CLOVA

딥러닝 음성 합성 기초편

송은우 / HDTS

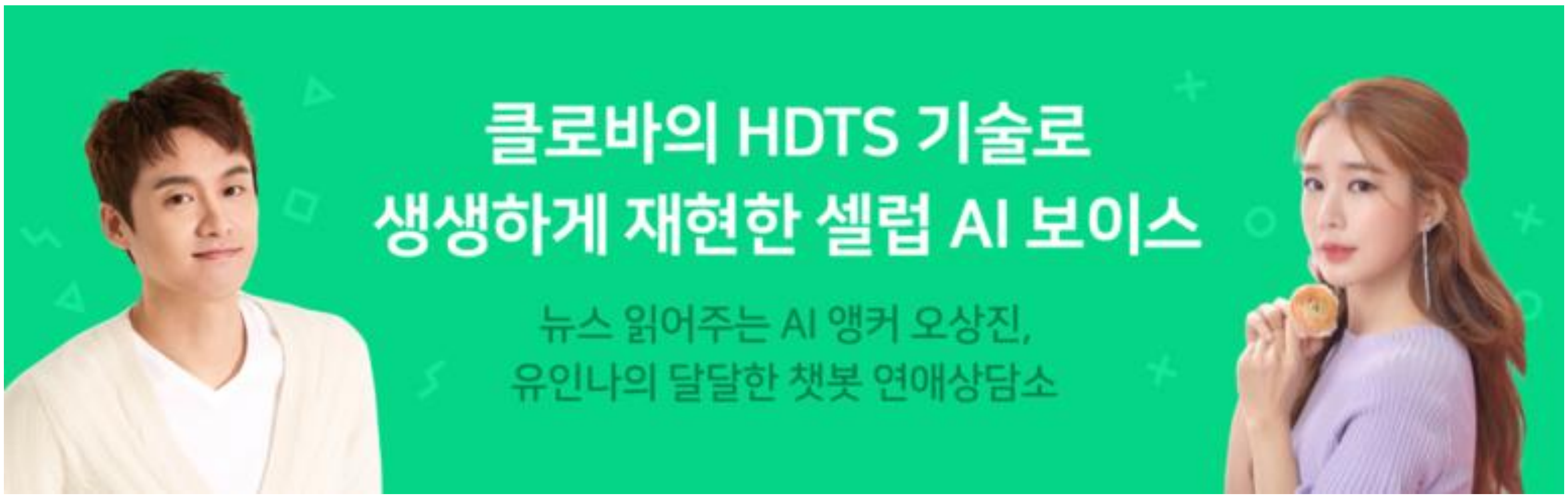
CLOVA

딥러닝 음성 합성 기초편

송은우 / HDTS

Introduction

Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.



Introduction

Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.

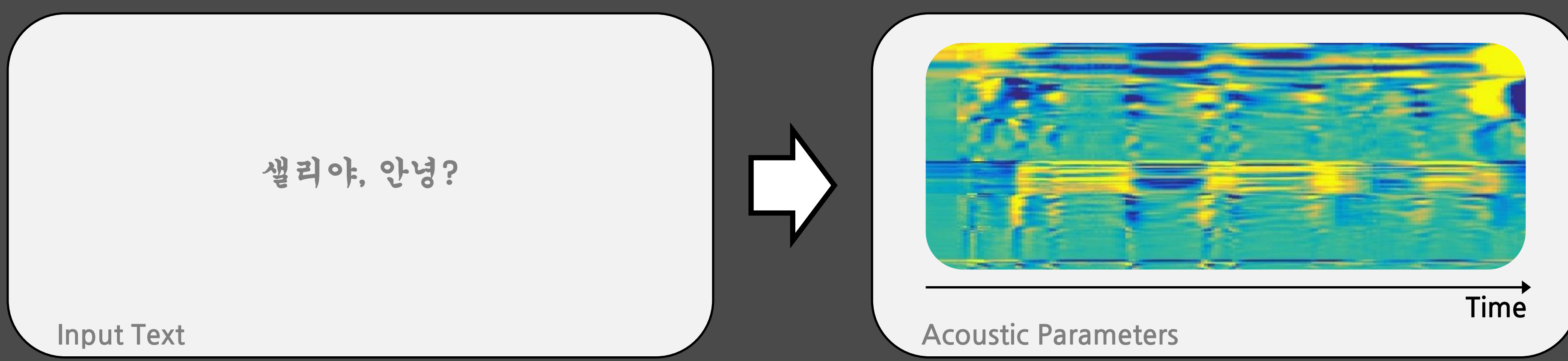
DNN TTS = Acoustic model + Vocoder

Introduction

Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.

톤의 높낮이, 음색, 어조, 강세 등

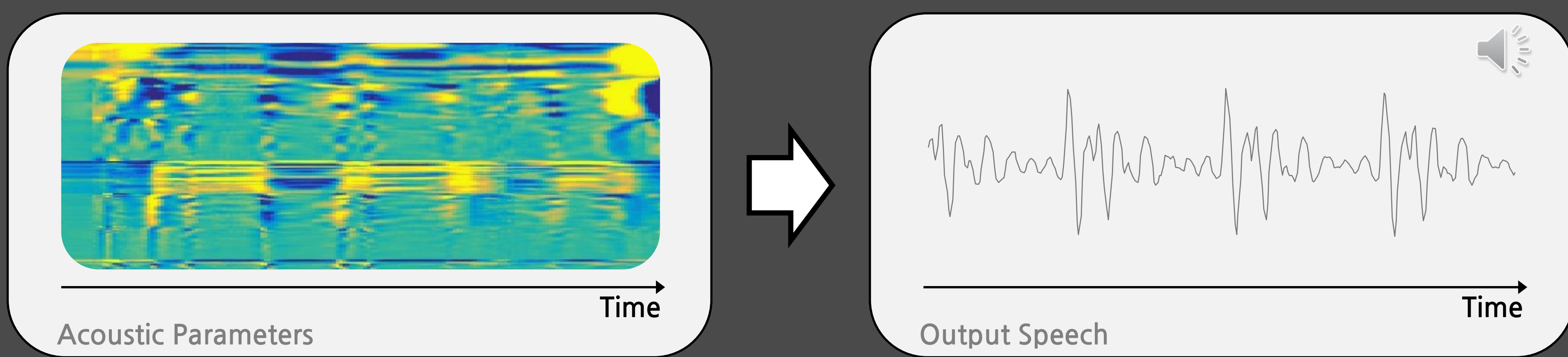
텍스트에서 **Acoustic Parameter** 를 추정



Introduction

Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.

Acoustic Parameter 에서 음성 신호를 생성



Introduction

Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.

Acoustic Parameter 에서 음성 신호를 추정

본 발표에서는 TTS 엔진의 핵심 요소인

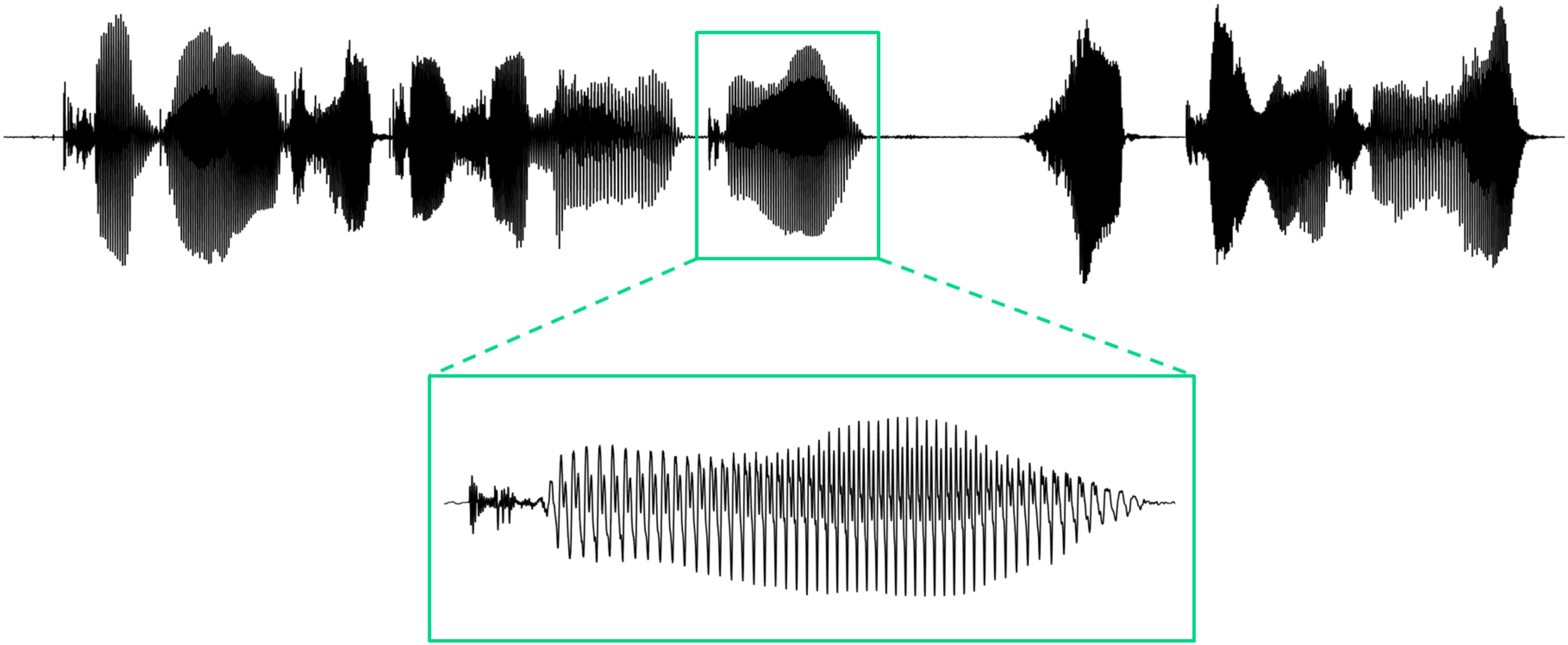
Acoustic Model & Vocoder 기술을 정리하고자 합니다.

Speech fundamentals

What is speech ?

