딥러닝 기반 사용자 적응형 음성합성 시스템

VOICE & DIALOGUE (TRACK A)
15:00 ~ 15:25

송은우

NAVER

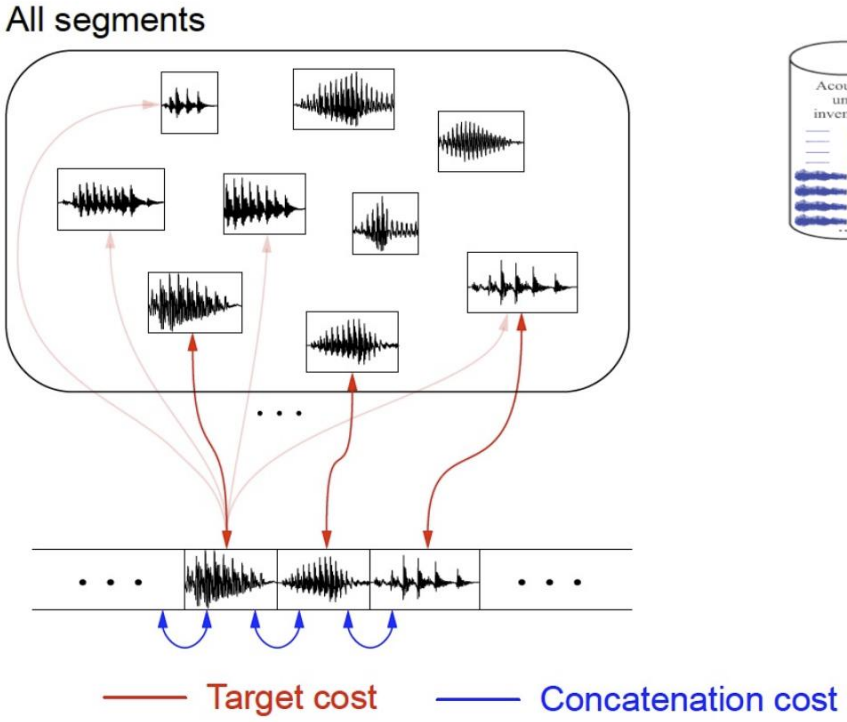
TTS AT NAVER

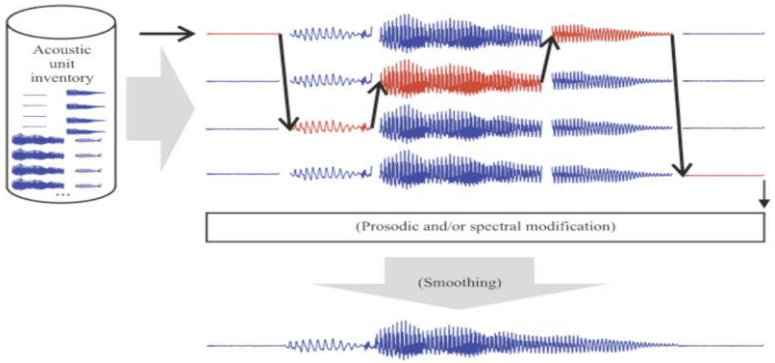
nVoice applications

Clova	Papago	NAVER i	Whale
	papago		whale
유인나 오디오북	사전 예문읽기	뉴스 본문듣기	Open API
	Dict	NEWS	NAVER open API

UNIT-SELECTION SPEECH SYNTHESIS SYSTEM (UTS)

Concatenate speech units (waveform) from a LARGE database





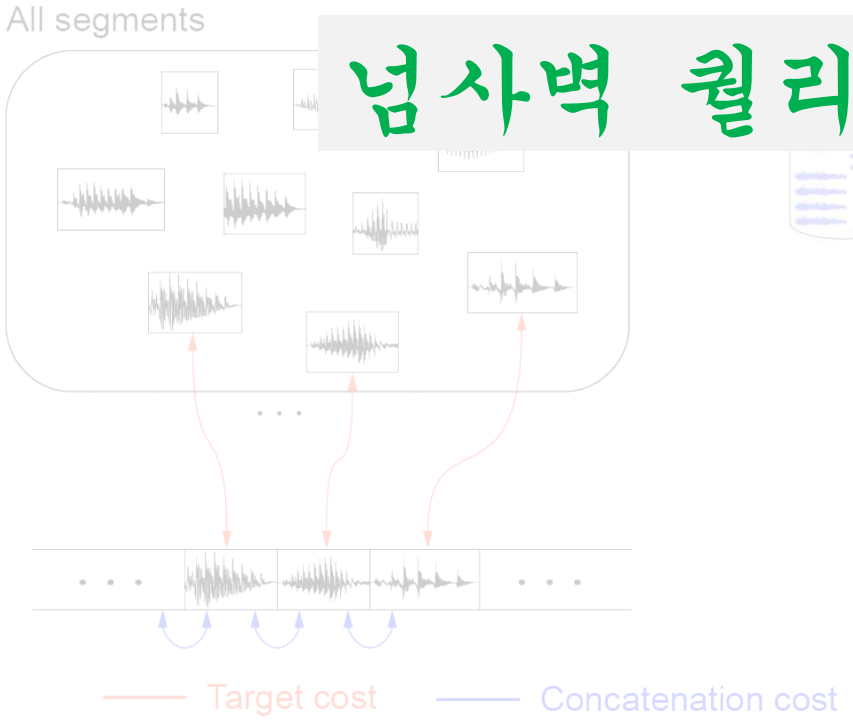
$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^{target}(t_i, u_i) + \sum_{i=2}^{n} C^{join}(u_{i-1}, u_i)$$

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

출처:
T. Keiichi, and H. Zen. "Fundamentals and recent advances in HMM-based speech synthesis." *Tutorial of INTERSPEECH,* 2009.

UNIT-SELECTION SPEECH SYNTHESIS SYSTEM (UTS)

Concatenate speech units (waveform) from a LARGE database



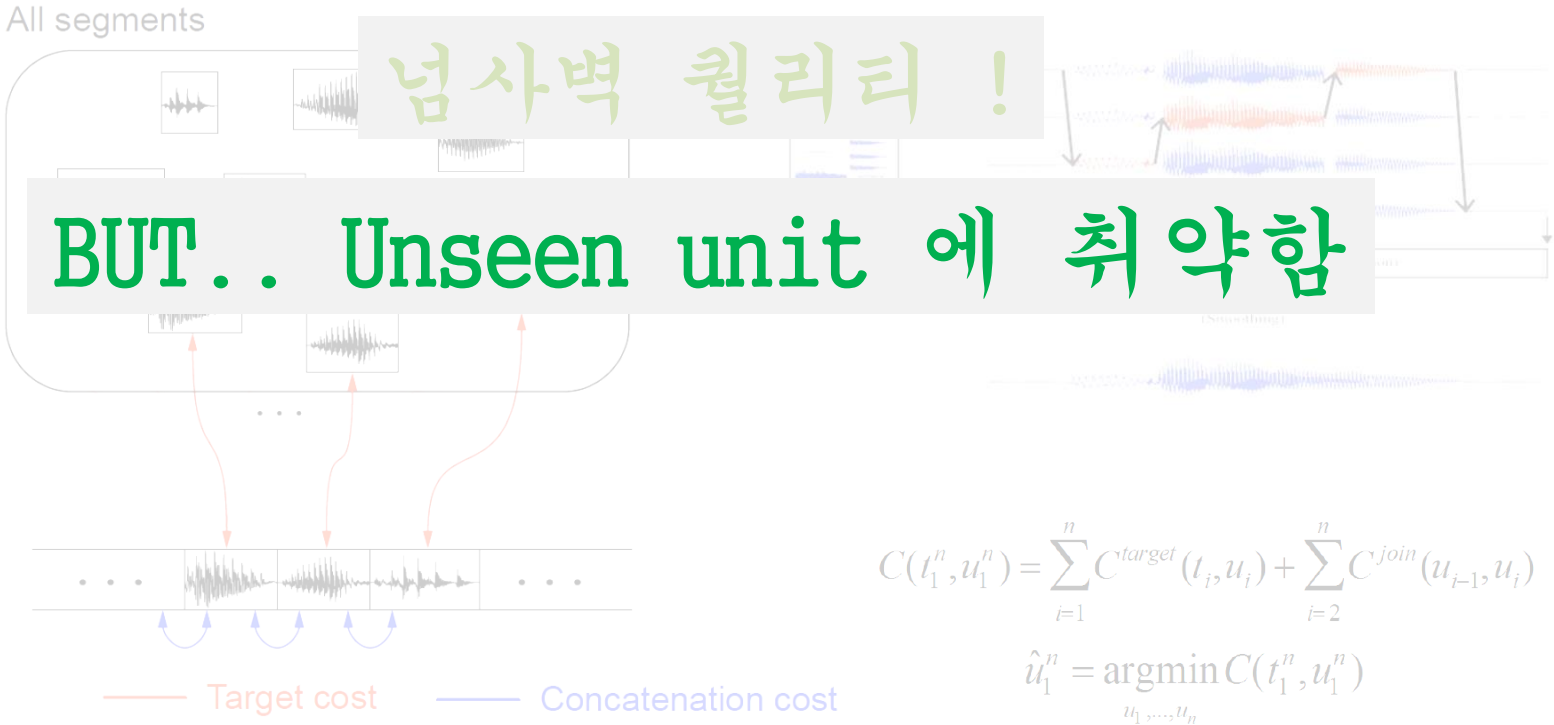
$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^{target}(t_i, u_i) + \sum_{i=2}^{n} C^{join}(u_{i-1}, u_i)$$

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

출처:
T. Keiichi, and H. Zen. "Fundamentals and recent advances in HMM-based speech synthesis." *Tutorial of INTERSPEECH,* 2009.

