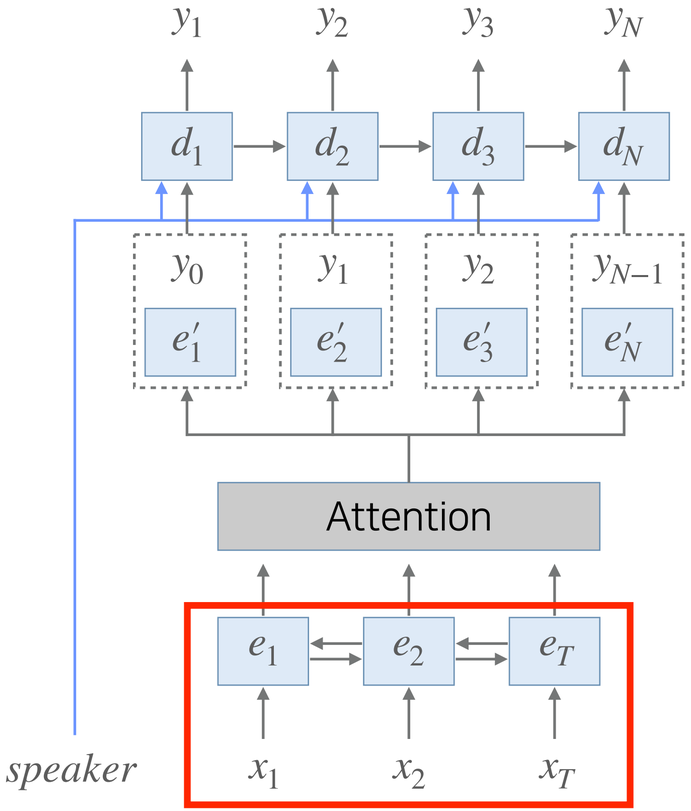
다국어 TTS를 쓰면 뭐가 좋을까?

- 원래의 목소리를 살려서 다른 나라 말을 할 수 있다.
  - 컨텐츠 수출 용이
  - 자연스러운 통역
  - ...

End-to-end 다국어 TTS 구조

- 전체적인 구조는 동일.
- 다화자이므로 화자정보 input이 있어야 함.

- 음소 입력을 벡터로 바꿔주는 부분만 다름.
  - Embedding layer



End-to-end 다국어 TTS

Embedding
Dictionary

a: (0.4, 0.6, 0.4, 0.5)

b: (0.4, 0.5, 0.4, 0.2)

...

z: (0.7, 0.6, 0.5, 0.5)

Embedding
matrix

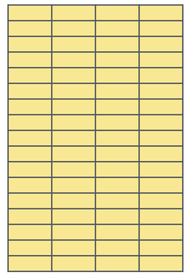
0.4	0.6	0.4	0.5
0.4	0.5	0.4	0.2
...	...	...	...
...	...	...	...
...	...	...	...
0.7	0.6	0.5	0.5

End-to-end 다국어 TTS

DEVIEW 2019

1개국어 Embedding matrix

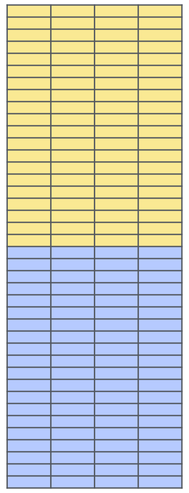
a: (0.4, 0.6, 0.4, 0.5)
b: (0.4, 0.5, 0.4, 0.2)
...
z: (0.7, 0.6, 0.5, 0.5)



2개국어 Embedding matrix

a: (0.4, 0.6, 0.4, 0.5)
b: (-0.4, 0.5, 0.4, 0.2)
...
z: (0.7, 0.6, -0.5, 0.5)

ㄱ: (0.1, 0.6, 0.4, 0)
ㄴ: (0.3, 0.9, 0.3, 0.2)
...
ㅣ: (0.8, 0.6, 0.5, -0.1)

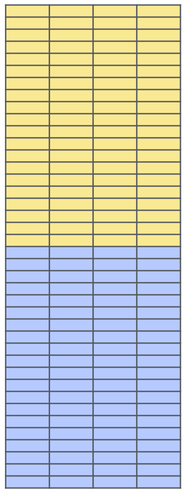


End-to-end 다국어 TTS

2개국어 Embedding matrix

a: (0.4, 0.6, 0.4, 0.5)
b: (-0.4, 0.5, 0.4, 0.2)
...
z: (0.7, 0.6, -0.5, 0.5)

ㄱ: (0.1, 0.6, 0.4, 0)
ㄴ: (0.3, 0.9, 0.3, 0.2)
...
ㅣ: (0.8, 0.6, 0.5, -0.1)



서로 다른 언어의 embedding이
같은 space안에서 학습됨.