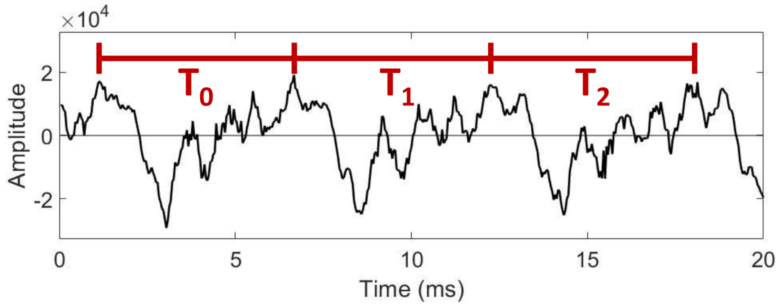
Input Text → Acoustic Model → Acoustic Parameters → Vocoder → Output Speech

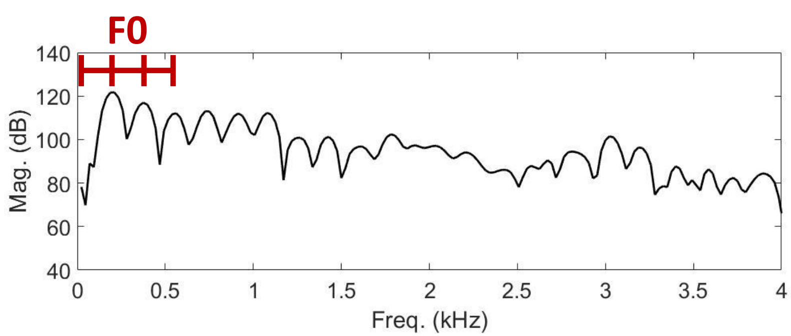
Speech waveform



Pitch period

음성의 **주기성**을 나타내는 파라미터: 음성의 **톤**을 결정합니다 (ex. 하이톤, 중저음).





Pitch period = $T\_0 \approx T\_1 \approx T\_2$

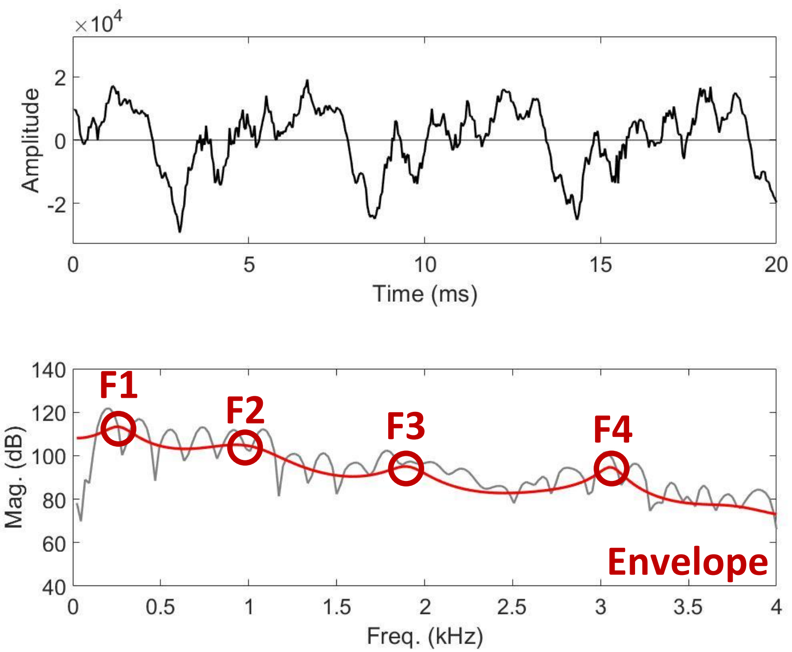
- Long-term period of speech (time-domain)

Fundamental frequency (F0) = $1/T\_0$

- 1 / PP (frequency-domain)
- Female voice: Ave. 200 Hz
- Male voice : Ave. 100 Hz

Formant frequency

음색을 나타내는 파라미터: 음성의 **발음**을 결정합니다 (ex. 아 / 에 / 이 / 오 / 우).



Pitch period = $T_0 \approx T_1 \approx T_2$

- Long-term period of speech (time-domain)

Fundamental frequency (F0) = $1/T_0$

- 1 / PP (frequency-domain)
- Female voice: Ave. 200 Hz
- Male voice : Ave. 100 Hz

Formant frequency (F1, F2, …)

- Vocal tract resonance

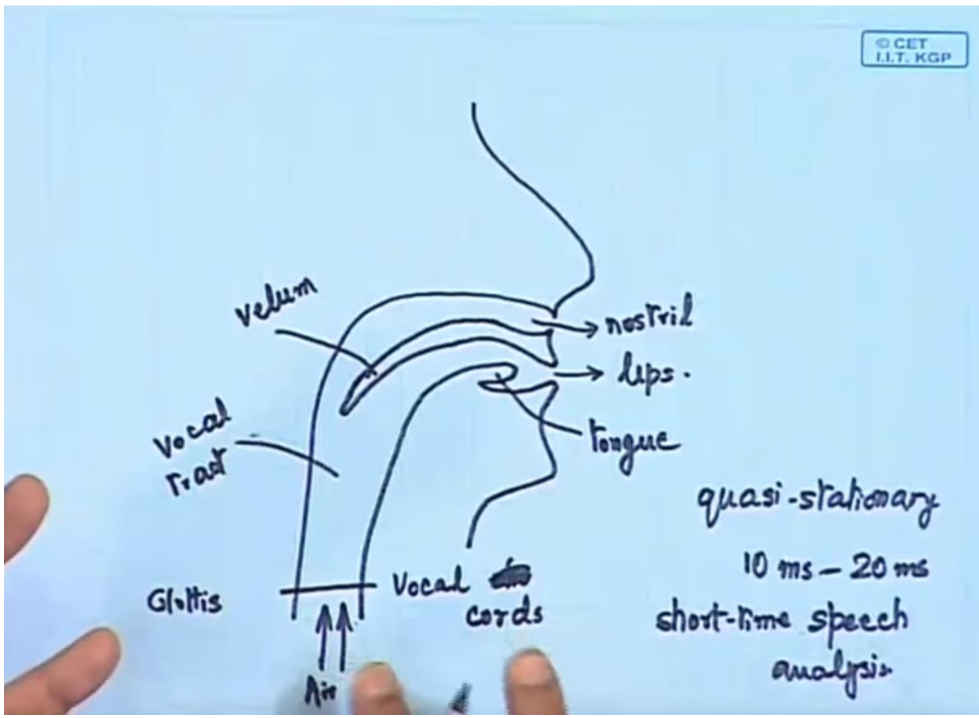
CLOVA

Speech fundamentals

How do we produce speech ?

How do we produce speech?

Speech Production Model



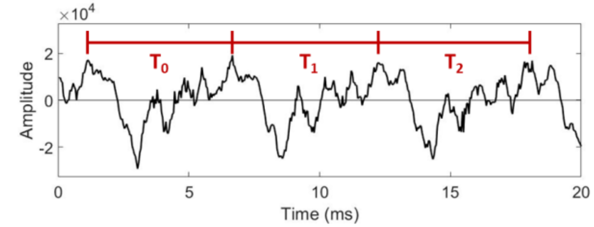
https://www.youtube.com/watch?v=X_JvfZiGEek

Source-filter model

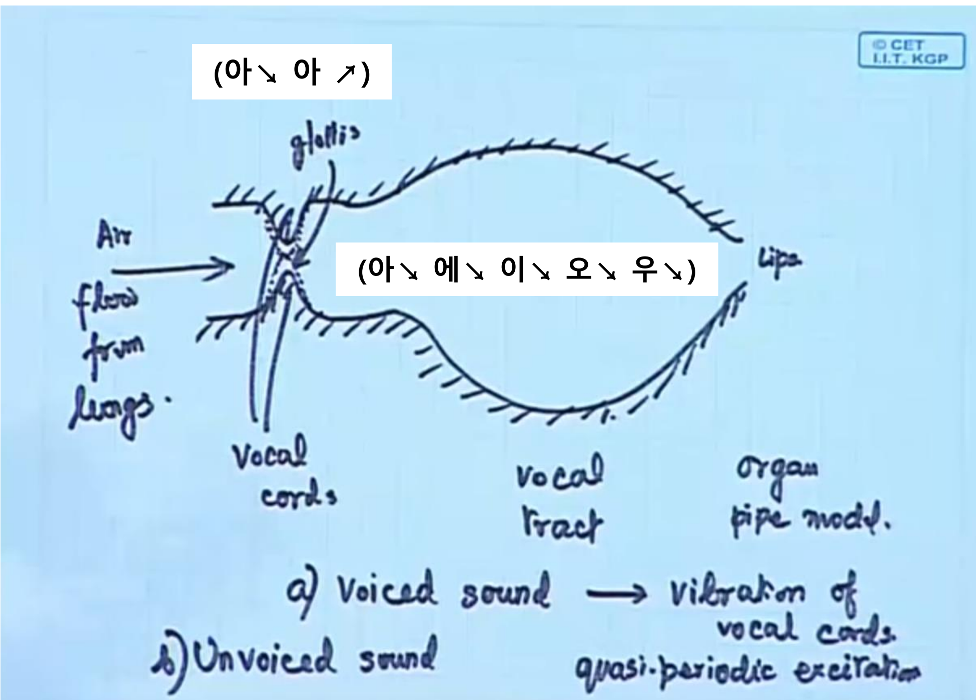
- Lung
  - Power supply
- Glottis ≈ vocal cords ≈ vocal folds
  - Modulator (= source = excitation)
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract (from vocal folds to lips)
  - Filter

Source → **Filter** → **Speech**

How do we produce speech?