UNIT-SELECTION SPEECH SYNTHESIS SYSTEM (UTS)

Concatenate speech units (waveform) from a LARGE database



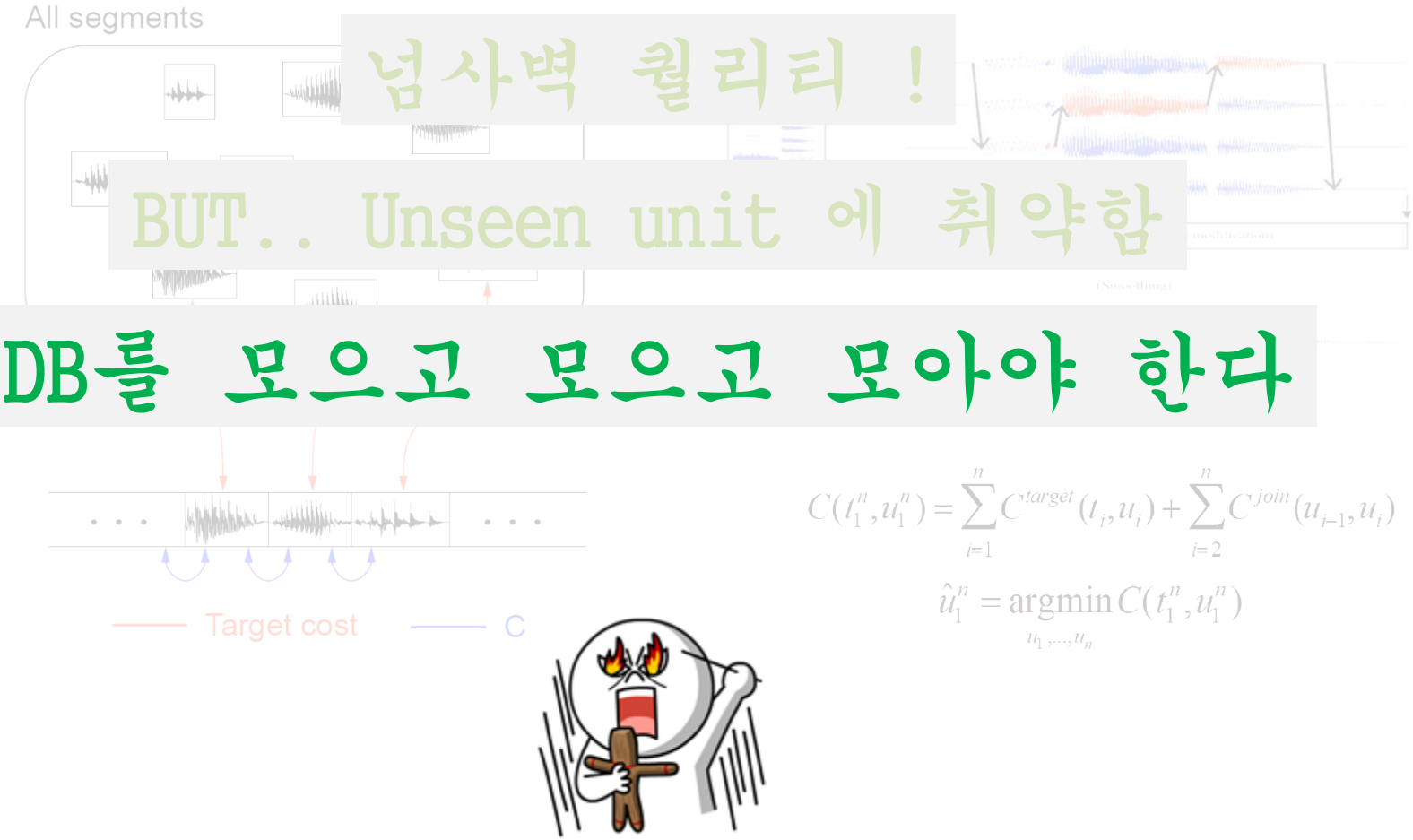
$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^{target}(t_i, u_i) + \sum_{i=2}^{n} C^{join}(u_{i-1}, u_i)$$

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

출처:
T. Keiichi, and H. Zen. "Fundamentals and recent advances in HMM-based speech synthesis." *Tutorial of INTERSPEECH,* 2009.

UNIT-SELECTION SPEECH SYNTHESIS SYSTEM (UTS)

Concatenate speech units (waveform) from a LARGE database



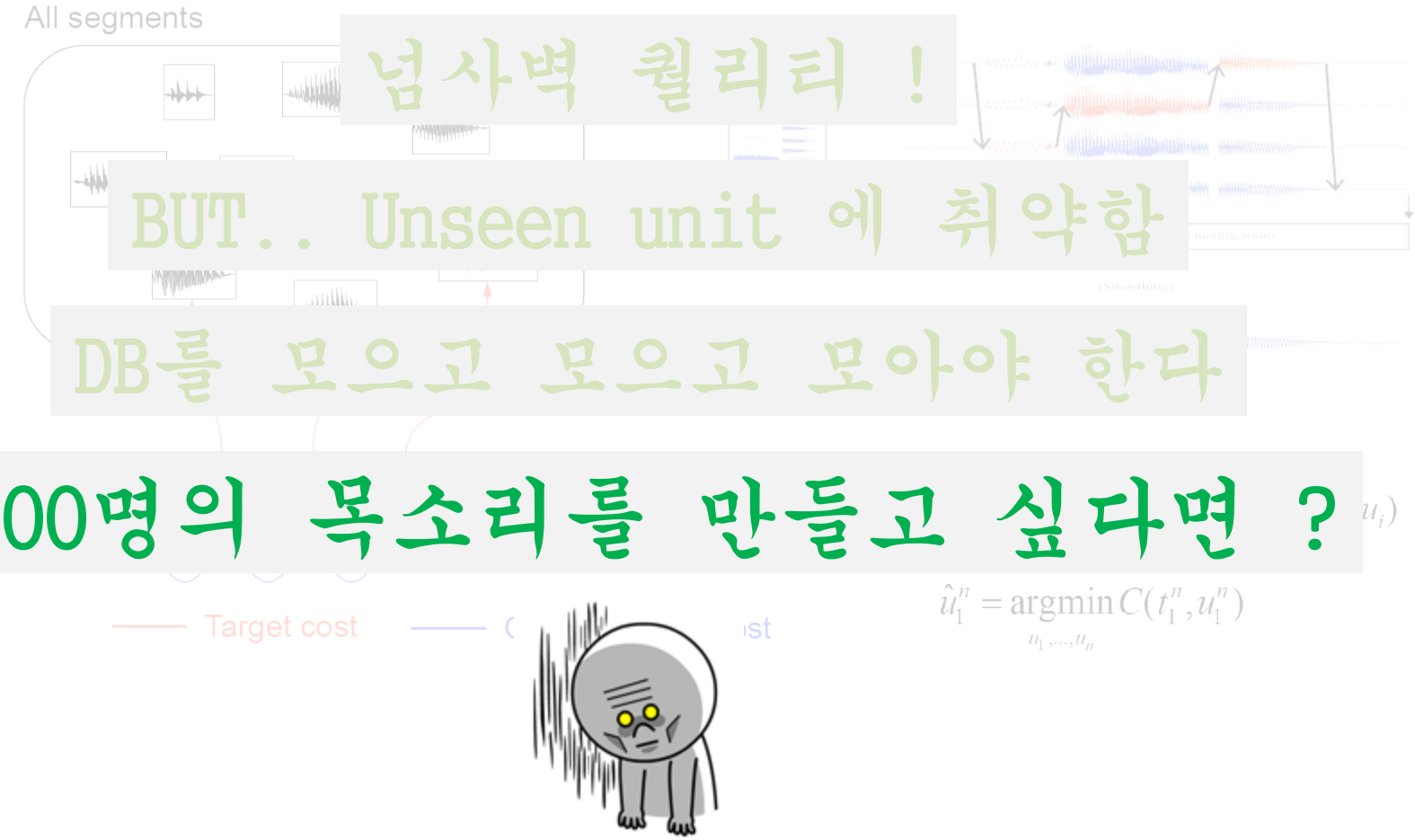
$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^{target}(t_i, u_i) + \sum_{i=2}^{n} C^{join}(u_{i-1}, u_i)$$

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

출처:
T. Keiichi, and H. Zen. "Fundamentals and recent advances in HMM-based speech synthesis." *Tutorial of INTERSPEECH,* 2009.

UNIT-SELECTION SPEECH SYNTHESIS SYSTEM (UTS)

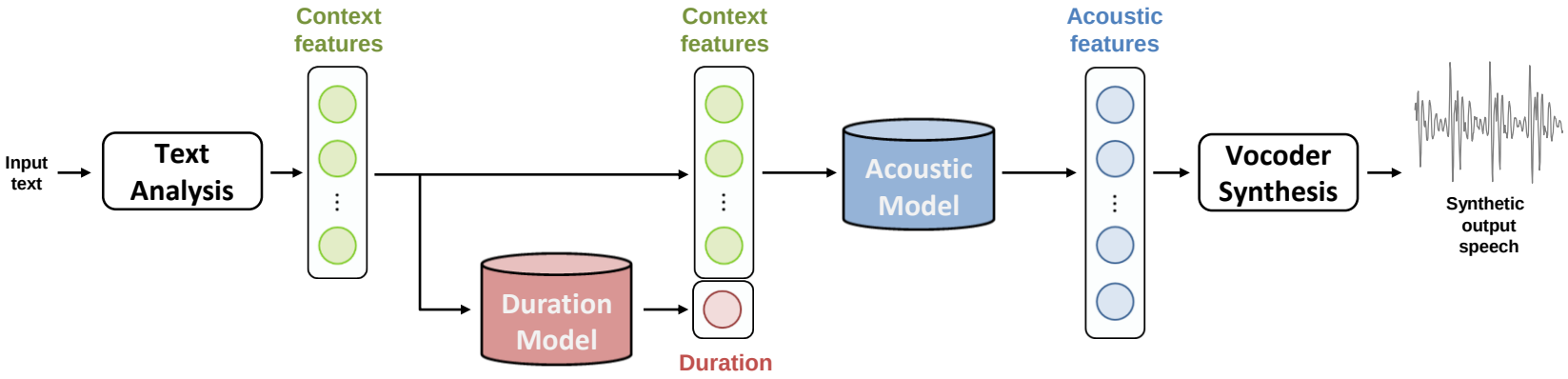
Concatenate speech units (waveform) from a LARGE database



출처:
T. Keiichi, and H. Zen. "Fundamentals and recent advances in HMM-based speech synthesis." *Tutorial of INTERSPEECH,* 2009.

DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

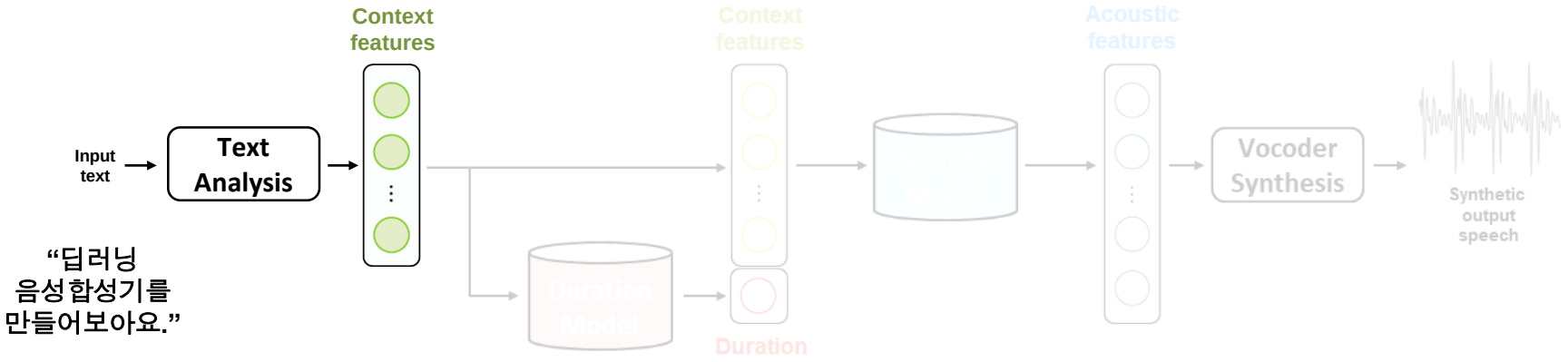
Generate speech parameters from input text and deep neural network



- NLP frontend : text analyzer
- Speech backend : vocoder synthesizer
- Deep learning model : duration & acoustic model

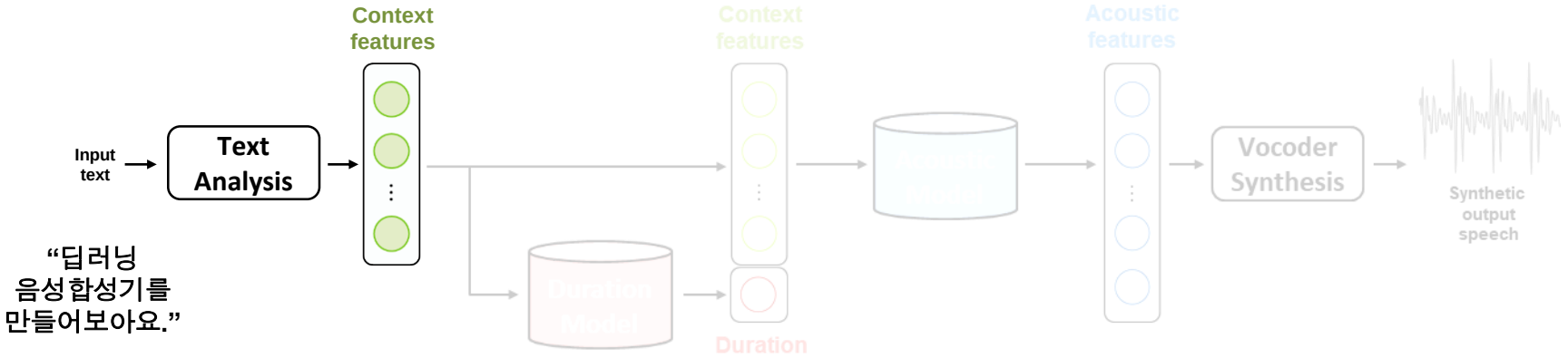
DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Text analyzer



DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Text analyzer



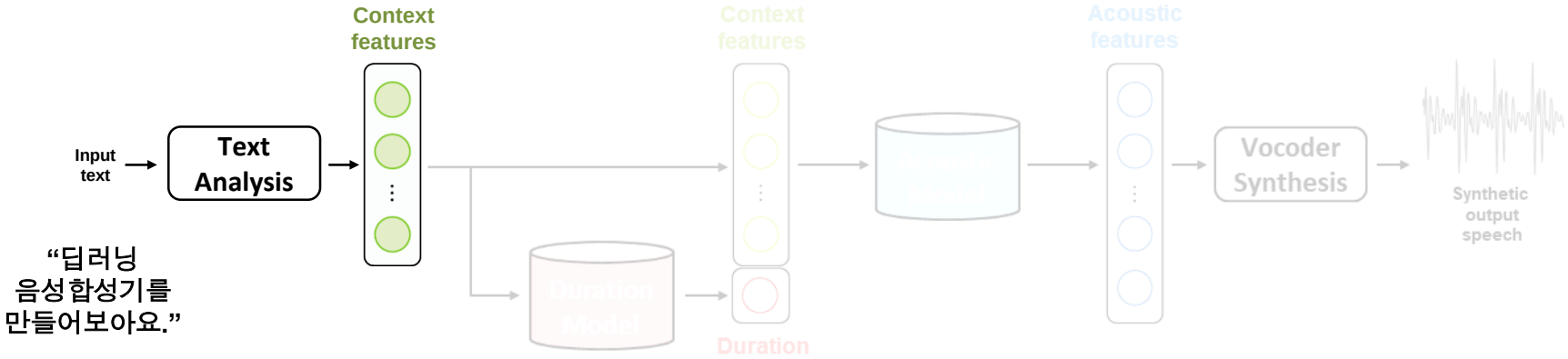
"딥러닝 음성합성기를 만들어보아요."

ㄷㅣㅁㄴㅓㄴㅣㅇ
ㅡㅁㅅㅓㅇㅎㅏㅂㅆㅓㅇㄱㅣㄹㄹ
ㅁㅏㄴㄷㅡㄹㅓㅂㅗㅏㅛ

Text Analysis

DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Text analyzer

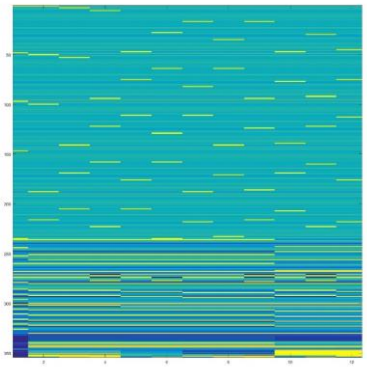


"딥러닝
음성합성기를
만들어보아요."

(묵음)ㄷㅣㅁㄴㅓㄴㅣㅇㅡㅁㅅ...

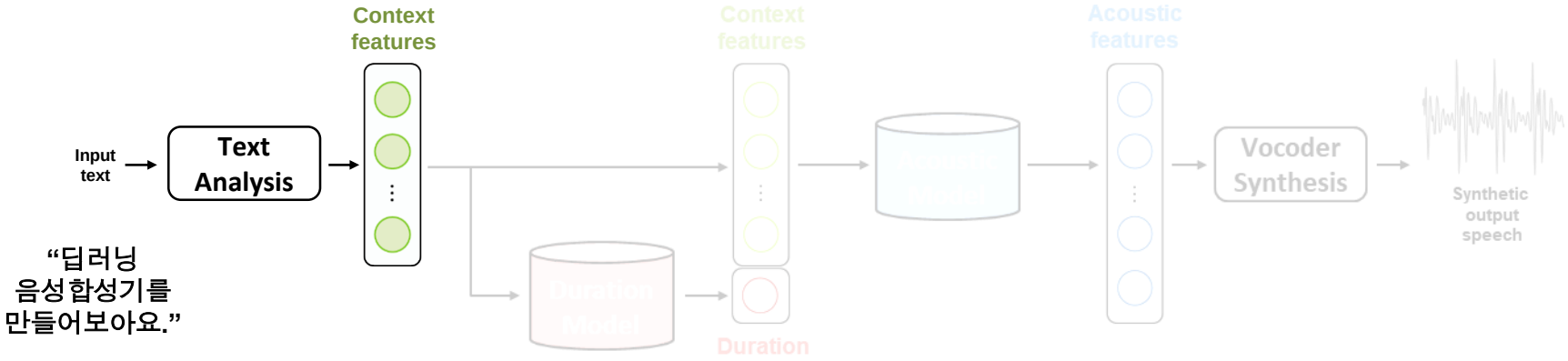
ㄷㅣㅁㄴㅓㄴㅣㅇ
ㅡㅁㅅㅓㅇㅎㅏㅂㅆㅓㅇㄱㅣㄹㄹ
ㅁㅏㄴㄷㅡㄹㅓㅂㅗㅏㅛ

Text Analysis



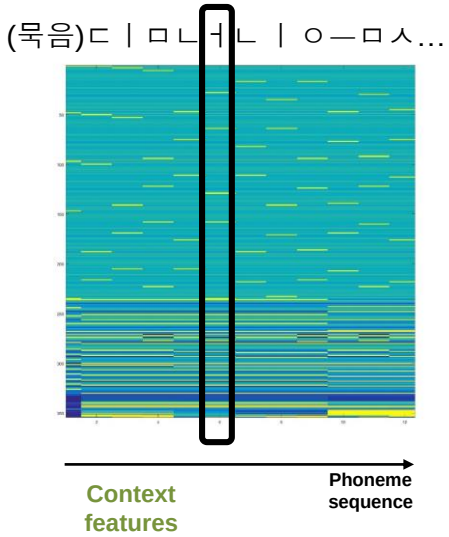
DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Text analyzer



ㄷㅣㅁㄴㅓㄴㅣㅇ
ㅡㅁㅅㅓㅇㅎㅏㅂㅆㅓㅇㄱㅣㄹㄹ
ㅁㅏㄴㄷㅡㄹㅓㅂㅗㅏㅛ

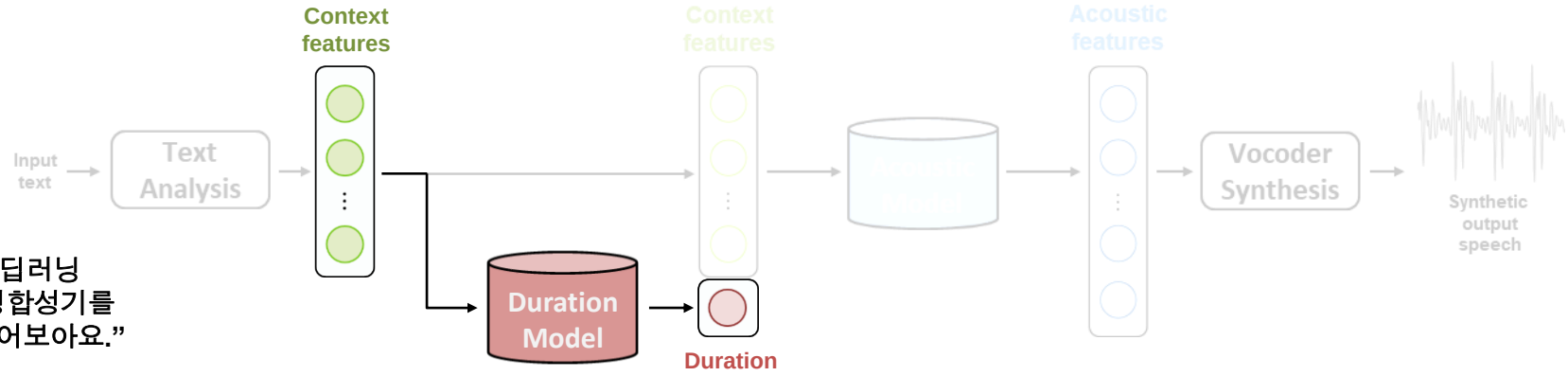
Text Analysis



Phoneme ID
Break
Punctuation
# of word in the current sentence
Position of word in the current sentence
# of phoneme in the current word
Position of phoneme in the current word
.
.
.