화자를 고정하고 다른 언어를 입력하여
다국어로 말할 수 있게 학습됨.

End-to-end 다국어 TTS 학습

- Objective function
  - 원래와 동일한 L1 loss (원본음성과 합성음성간 차이)

- 학습 데이터
  - 한국어 화자 + 영어 화자 동시에 사용
  - 모든 화자는 1개국어만 구사

다국어 TTS의 활용

DEVIEW
2019

앞서 본 모델의 문제점?

- 억양에 대한 조종을 할 수 없음

	한국어	영어
한국어 억양	O	O
영어 억양	X	X

	한국어	영어
한국어 억양	X	X
영어 억양	O	O

앞서 본 모델의 문제점?

DEVIEW 2019

- 억양에 대한 조종을 할 수 없음

	한국어	영어
한국어 억양	O	O
영어 억양	O	O

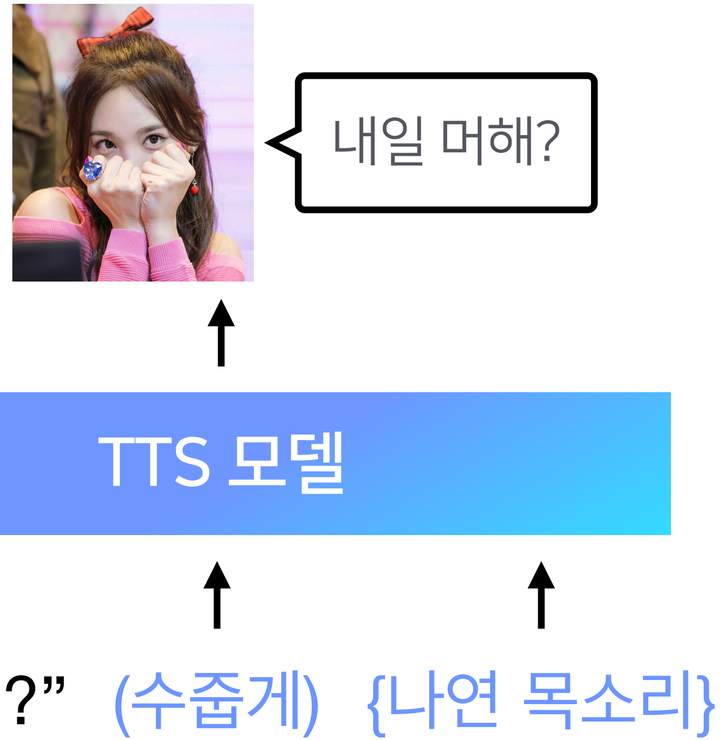
	한국어	영어
한국어 억양	O	O
영어 억양	O	O

비슷한 문제를 어떻게 풀었더라...?

DEVIEW 2019

- 일반적인 TTS의 입출력
  Input: 텍스트
  Output: 음성

- 개선된 TTS의 입출력
  Input: 텍스트, 스타일, 화자(목소리), …
  Output: 음성



Source: http://mcjoun.tistory.com/92

개선된 end-to-end 다국어 TTS

DEVIEW 2019

Input: 텍스트, 스타일, 화자(목소리), 언어

Output: 음성

What are you doing tomorrow?

TTS 모델

(수줍게) {나연 목소리} [영어억양]

“What are you doing tomorrow?”