Speech Production Model



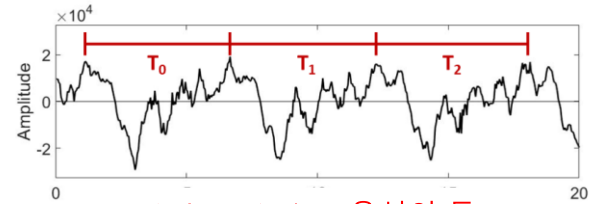
https://www.youtube.com/watch?v=X_JvfZiGEek

Source-filter model

- Lung
  - Power supply
- Glottis ≈ vocal cords ≈ vocal folds
  - Modulator (= source = excitation)
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract (from vocal folds to lips)
  - Filter

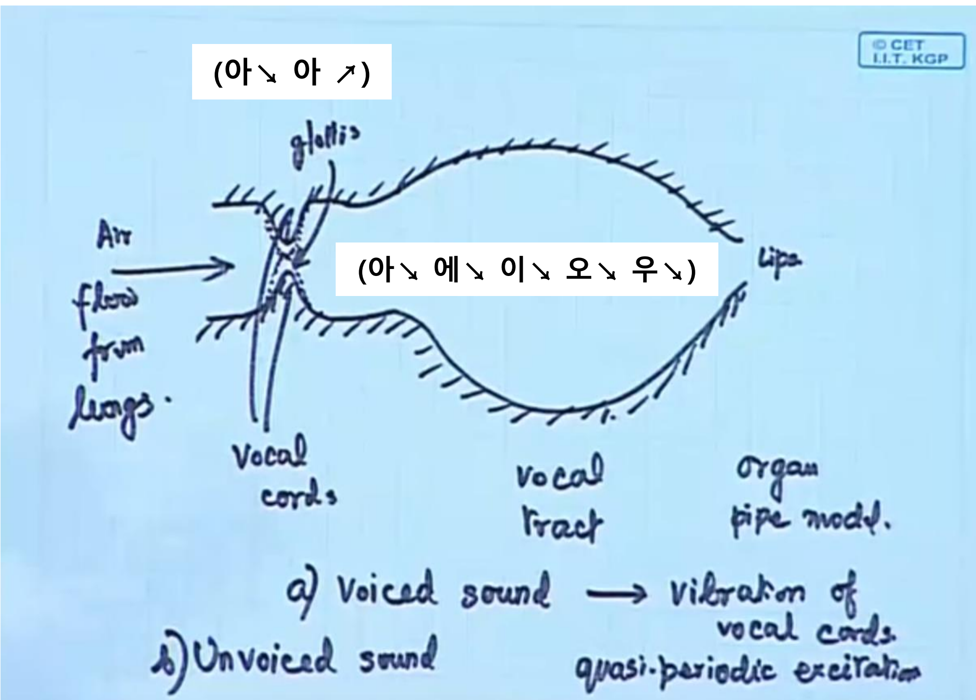
Source → **Filter** → **Speech**

How do we produce speech?



Pitch period → 음성의 톤

Speech Production Model



https://www.youtube.com/watch?v=X_JvfZiGEek

Source-filter model

- Lung
  - Power supply
- Glottis ≈ vocal cords ≈ vocal folds
  - Modulator (= source = excitation)
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract (from vocal folds to lips)
  - Filter

Source → Filter → Speech

How do we produce speech?

Speech Production Model: Linear Prediction

Linear prediction

- Representation of speech
    - Weighted sum. of previous samples.
        - $\hat{s}(n) = \sum_{k=1}^{p} a(k)s(n-k)$
- Prediction error
    - Time-domain
        - $e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a(k)s(n-k)$
    - Minimizing mean square error
        - $\underset{a_k}{\operatorname{argmin}} E\left\{\left\| s(n) - \sum_{k=1}^{p} a(k)s(n-k)\right\|^2\right\}$

Source-filter model

- Lung
    - Power supply
- Glottis ≈ vocal cords ≈ vocal folds
    - Modulator (= source = excitation)
    - Voiced sound : quasi-periodic
    - Unvoiced sound : noisy
- Vocal tract (from vocal folds to lips)
    - Filter

Source → $\boxed{\textbf{Filter}}$ → **Speech**

How do we produce speech?

Speech Production Model: Linear Prediction

Linear prediction

- Representation of speech
  - Weighted sum. of previous samples.
    - $\hat{s}(n) = \sum_{k=1}^{p} a(k)s(n-k)$
- Prediction error
  - Frequency-domain
    - $E(z) = S(z) - \sum_{k=1}^{p} a(k)z^{-k}S(z)$
      $= S(z)\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right)$
    - $S(z) = \frac{E(z)}{1-\sum_{k=1}^{p} a_k z^{-k}} = \frac{E(z)}{A(z)} = E(z)H(z)$
    - $20\log|S(z)| = 20\log|E(z)| + 20\log|H(z)|$

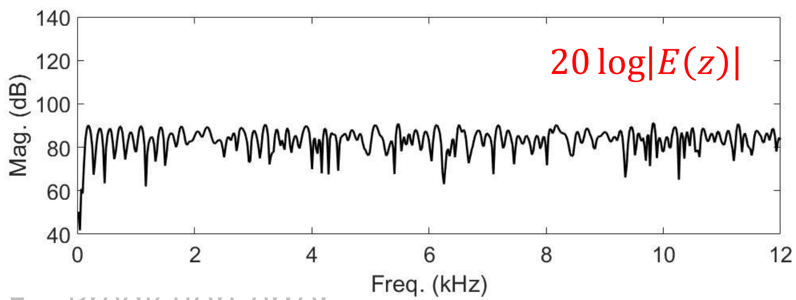
Source-filter model

- Lung
  - Power supply
- Glottis ≈ vocal cords ≈ vocal folds
  - Modulator (= source = excitation)
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract (from vocal folds to lips)
  - Filter

Source → $\boxed{\textbf{Filter}}$ → **Speech**

How do we produce speech?

Speech Production Model: Linear Prediction



- Prediction error
  - Frequency-domain
    - $E(z) = S(z) - \sum_{k=1}^{p} a(k) z^{-k} S(z)$
      $= S(z)\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right)$
    - $S(z) = \frac{E(z)}{1-\sum_{k=1}^{p} a_k z^{-k}} = \frac{E(z)}{A(z)} = E(z)H(z)$
    - $20\log|S(z)| = 20\log|E(z)| + 20\log|H(z)|$

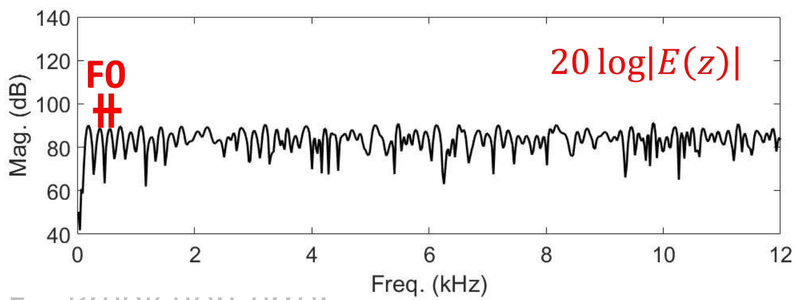
Source-filter model

- Lung
  - Power supply
- Glottis ≈ vocal cords ≈ vocal folds
  - Modulator (= source = excitation)
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract (from vocal folds to lips)
  - Filter

Source → Filter → Speech

How do we produce speech?

Speech Production Model: Linear Prediction



- Prediction error
  - Frequency-domain
    - $E(z) = S(z) - \sum_{k=1}^{p} a(k) z^{-k} S(z)$

      $= S(z)\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right)$

    - $S(z) = \frac{E(z)}{1-\sum_{k=1}^{p} a_k z^{-k}} = \frac{E(z)}{A(z)} = E(z)H(z)$

    - $20\log|S(z)| = 20\log|E(z)| + 20\log|H(z)|$

Source-filter model

- Lung
  - Power supply
- Glottis ≈ vocal cords ≈ vocal folds
  - Modulator (= source = excitation)
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract (from vocal folds to lips)
  - Filter

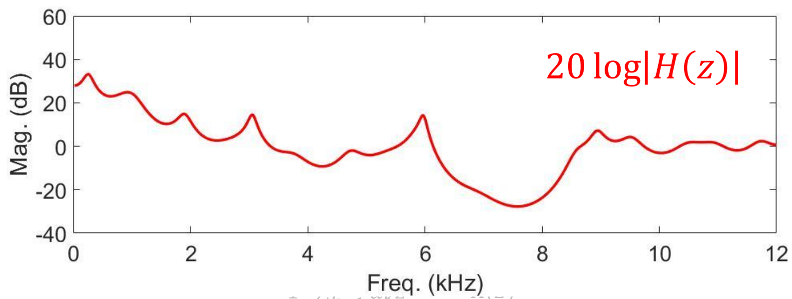
Source → **Filter** → **Speech**

How do we produce speech?

Speech Production Model: Linear Prediction

Linear prediction

- Representation of speech
  - Weighted sum. of previous samples.
    - $\hat{s}(n) = \sum_{k=1}^{p} a(k)s(n-k)$
- Prediction error



- $20\log|S(z)| = 20\log|E(z)| + 20\log|H(z)|$

Source-filter model

- Lung
  - Power supply
- Glottis ≈ vocal cords ≈ vocal folds
  - Modulator (= source = excitation)
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract (from vocal folds to lips)
  - Filter

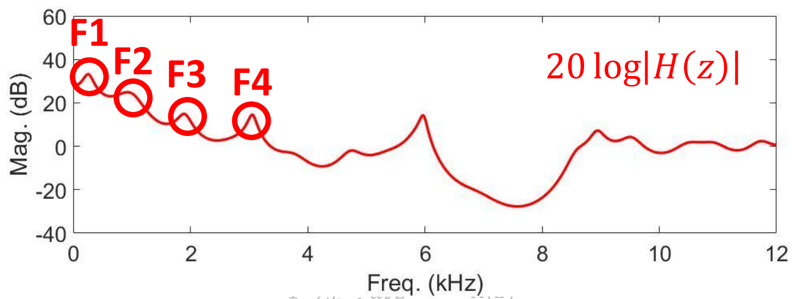
Source → $\boxed{\textbf{Filter}}$ → Speech

How do we produce speech?