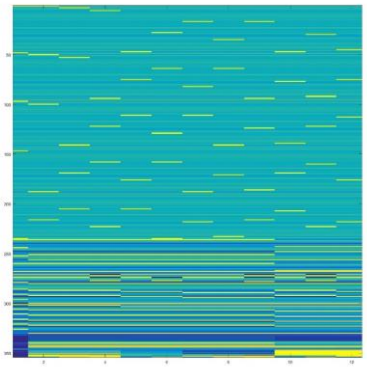
DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Duration model



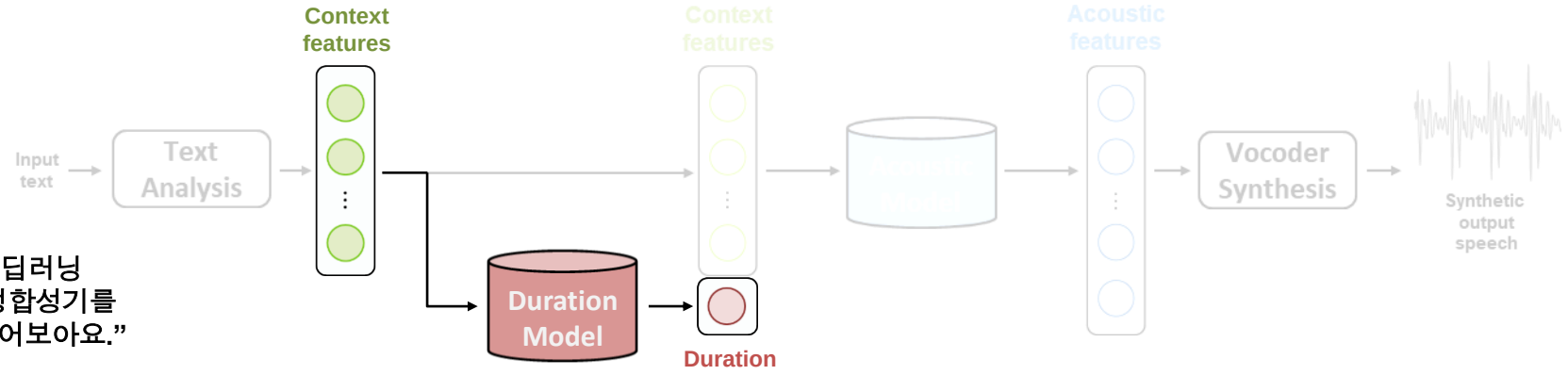
"딥러닝 음성합성기를 만들어보아요."

(묵음)ㄷㅣㅁㄴㅓㄴㅣㅇㅡㅁㅅ...

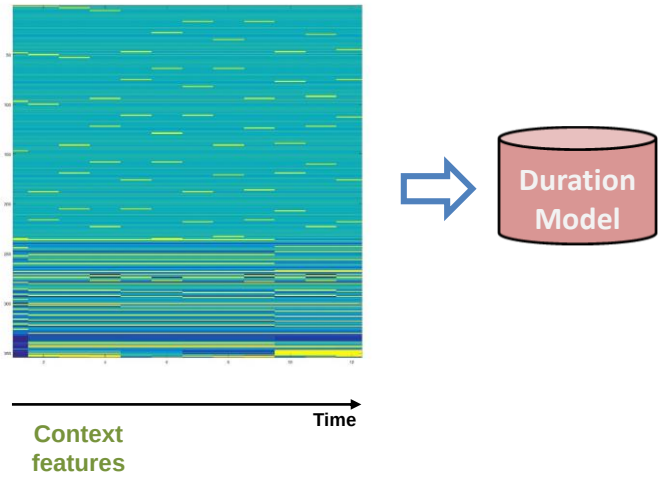


DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Duration model

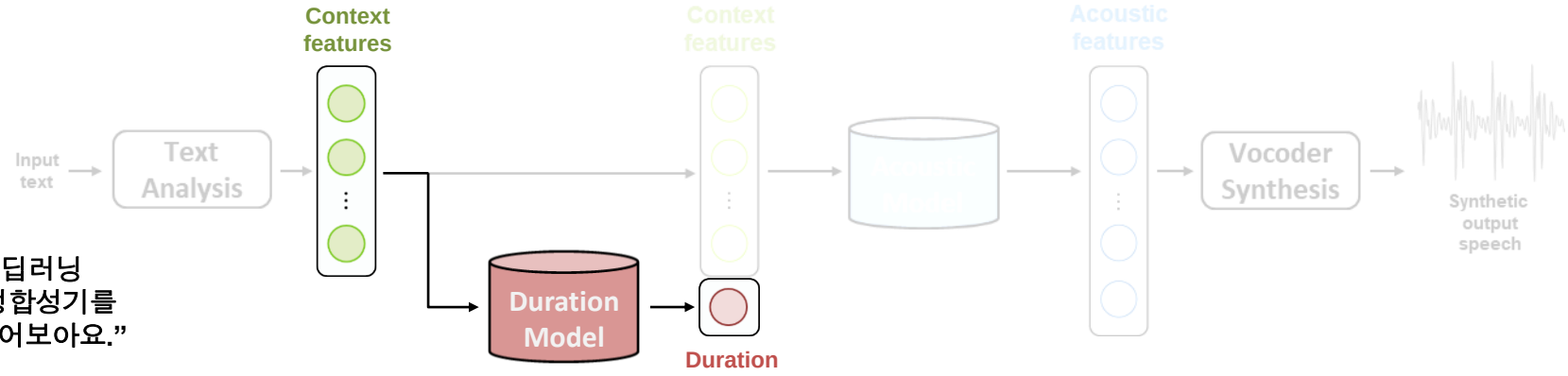


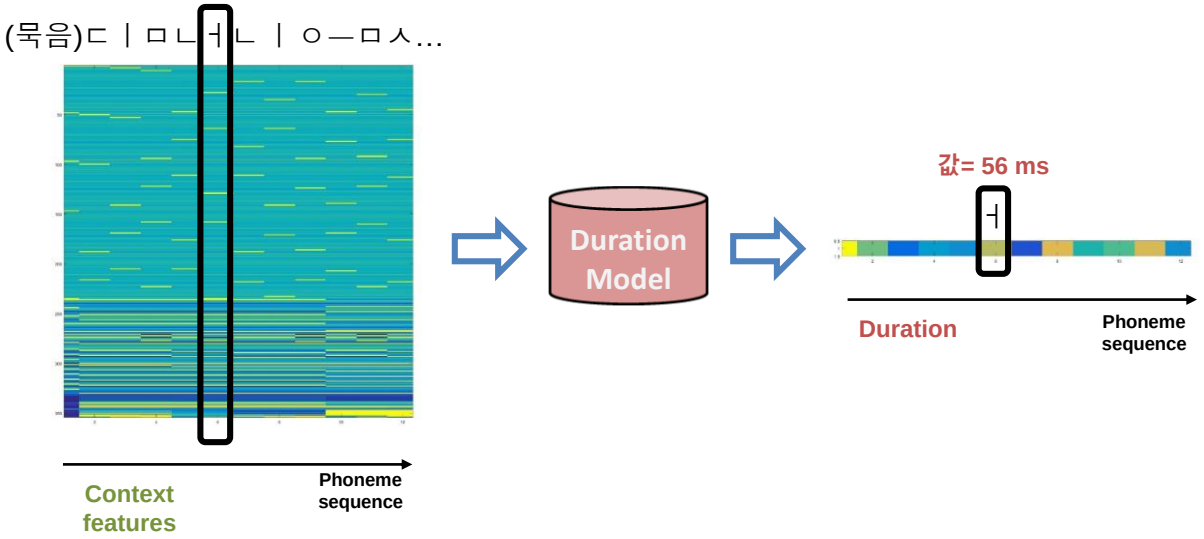
(묵음)ㄷㅣㅁㄴㅓㄴㅣㅇㅡㅁㅅ…



DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

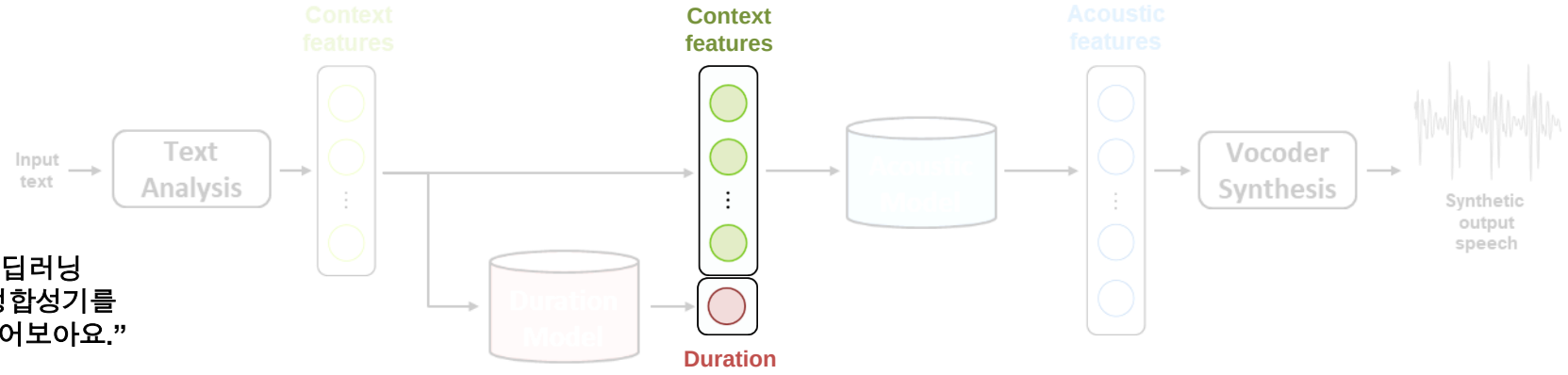
Duration model





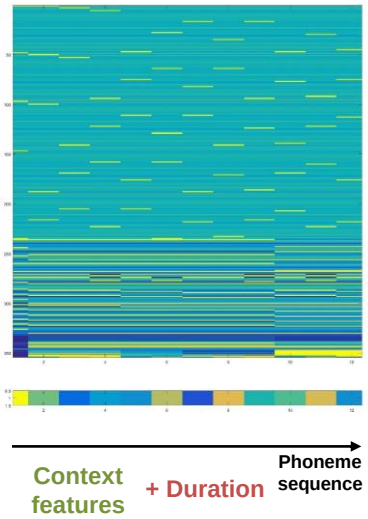
DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Linguistic upsampler