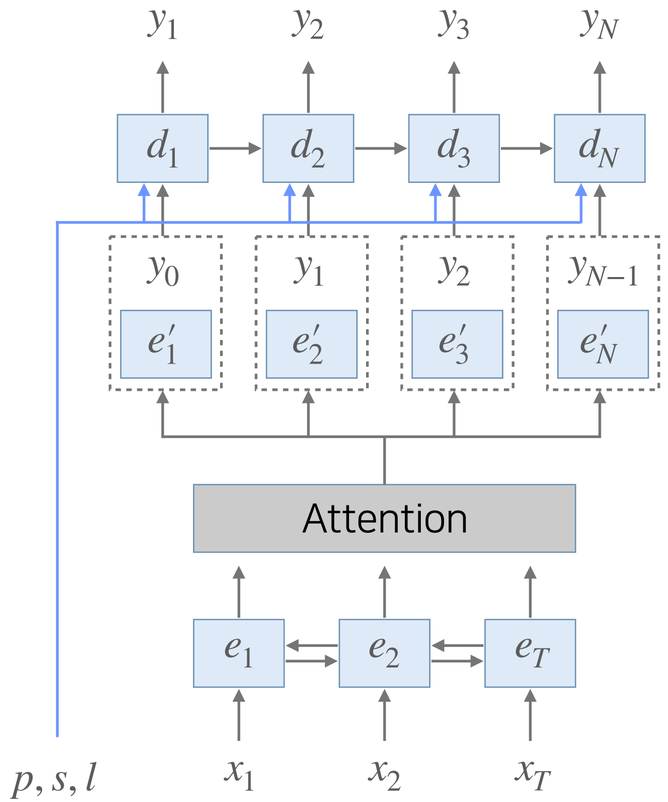
Source: http://mcjoun.tistory.com/92

유창함을 조종할 수 있는 다국어 TTS

DEVIEW 2019

Decoder

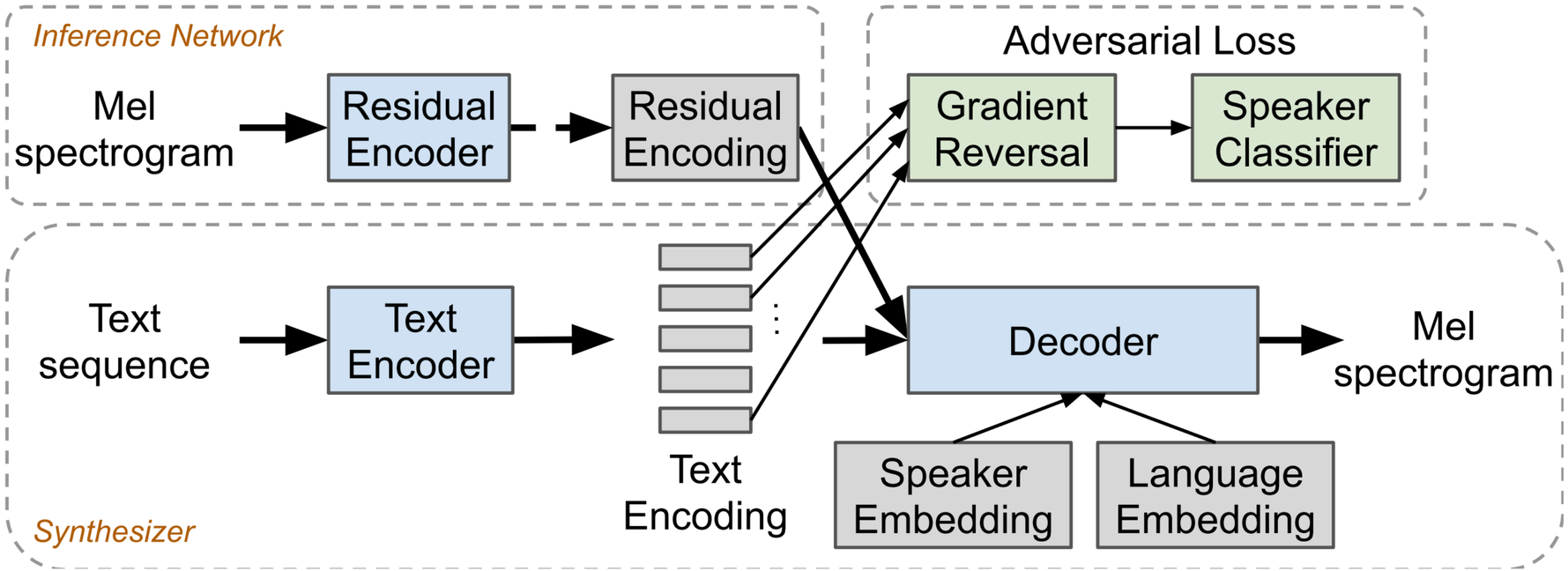
Encoder



x	텍스트 입력
y	음성 출력
p	스타일 정보
s	화자 정보
l	언어 정보
T	텍스트 길이