Speech Production Model: Linear Prediction

Linear prediction

- Representation of speech
  - Weighted sum. of previous samples.
    - $\hat{s}(n) = \sum_{k=1}^{p} a(k)s(n-k)$
- Prediction error



- $20\log|S(z)| = 20\log|E(z)| + 20\log|H(z)|$

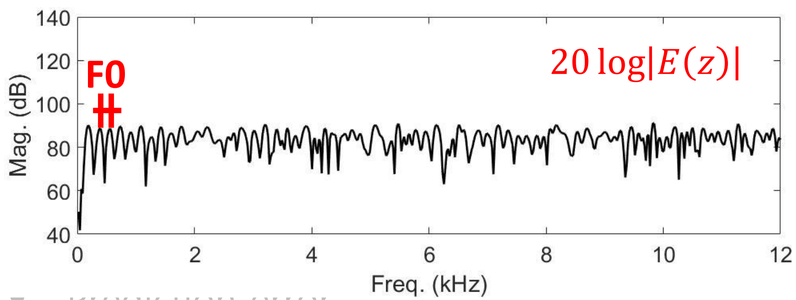
Source-filter model

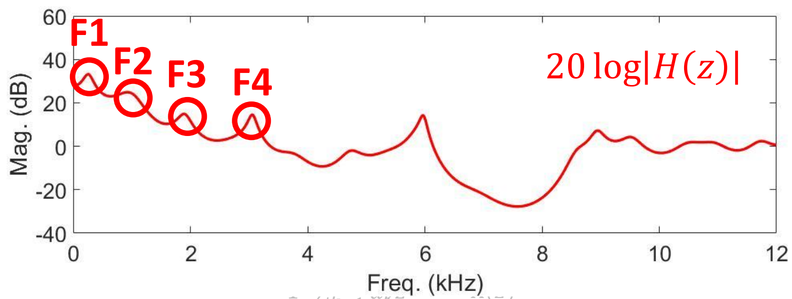
- Lung
  - Power supply
- Glottis ≈ vocal cords ≈ vocal folds
  - Modulator (= source = excitation)
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract (from vocal folds to lips)
  - Filter

Source → **Filter** → Speech

How do we produce speech?

Speech Production Model: Linear Prediction





- $20\log|S(z)| = 20\log|E(z)| + 20\log|H(z)|$

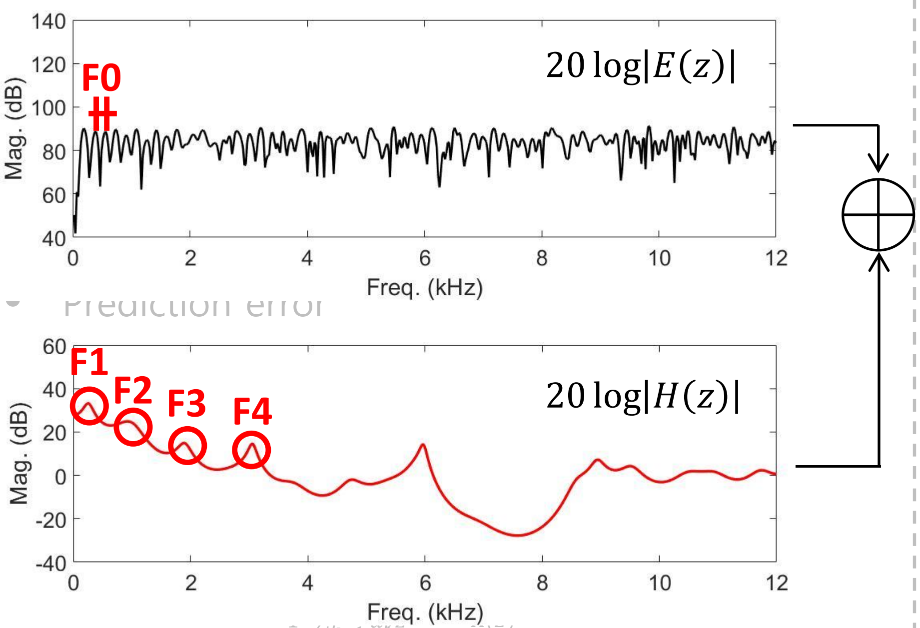
Source-filter model

- Lung
  - Power supply
- Glottis ≈ vocal cords ≈ vocal folds
  - Modulator (= source = excitation)
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract (from vocal folds to lips)
  - Filter

Source → Filter → Speech

How do we produce speech?

Speech Production Model: Linear Prediction



- $\color{red}{20\log|S(z)|} = 20\log|E(z)| + 20\log|H(z)|$

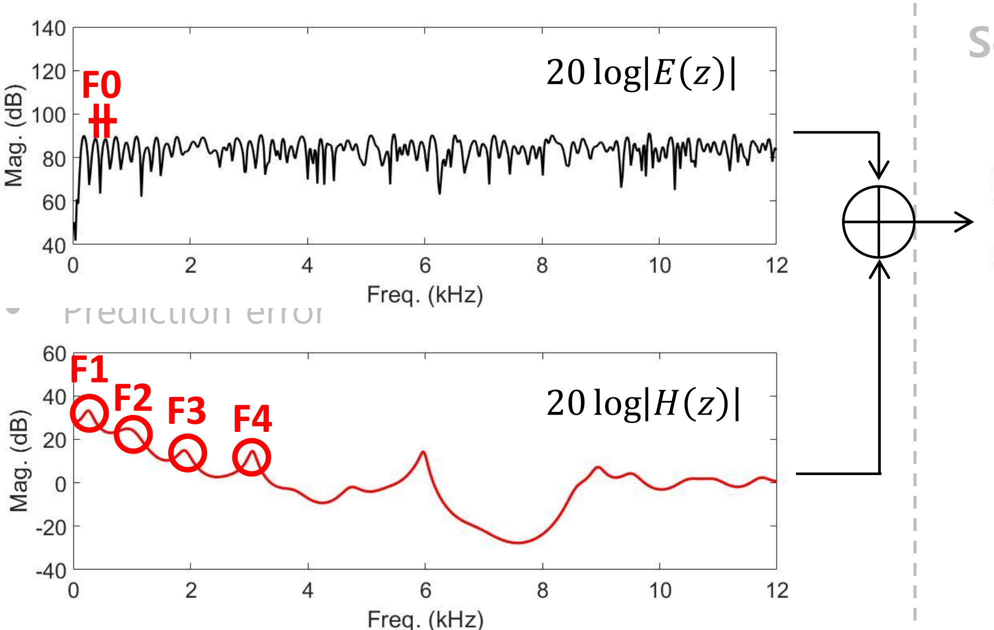
Source-filter model

- Lung
  - Power supply
- Glottis ≈ vocal cords ≈ vocal folds
  - Modulator (= source = excitation)
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy
- Vocal tract (from vocal folds to lips)
  - Filter

Source → **Filter** → Speech

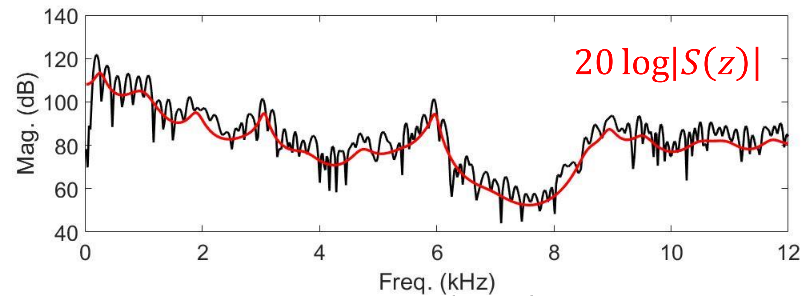
How do we produce speech?

Speech Production Model: Linear Prediction



- Prediction error

Source-filter model



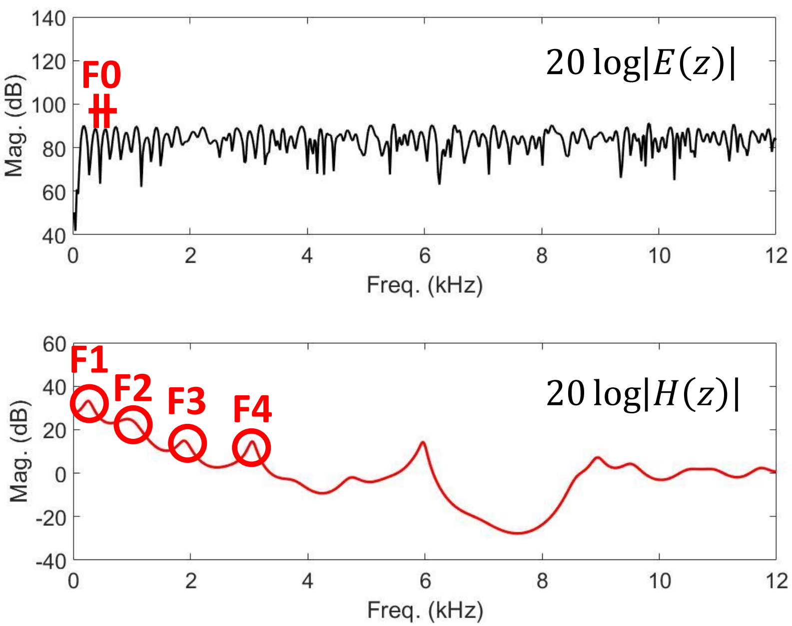
- Unvoiced sound : noisy
- Vocal tract (from vocal folds to lips)
  - Filter

- $20\log|S(z)| = 20\log|E(z)| + 20\log|H(z)|$

Source → **Filter** → **Speech**

Summary

Pitch Period (or F0) 와 Linear Prediction 을 꼭 기억해 주세요!