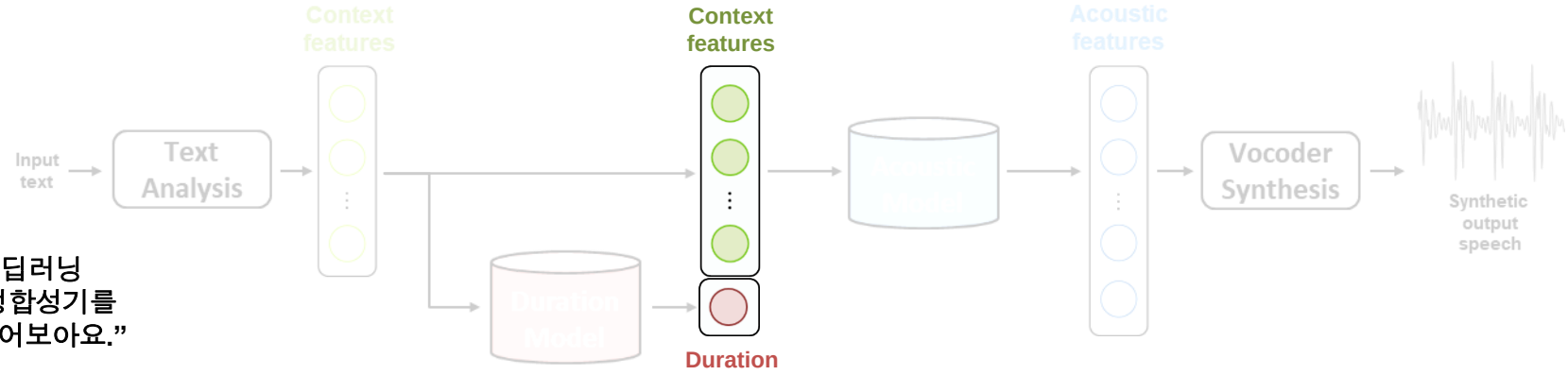
"딥러닝 음성합성기를 만들어보아요."

(묵음)ㄷㅣㅁㄴㅓㄴㅣㅇㅡㅁㅅ...



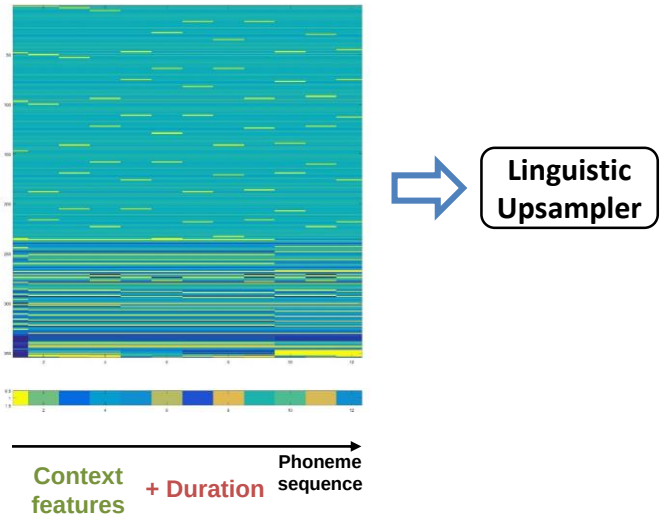
DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Linguistic upsampler



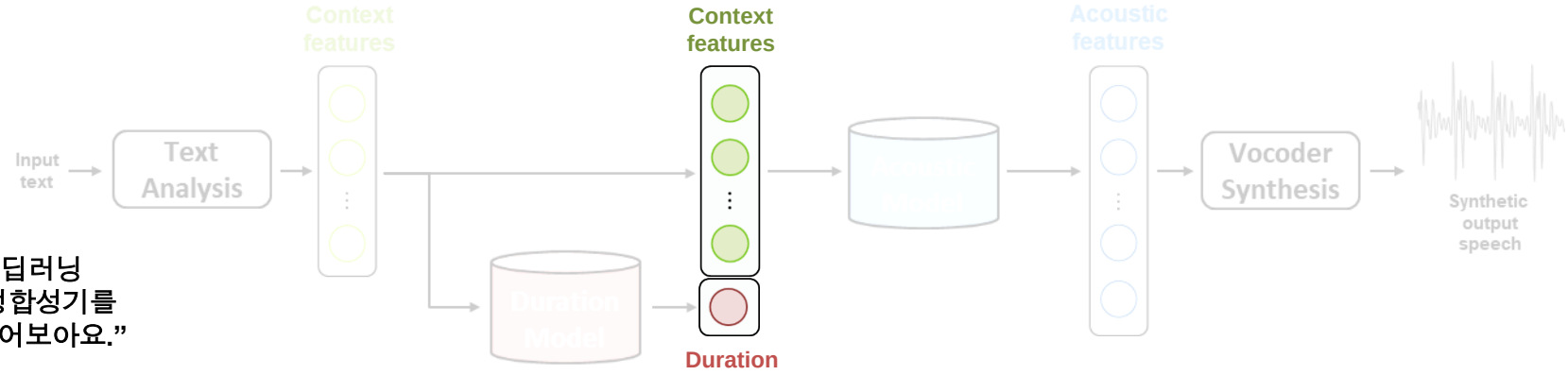
"딥러닝 음성합성기를 만들어보아요."

(묵음)ㄷㅣㅁㄴㅓㄴㅣㅇㅡㅁㅅ…

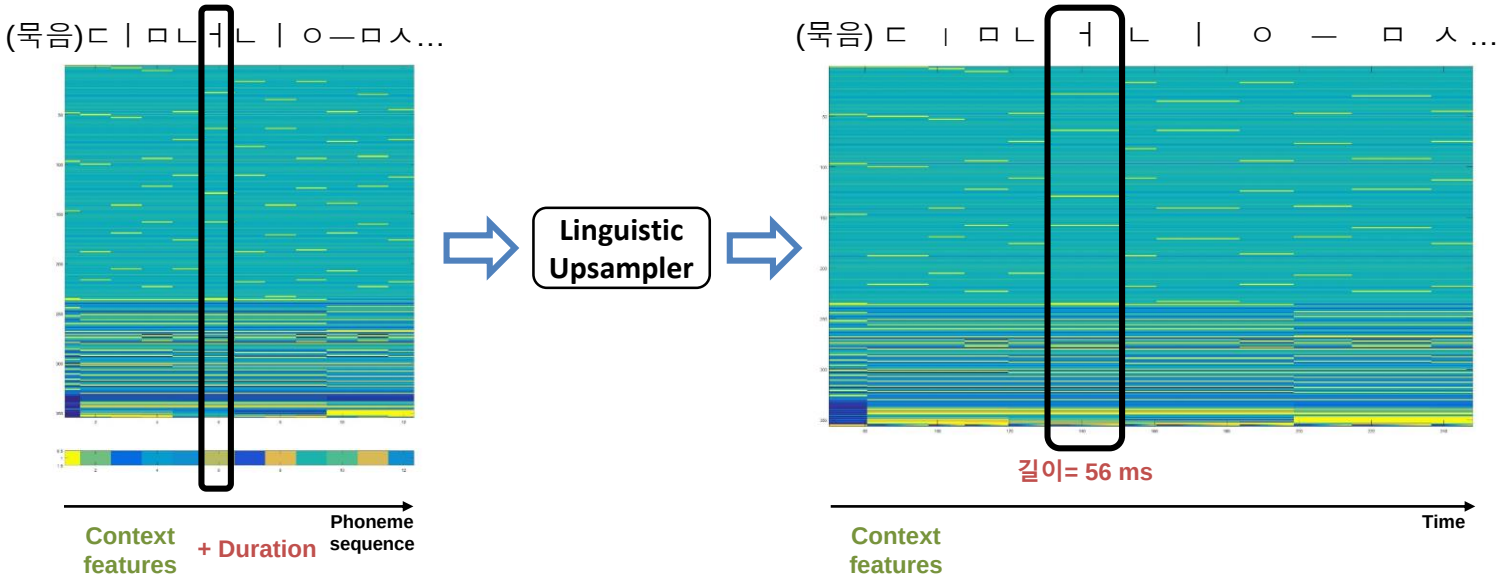


DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Linguistic upsampler

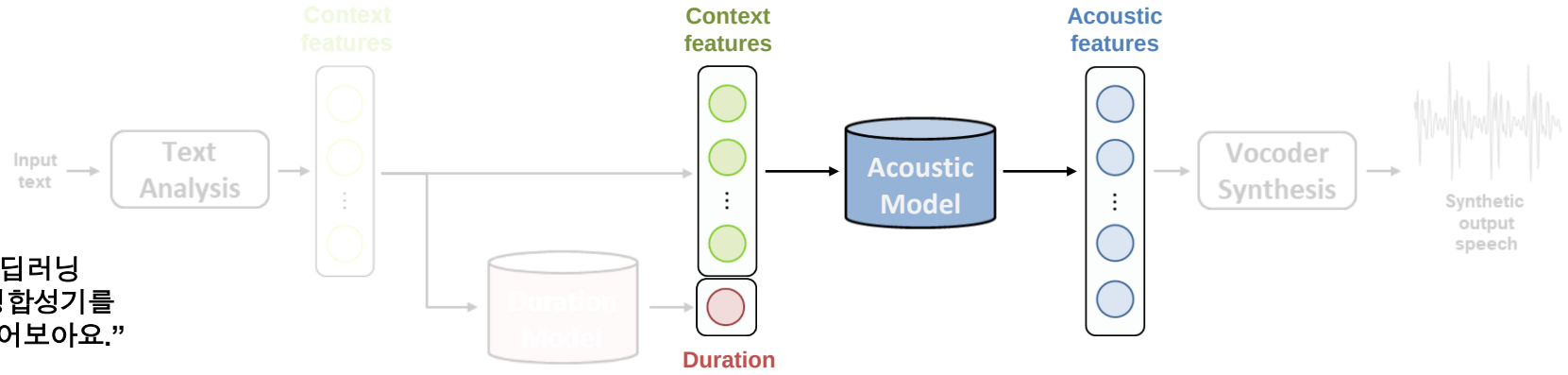


"딥러닝 음성합성기를 만들어보아요."