N	음성 길이
⬚	Concatenate

언어 embedding 학습을 위한 trick

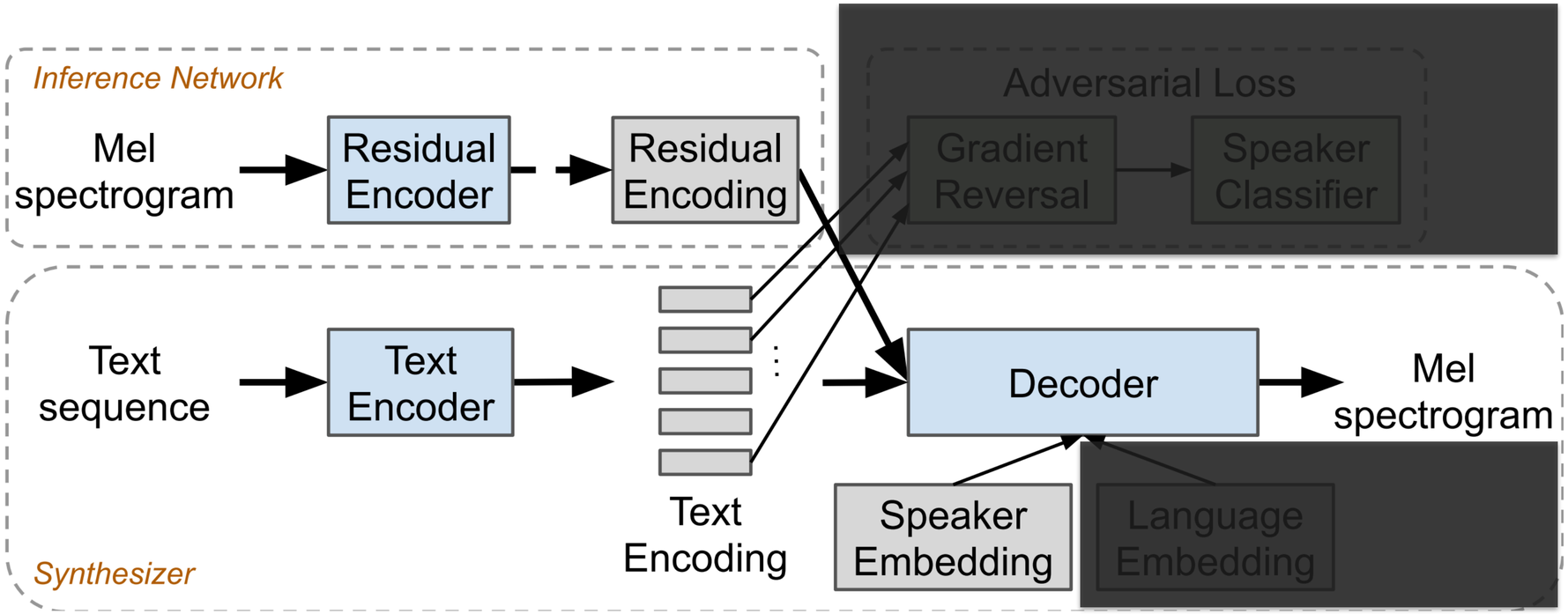
- Adversarial loss의 활용



Source: Zhang, Y. et al., Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. Proc. Interspeech 2019