Pitch period

- Long-term period of speech (time-domain)

Fundamental frequency (F0)

- 1 / PP (frequency-domain)

Harmonic spectrum

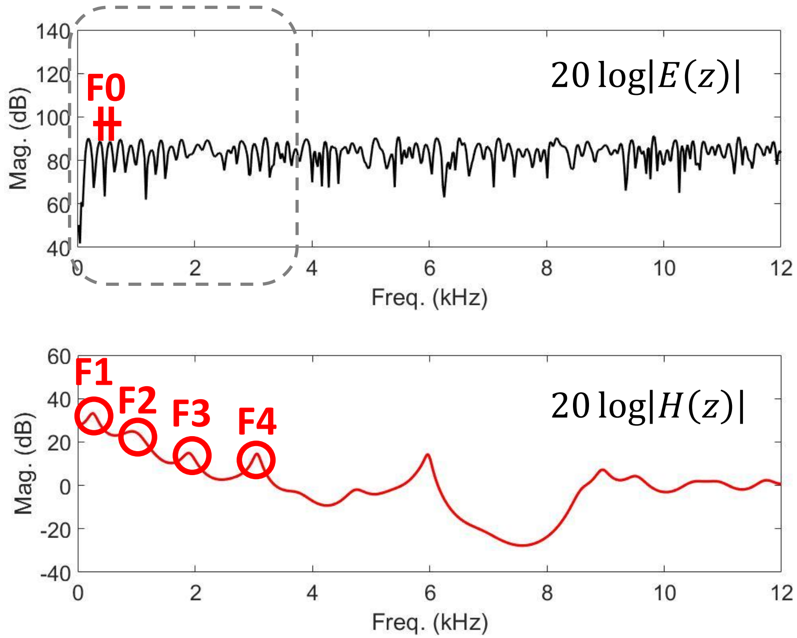
- Multiple peaks of speech spectrum (interval=F0)

Formant frequency (F1, F2, ...)

- Vocal tract resonance

Summary

Pitch Period (or F0) 와 Linear Prediction 을 꼭 기억해 주세요!



Source-filter model

- Glottis ≈ vocal cords ≈ vocal folds
  - Excitation = linear prediction residual

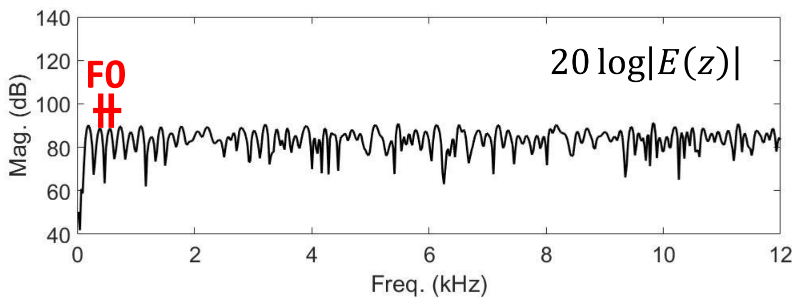
→ Vocal cords movement determines F0
(아↘ 아 ↗)

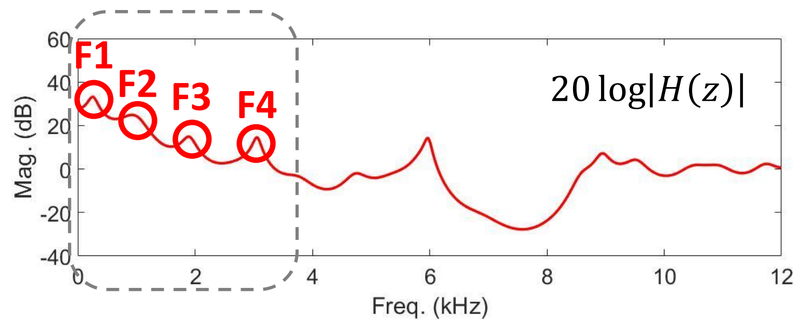
- Vocal tract (from vocal folds to lips)
  - Linear prediction filter

→ LP spectrum determines fomant structure
(아↘ 에↘ 이↘ 오↘ 우↘)

Summary

Pitch Period (or F0) 와 Linear Prediction 을 꼭 기억해 주세요!





Source-filter model

- Glottis ≈ vocal cords ≈ vocal folds
  - Excitation = linear prediction residual

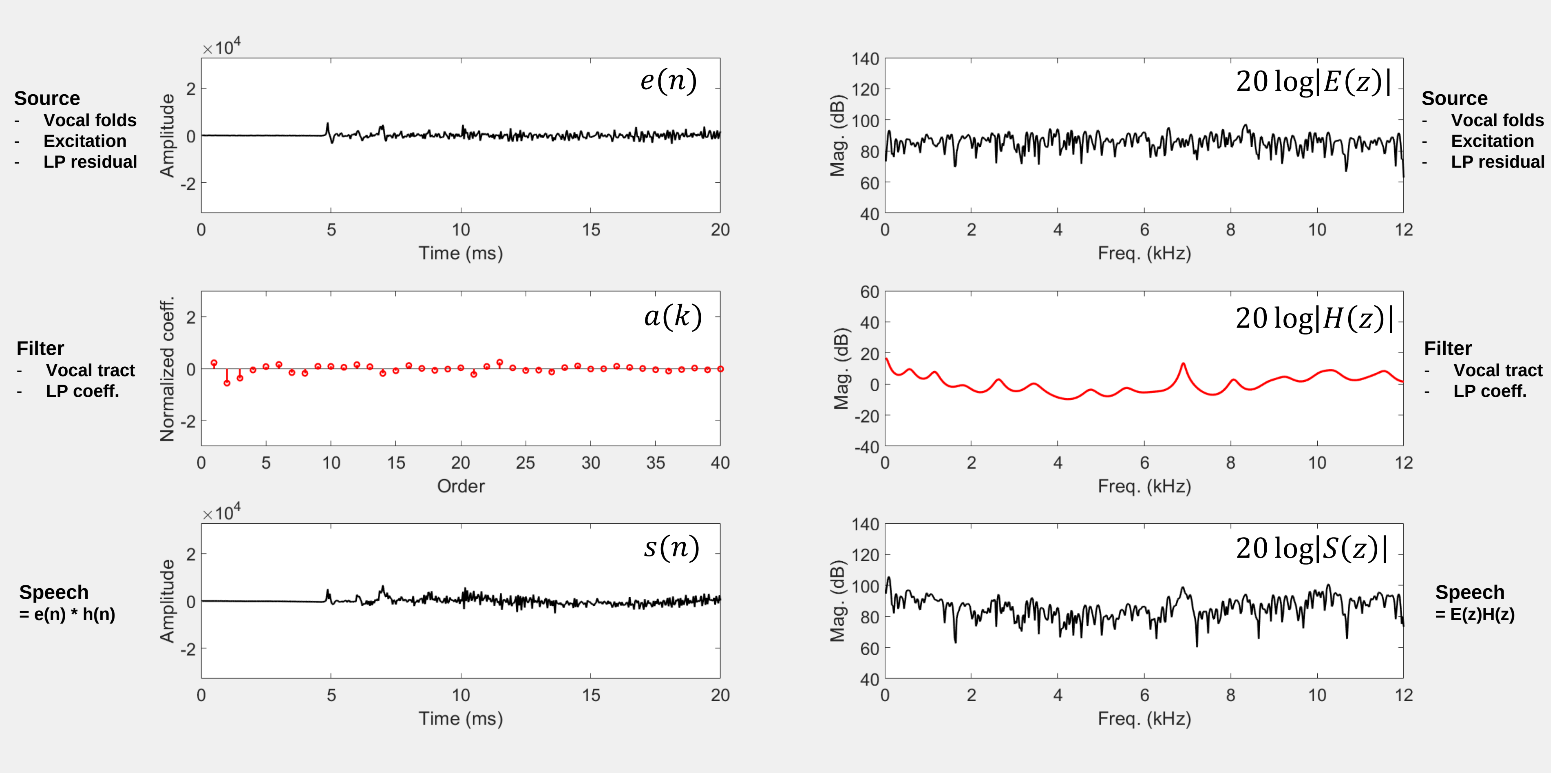
  → Vocal cords movement determines F0
  (아↘ 아 ↗)

- Vocal tract (from vocal folds to lips)
  - Linear prediction filter

  → LP spectrum determines fomant structure
  (아↘ 에↘ 이↘ 오↘ 우↘)

Summary

Time-frequency analysis of speech production model



CLOVA

Vocoding model

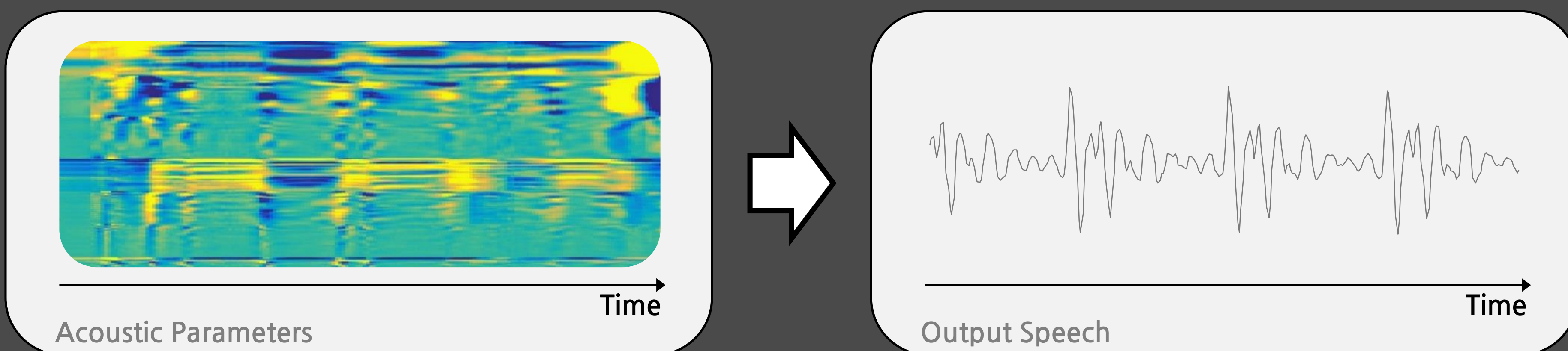
Parametric LPC vocoder

Recall

Text-to-speech (TTS) 란 기계가 사람처럼 텍스트를 읽어주는 기술입니다.