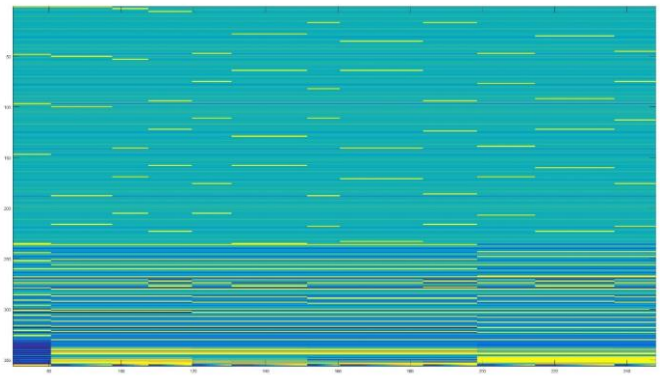
DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Acoustic model



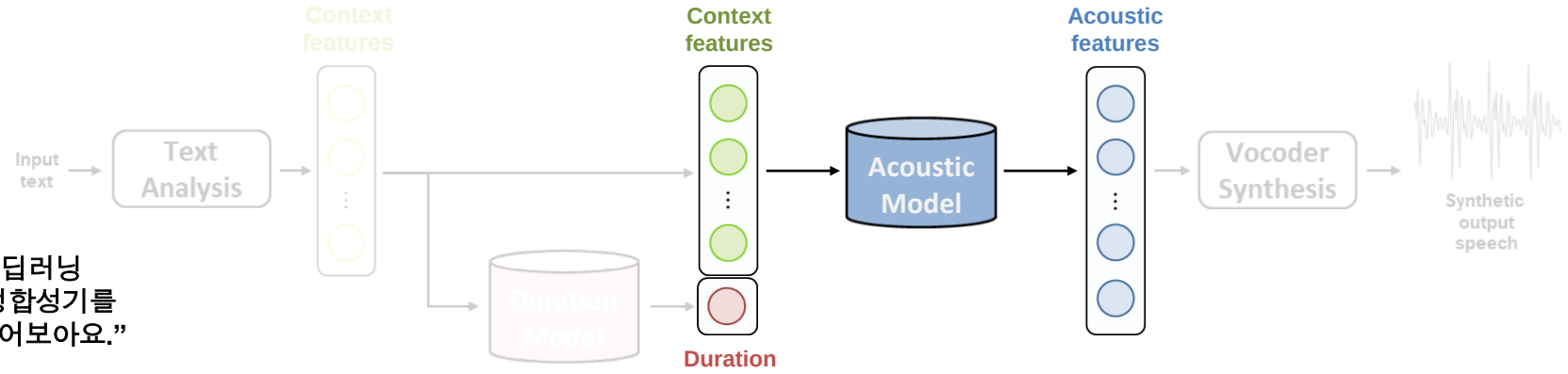
"딥러닝 음성합성기를 만들어보아요."

(묵음) ㄷ ㅣ ㅁㄴ ㅓ ㄴ ㅣ ㅇ ㅡ ㅁ ㅅ …



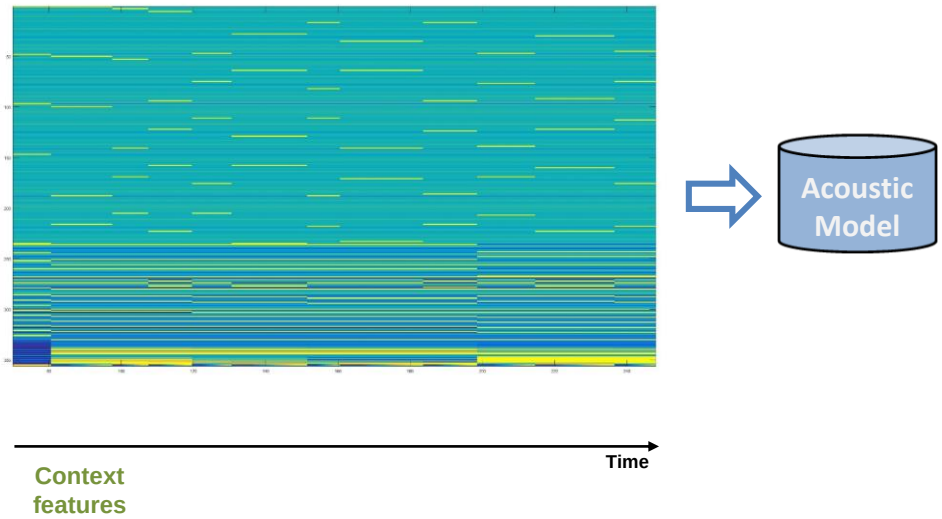
DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Acoustic model



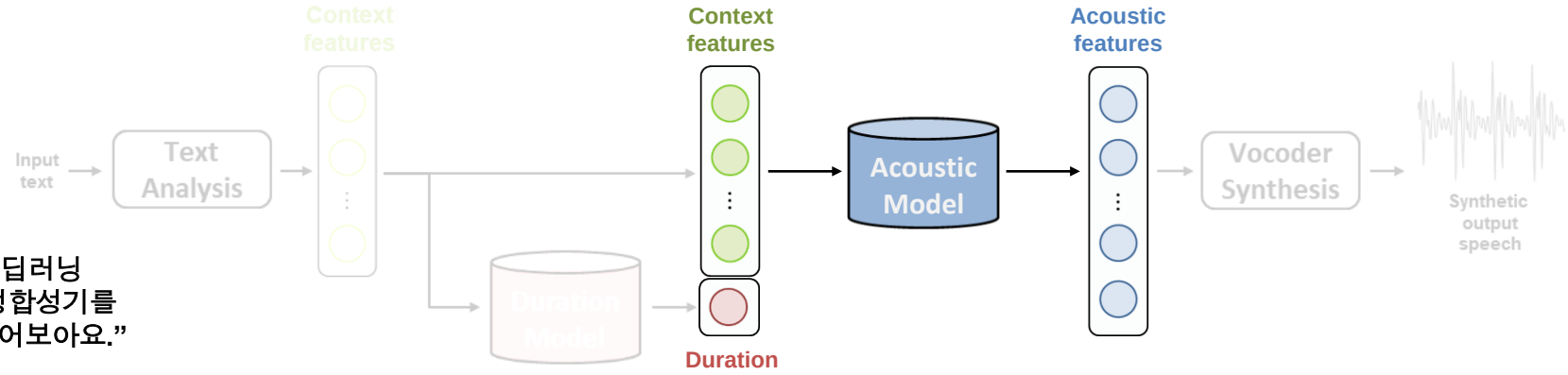
"딥러닝 음성합성기를 만들어보아요."

(묵음) ㄷ ㅣ ㅁㄴ ㅓ ㄴ ㅣ ㅇ ㅡ ㅁ ㅅ …

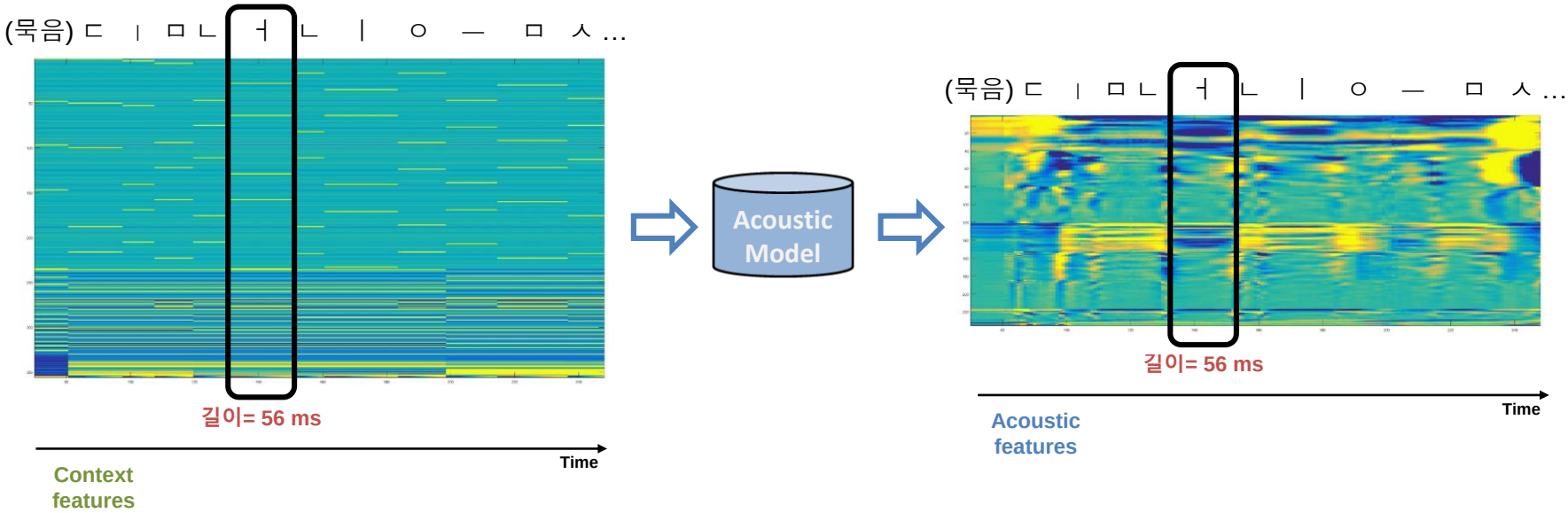


DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Acoustic model

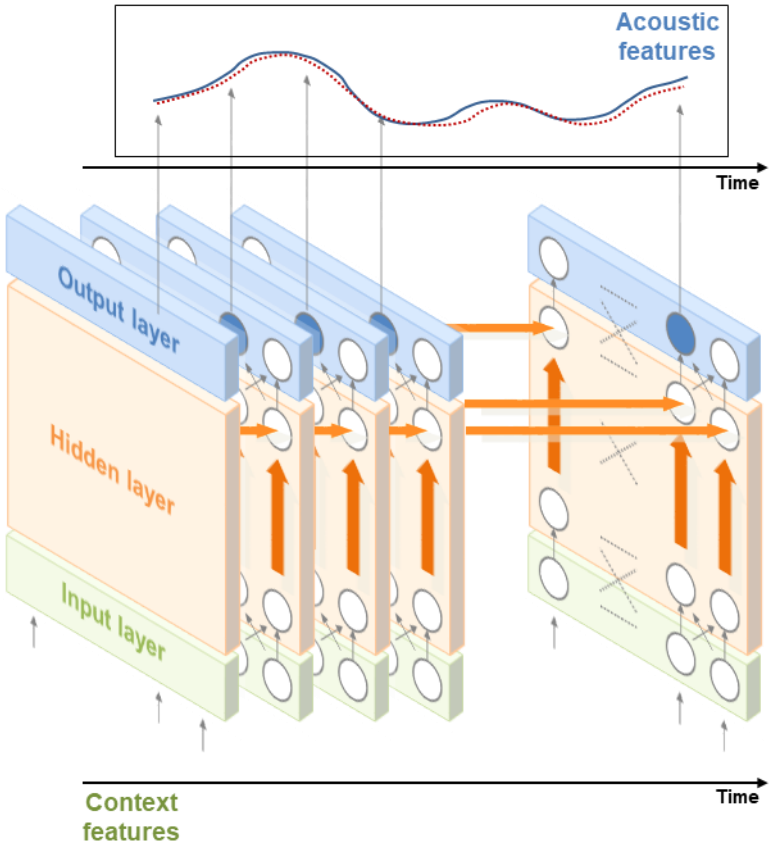


"딥러닝 음성합성기를 만들어보아요."



DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

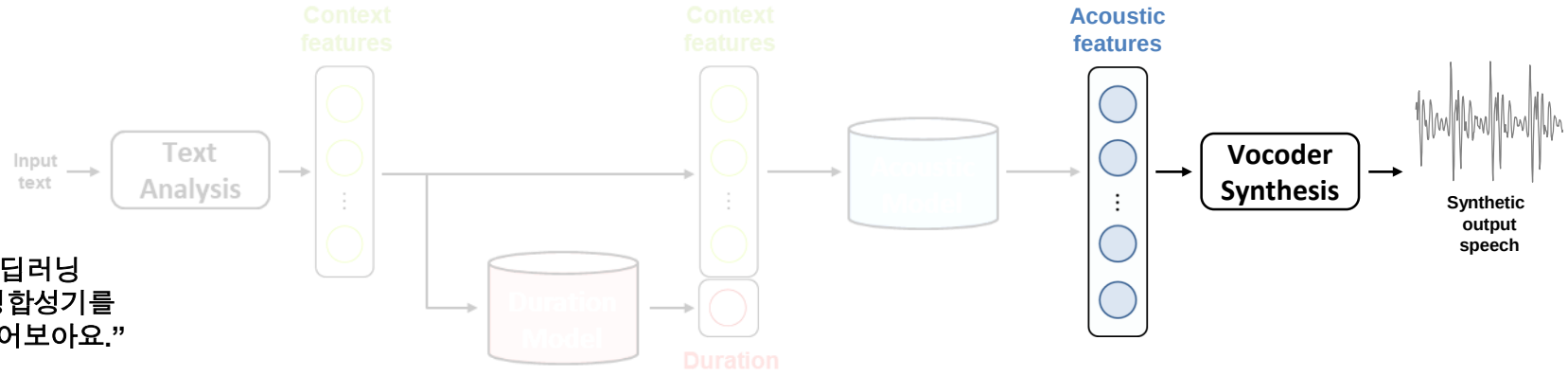
Multiple feed-forward and LSTM layers

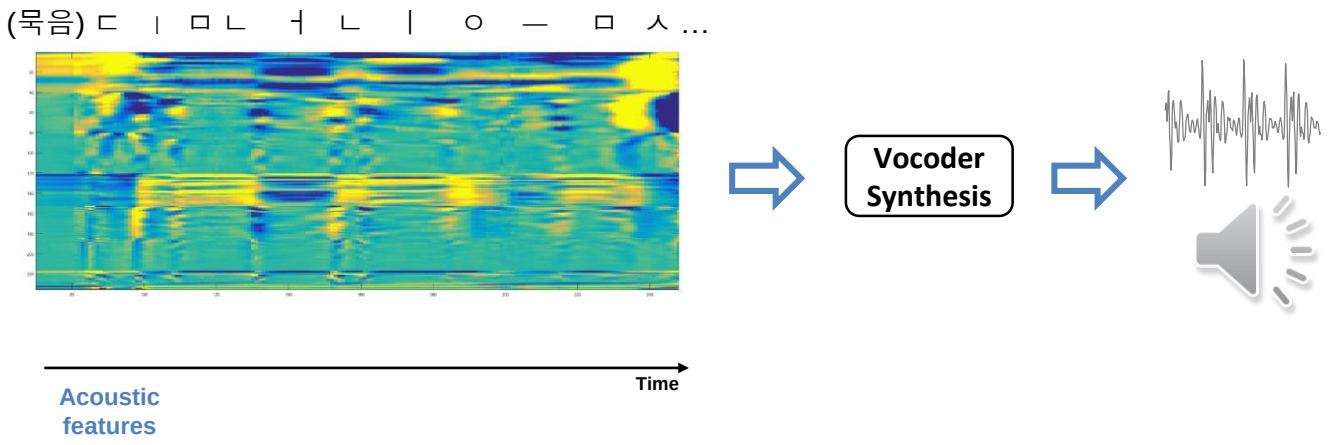


$$\text{AcousticFeature} = F_{DNN}(\text{ContextFeature})$$

DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

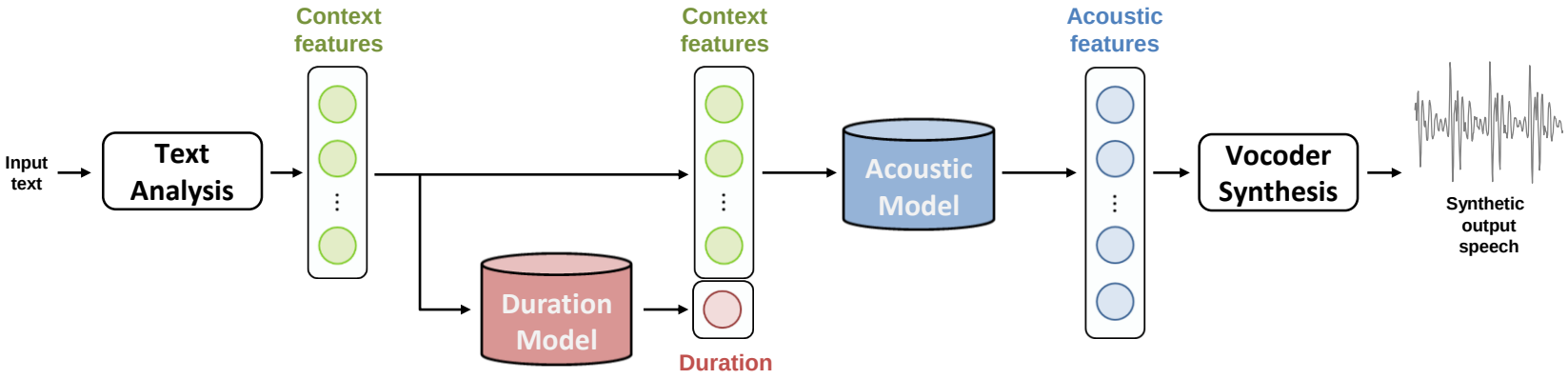
Vocoder synthesis





DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

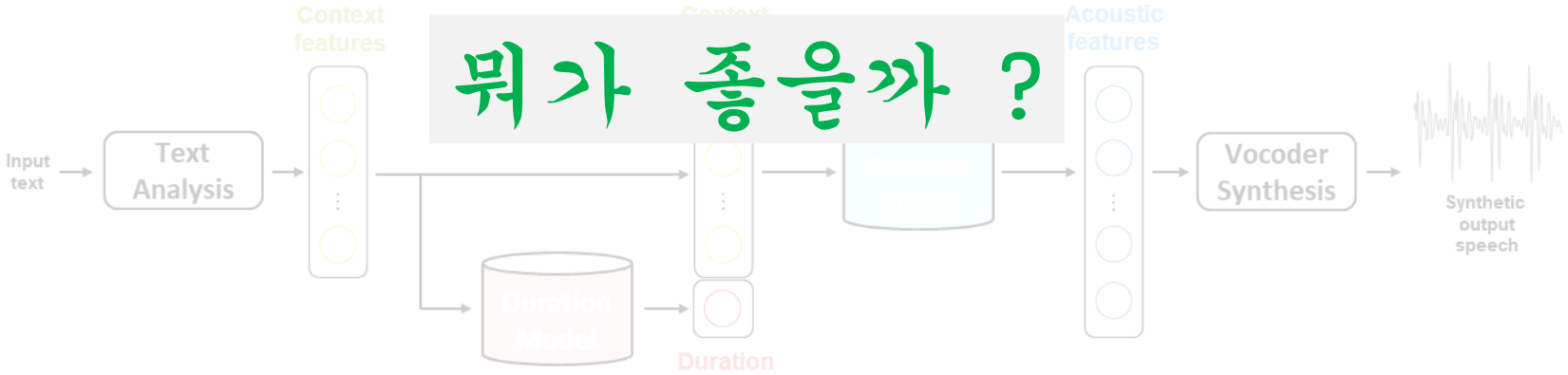
Generate speech parameters from input text and deep neural network



- NLP frontend : text analyzer
- Speech backend : vocoder synthesizer
- Deep learning model : duration & acoustic model

DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Generate speech parameters from input text and deep neural network



DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Generate speech parameters from input text and deep neural network

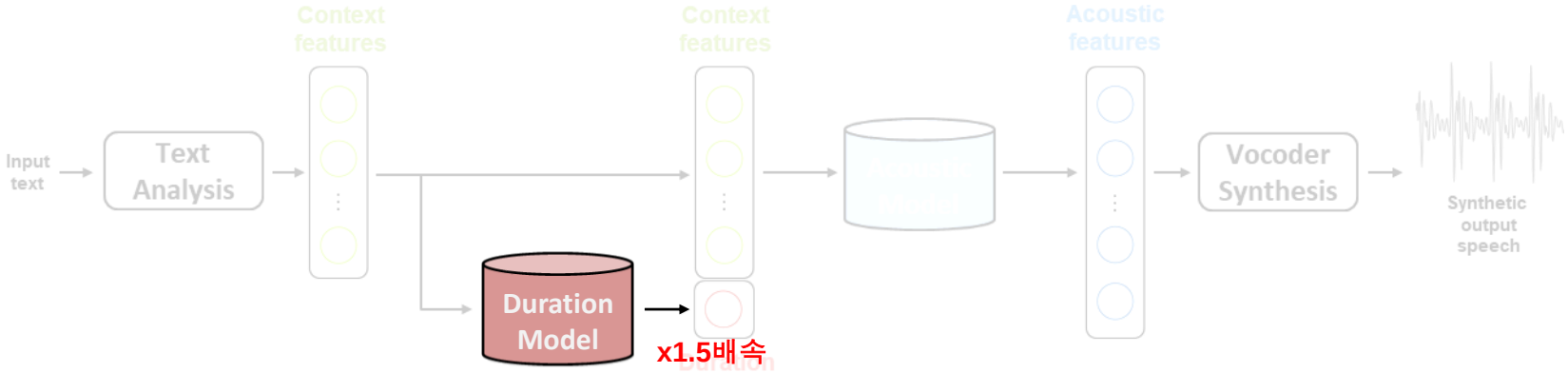
DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Generate speech parameters from input text and deep neural network