언어 embedding 학습을 위한 trick

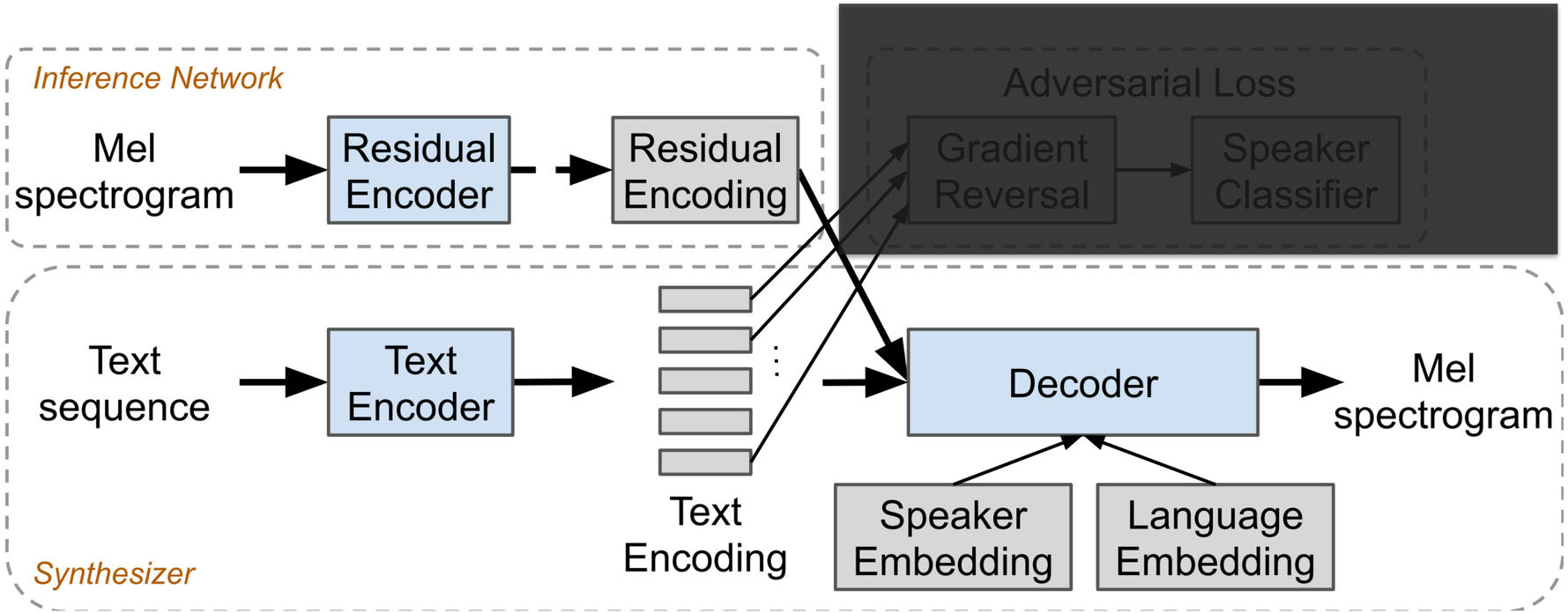
- Adversarial loss의 활용



Source: Zhang, Y. et al., Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. Proc. Interspeech 2019

언어 embedding 학습을 위한 trick

- Adversarial loss의 활용



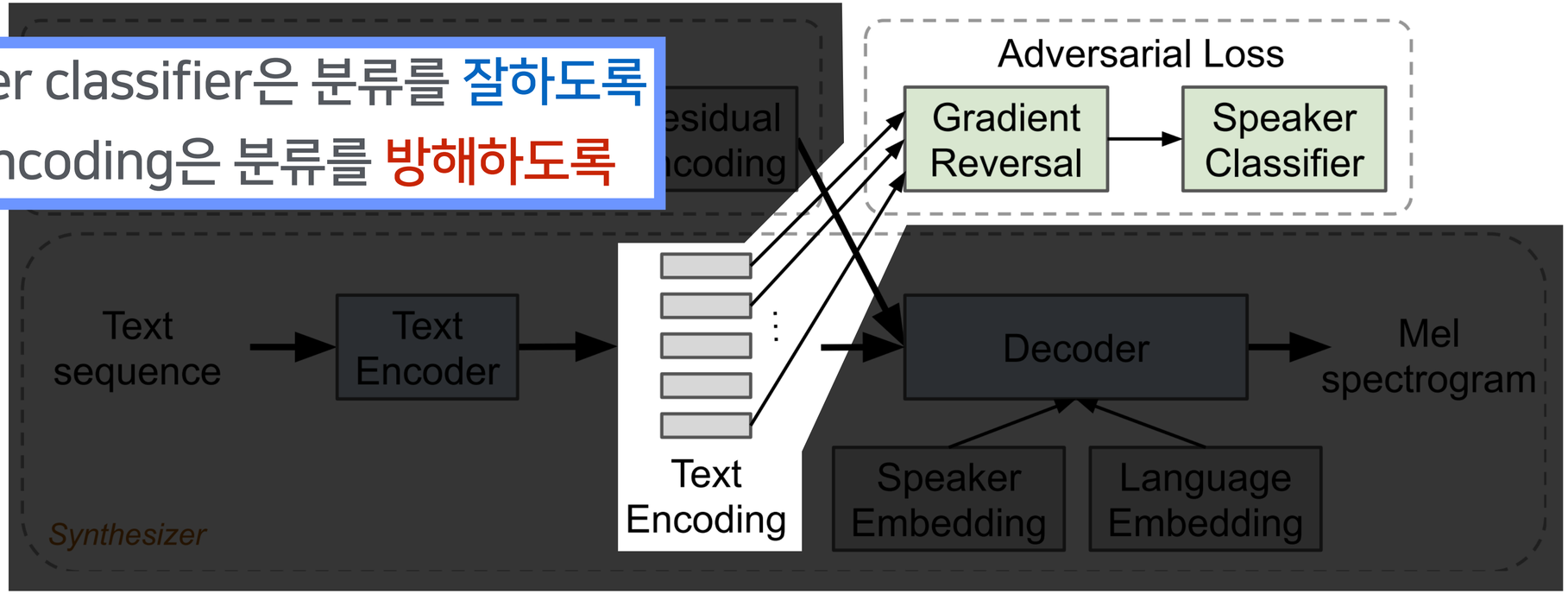
Source: Zhang, Y. et al., Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. Proc. Interspeech 2019