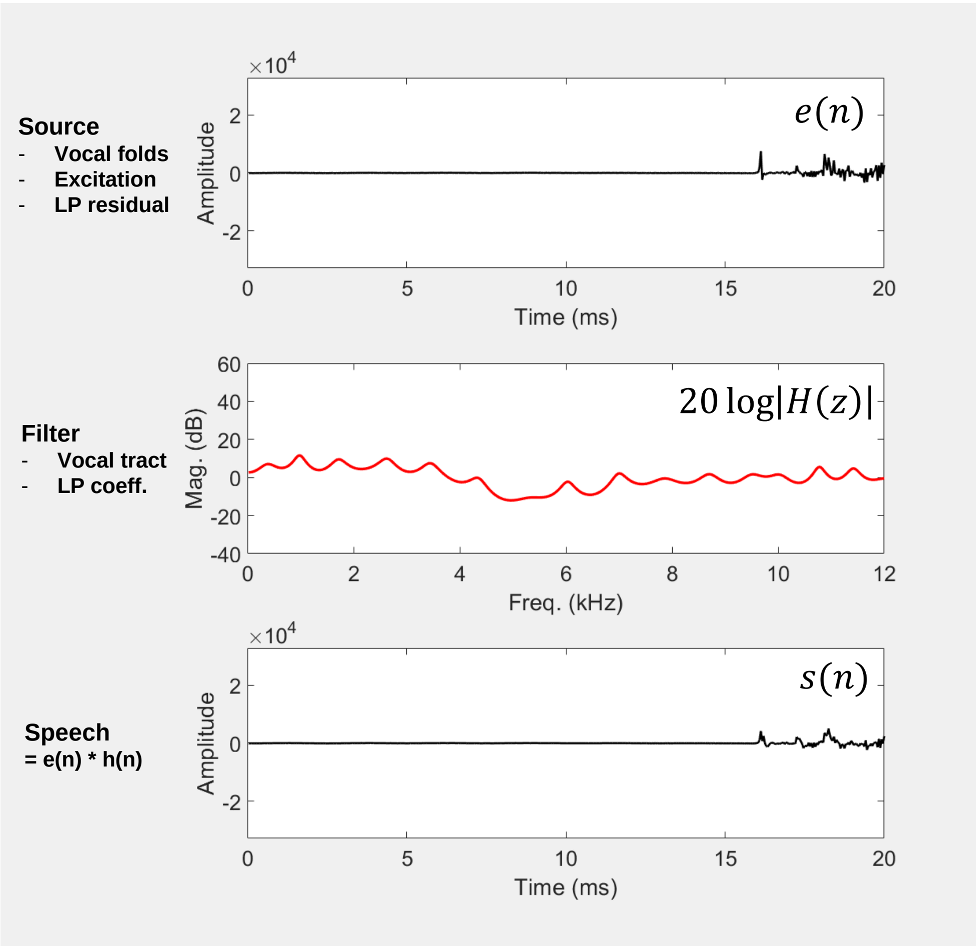
Acoustic Parameter 에서 음성 신호를 생성

Input Text → Acoustic Model → Acoustic Parameters → Vocoder → Output Speech



Recall

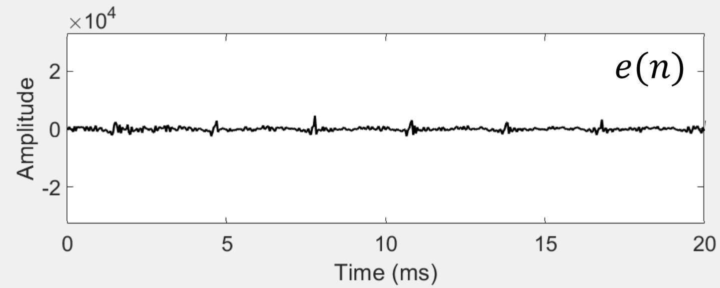
20 ms 음성 신호를 어떻게 만들 수 있을까요?



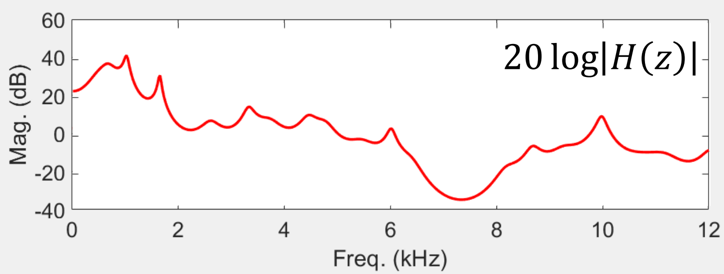
Recall

LP coefficients 40 개

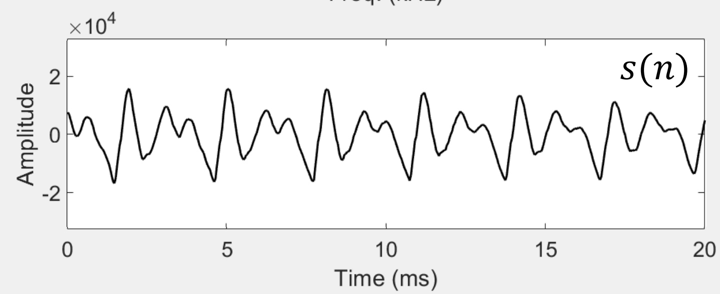
Source
- **Vocal folds**
- **Excitation**
- **LP residual**

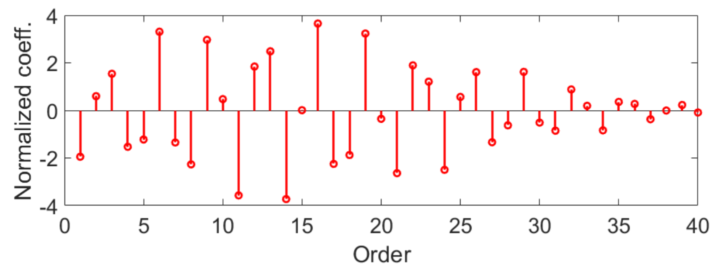


Filter
- **Vocal tract**
- **LP coeff.**



Speech
= e(n) * h(n)