Flexible to change voice characteristics

DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

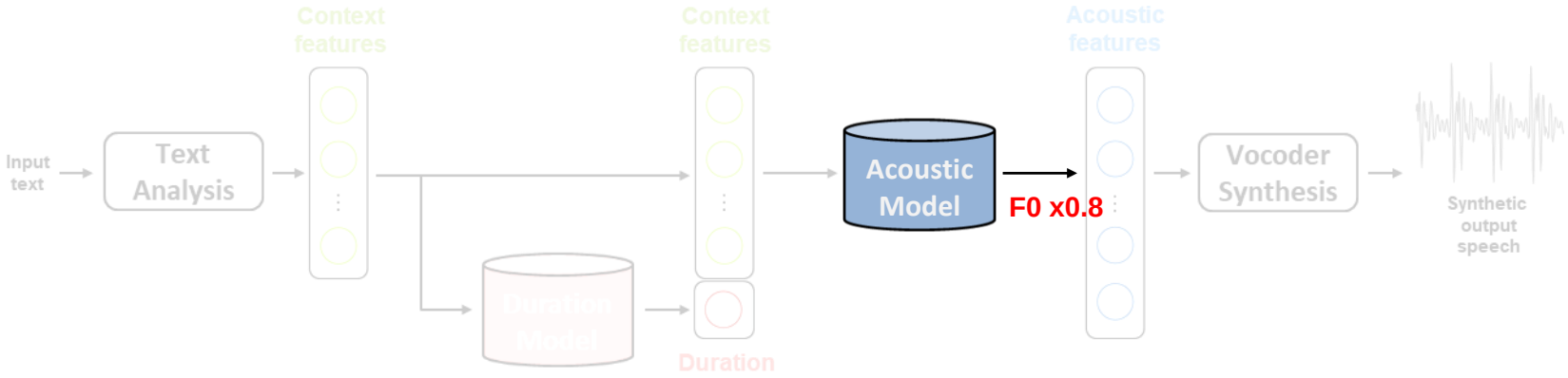
Applications : speech rate



간장 공장 공장장은 강 공장장이고, 된장 공장 공장장은 공 공장장이다.	내가 그린 기린 그림은 긴 기린 그림이고, 네가 그린 기린 그림은 안 긴 기린 그림이다.	좌로인정 우로인정 앞구르기인정 인정올리지말고 인정내려 인정안해서 후회한다면 후회 할 시간을 후회하는 각이고요.	오졌따리 오져따 쿵쿵따리쿵쿵따 산기슭이 인정하는 바이고요 슭곰발이 인정하는 바입니다.
x1.5배속			

DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Applications : speech tone



간장 공장 공장장은
강 공장장이고,
된장 공장 공장장은
공 공장장이다.

내가 그린 기린 그림은
긴 기린 그림이고,
네가 그린 기린 그림은
안 긴 기린 그림이다.

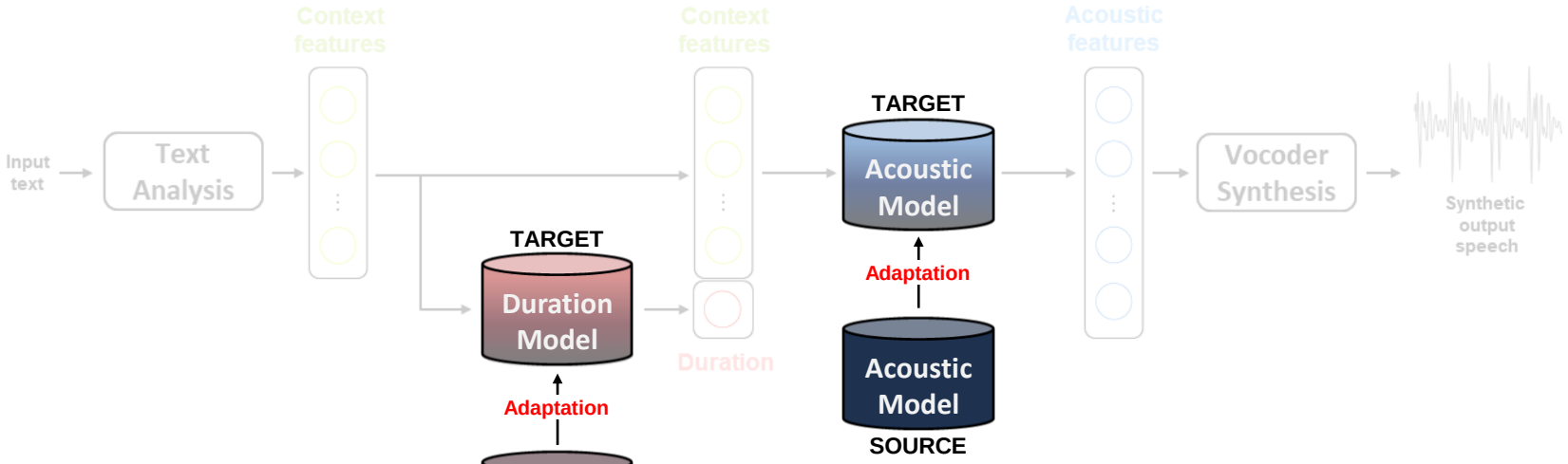
좌로인정 우로인정
앞구르기인정
인정올리지말고 인정내려
인정안해서 후회한다면
후회 할 시간을
후회하는 각이고요.

오졌따리 오져따
쿵쿵따리쿵쿵따
산기슭이
인정하는 바이고요
슭곰발이
인정하는 바입니다.

F0 x0.8

DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

Applications : speaker adaptation



간장 공장 공장장은
강 공장장이고,
된장 공장 공장장은
공 공장장이다.

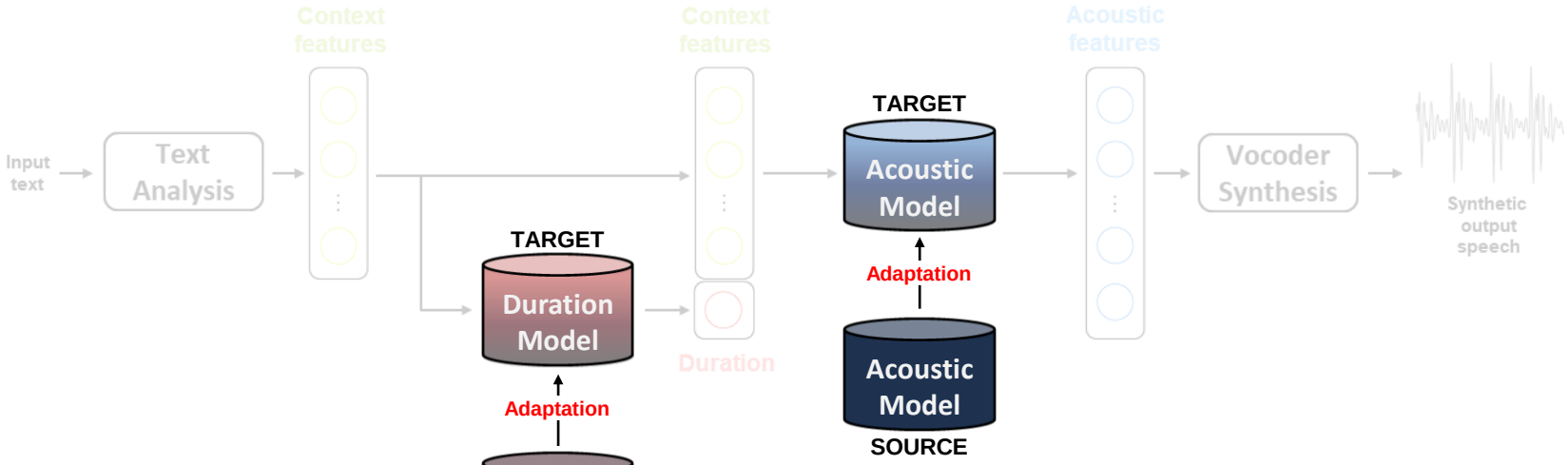
네가 그린 기린 그림은
안 긴 기린 그림이다.

좌로인정 우로인정
앞구르기인정
인정올리지말고 인정내려
인정안해서 후회한다면
후회 할 시간을
후회하는 각이고요.

오졌따리 오져따
쿵쿵따리쿵쿵따
산기슭이
인정하는 바이고요
슭곰발이
인정하는 바입니다.

DEEP-LEARNING SPEECH SYNTHESIS SYSTEM (DTS)

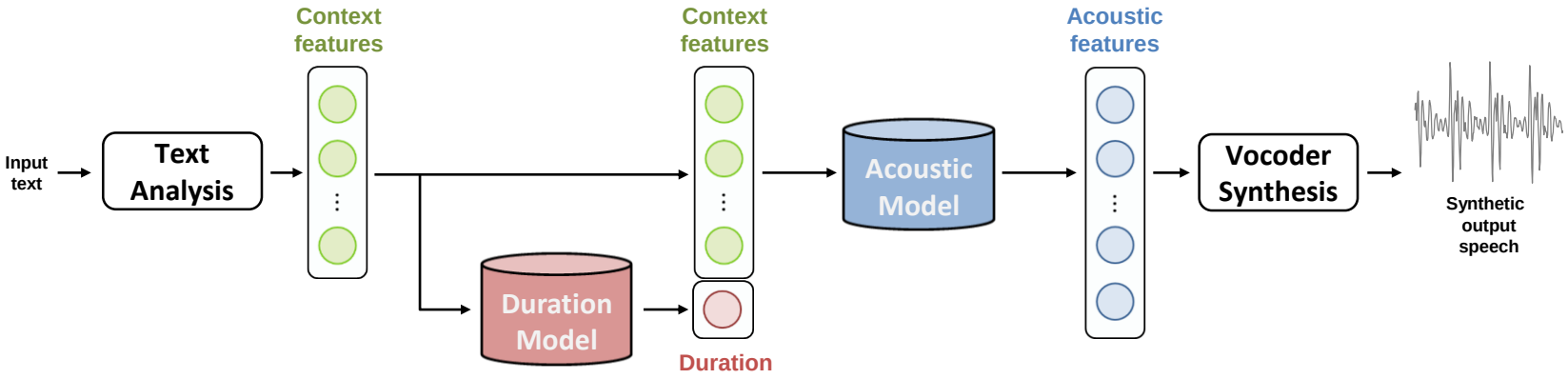
Applications : speaker adaptation



x1.5배속 F0 x0.8 x1.5배속 F0 x0.8

SUMMARY

Generate speech parameters from input text and deep neural network



- NLP frontend : text analyzer
- Speech backend : vocoder synthesizer
- Deep learning model : duration & acoustic model

- Advantages
  - Flexible to change voice characteristics

질문있어요 ?

CLOVA 채용부스로 오세요.