언어 embedding 학습을 위한 trick

DEVIEW 2019

- Adversarial loss의 활용

- Speaker classifier은 분류를 잘하도록
- Text-encoding은 분류를 방해하도록



Source: Zhang, Y. et al., Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. Proc. Interspeech 2019

Speaker adversarial loss는 왜?

- 우리가 가지고 있는 텍스트가 언어와 화자에 종속적임.

- 텍스트에 화자정보와 언어정보를 빼주어야 여러 조합의 음성 생성 가능.

Speaker adversarial loss는 왜?

DEVIEW 2019

가지고 있는 데이터

영문 텍스트

영어 억양
영어 화자

한글 텍스트

한국어 억양
한국어 화자

생성 하려는 데이터

영문 텍스트

영어 억양
한국어 화자

한국어 억양
한국어 화자

한글 텍스트

한국어 억양
영어 화자

영어 억양
영어 화자

유창함을 조종할 수 있는 다국어 TTS 데모

DEVIEW 2019

	한국어 화자	영어 화자
유창한 모국어		
유창한 외국어		
어눌한 외국어		
어눌한 모국어		

유창함을 조종할 수 있는 다국어 TTS 데모

DEVIEW 2019

	한국어 화자	영어 화자
유창한 모국어		
유창한 외국어		
어눌한 외국어		
어눌한 모국어		

유창함을 조종할 수 있는 다국어 TTS 데모

DEVIEW 2019

	한국어 화자	영어 화자
유창한 모국어		
유창한 외국어		
어눌한 외국어		
어눌한 모국어		

유창함을 조종할 수 있는 다국어 TTS 데모

DEVIEW 2019

	한국어 화자	영어 화자
유창한 모국어		
유창한 외국어		
어눌한 외국어		
어눌한 모국어		

유창함을 조종할 수 있는 다국어 TTS 데모

DEVIEW 2019

	한국어 화자	영어 화자
유창한 모국어		
유창한 외국어		
어눌한 외국어		
어눌한 모국어		

유창함을 조종할 수 있는 다국어 TTS 데모

DEVIEW 2019

	한국어 화자	영어 화자
유창한 모국어		
유창한 외국어		
어눌한 외국어		
어눌한 모국어		

유창함을 조종할 수 있는 다국어 TTS 데모

DEVIEW 2019

	한국어 화자	영어 화자
유창한 모국어		
유창한 외국어		
어눌한 외국어		
어눌한 모국어		

유창함을 조종할 수 있는 다국어 TTS 데모

DEVIEW 2019

	한국어 화자	영어 화자
유창한 모국어		
유창한 외국어		
어눌한 외국어		
어눌한 모국어		

오늘 기억할 내용

- 딥러닝 기반 TTS에서 목소리를 원하는 대로 컨트롤 하려고 한 연구들.
  - 화자 컨트롤
  - 스타일 컨트롤 (감정, 특정 부분 강조, …)
  - 언어 컨트롤 (언어의 유창한 정도)

- 컨트롤 할 정보에 대한 입력을 잘 정의해주고 모델링 해야함.

DEVIEW
2019

Q & A

DEVIEW
2019

Thank You