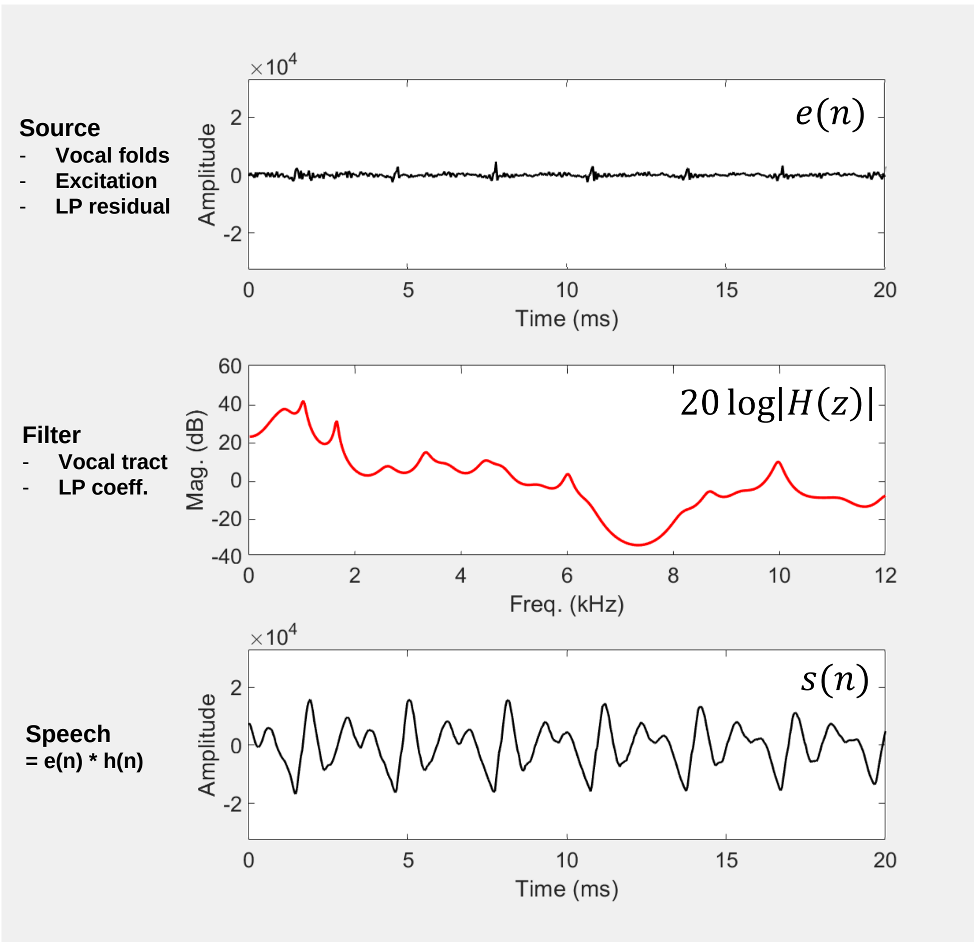


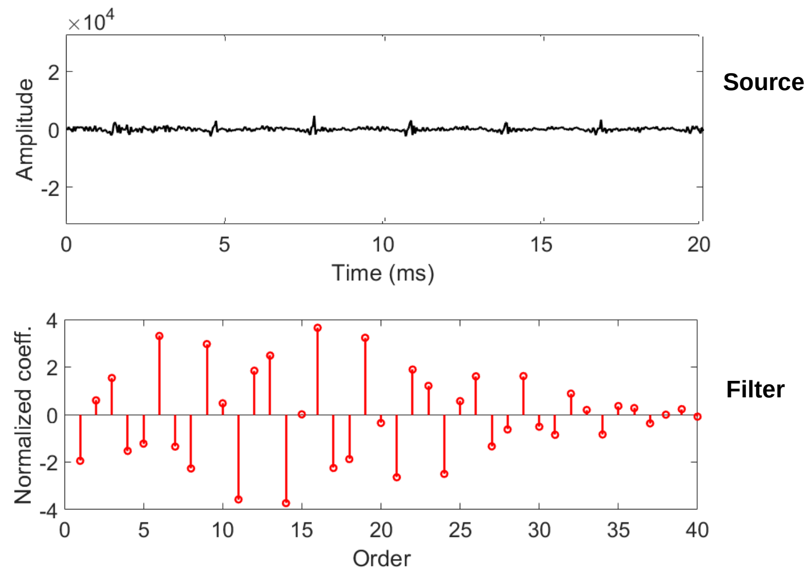


Filter

Recall

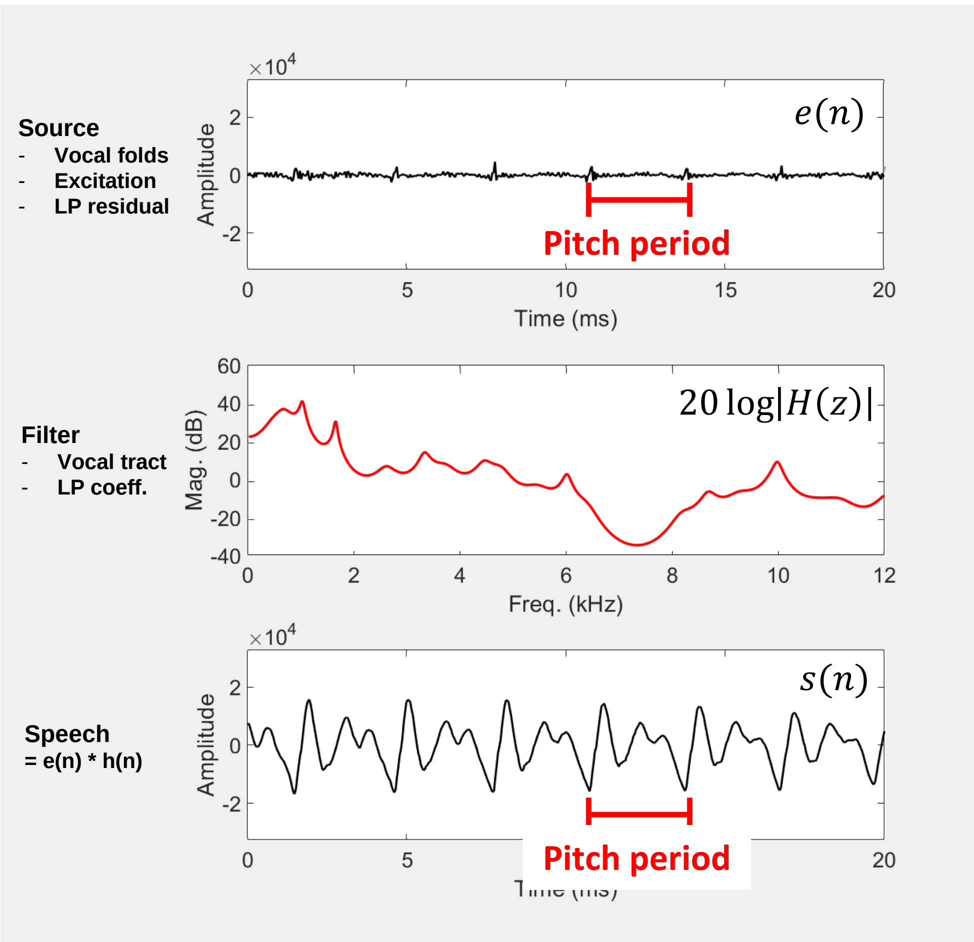
LP coefficients 40 개 + Excitation 20 ms

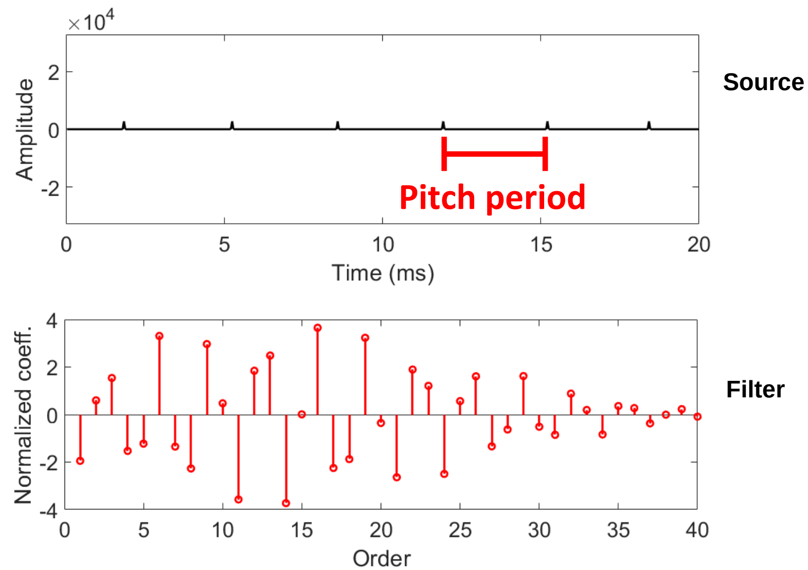




Recall

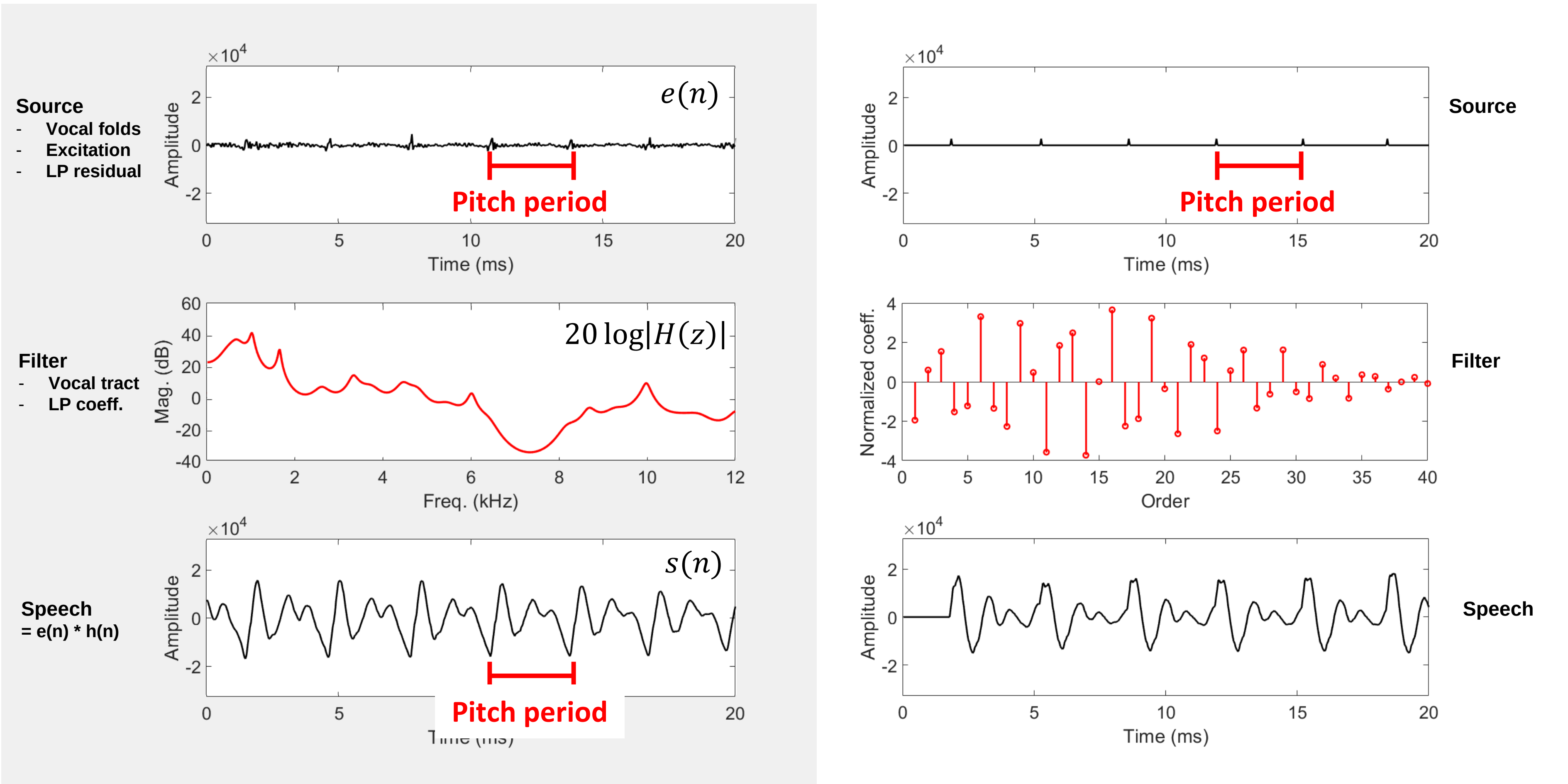
LP coefficients 40 개 + Excitation 20 ms (approximation using pitch period)





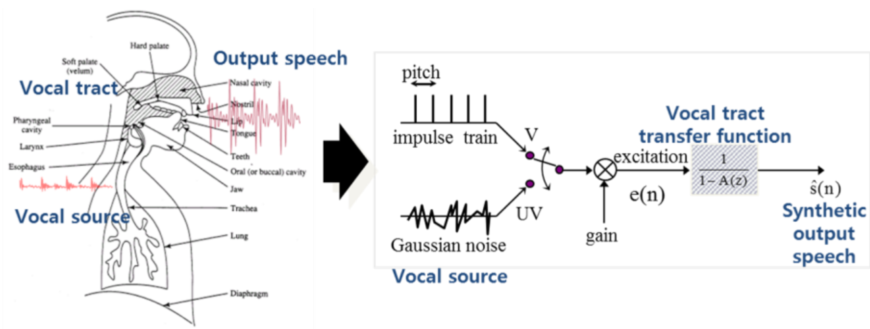
Recall

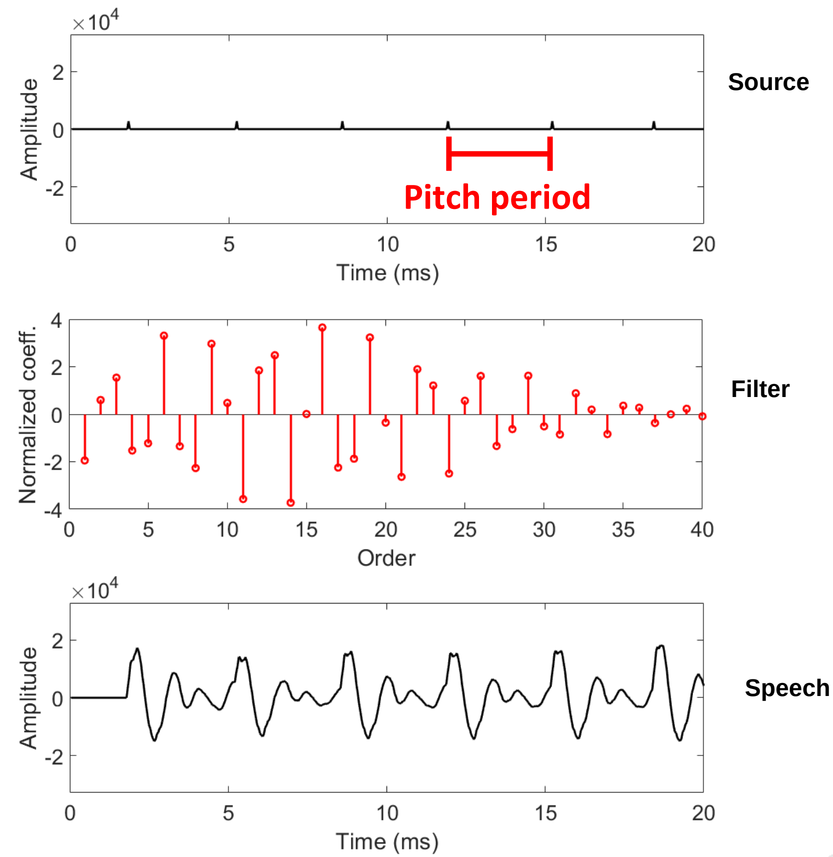
LP coefficients 40 개 + Excitation 20 ms (approximation using pitch period)



Parametric LPC synthesis

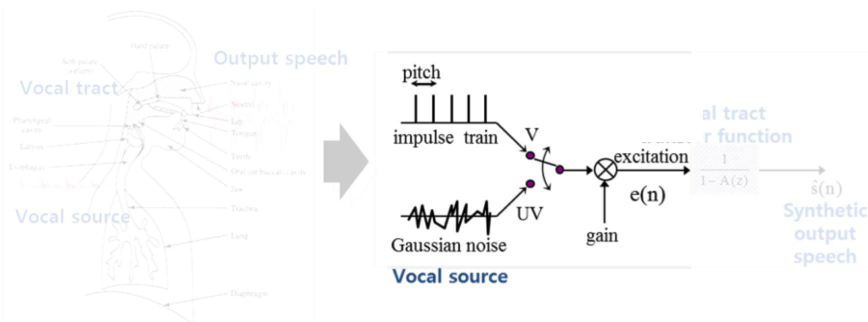
LP coefficient 와 approximated excitation 을 이용해서 **음성**을 만들 수 있습니다.





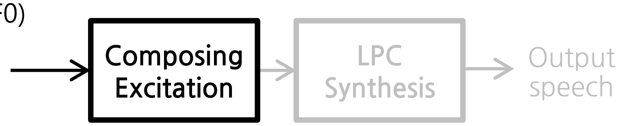
Parametric LPC synthesis

LP coefficient 와 **approximated excitation** 을 이용해서 **음성**을 만들 수 있습니다.



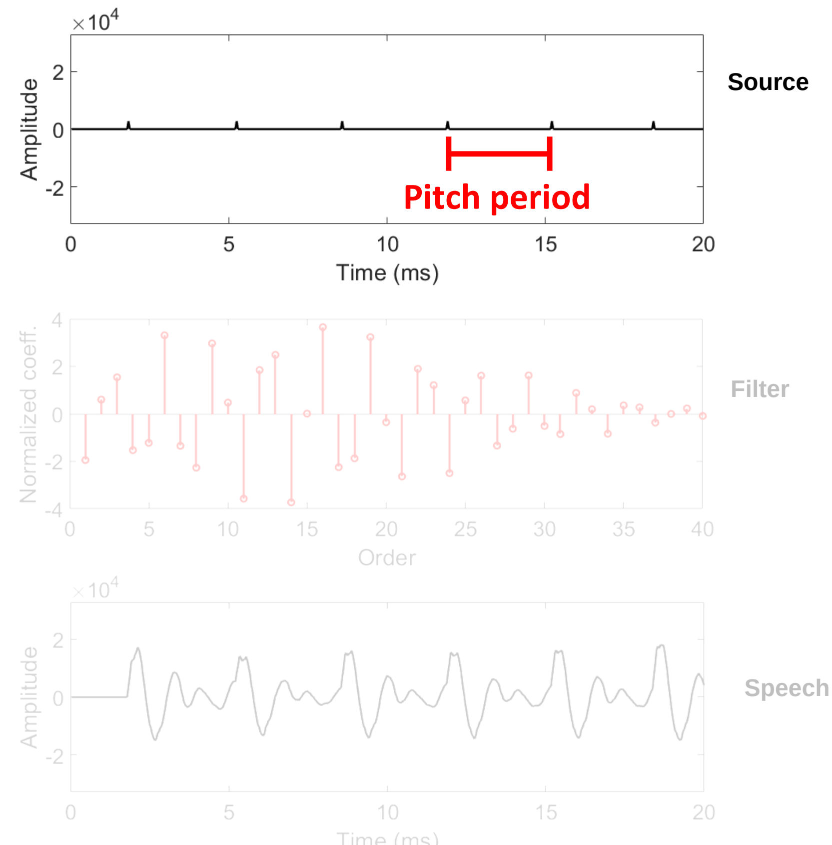
Excitation parameters

- Pitch period (or F0)
- Voicing flag
- Gain



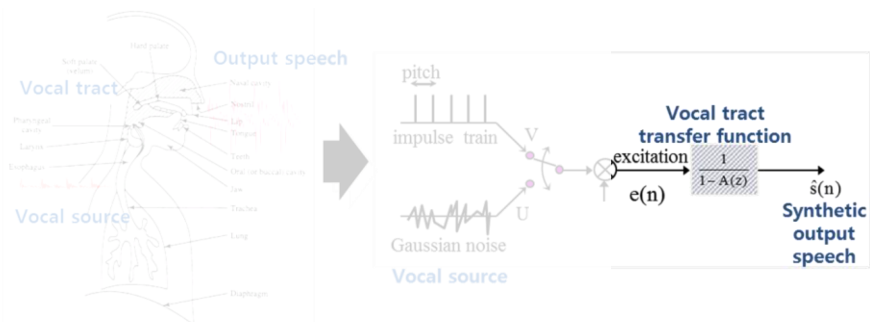
Spectral parameters

- LP coefficients



Parametric LPC synthesis

LP coefficient 와 approximated excitation 을 이용해서 **음성**을 만들 수 있습니다.



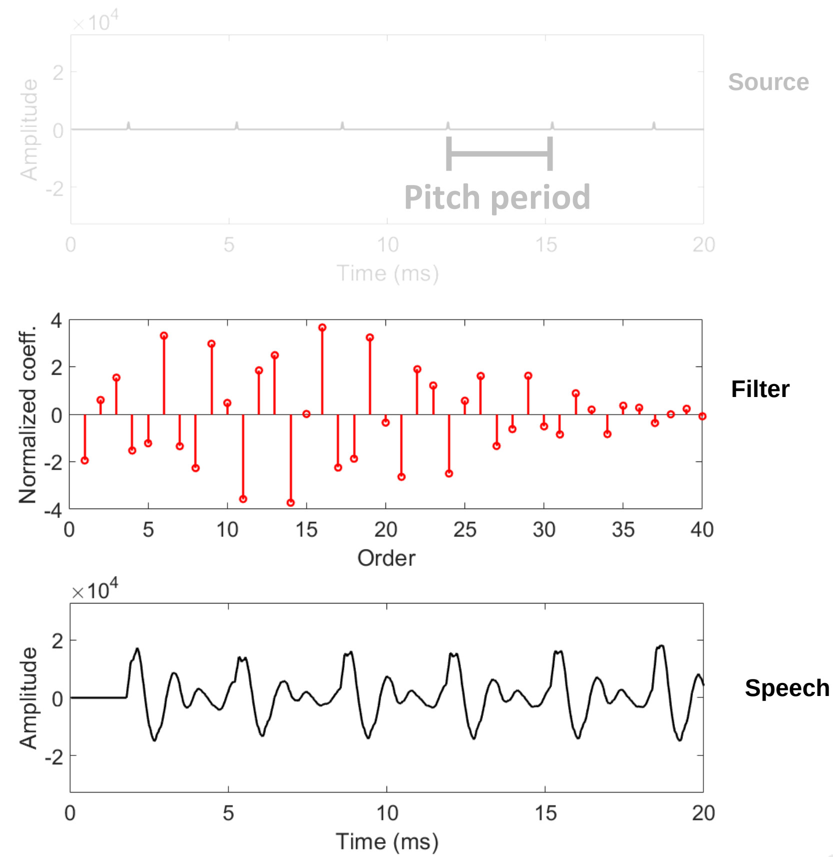
Excitation parameters

- Pitch period (or F0)
- Voicing flag
- Gain

Spectral parameters

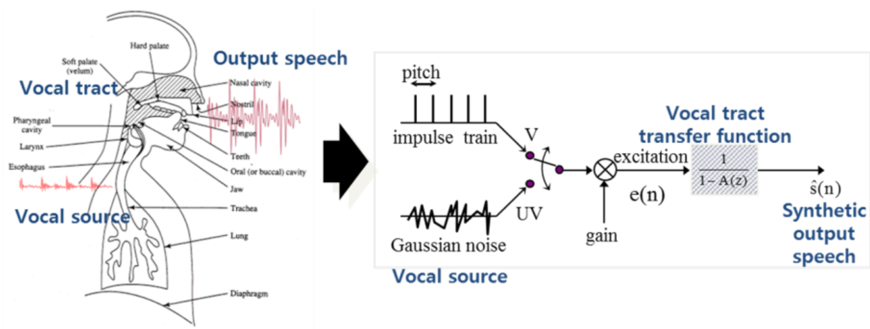
- LP coefficients

Composing Excitation → **LPC Synthesis** → Output speech



Parametric LPC synthesis

LP coefficient 와 approximated excitation 을 이용해서 **음성**을 만들 수 있습니다.



Excitation parameters

- Pitch period (or F0)
- Voicing flag
- Gain

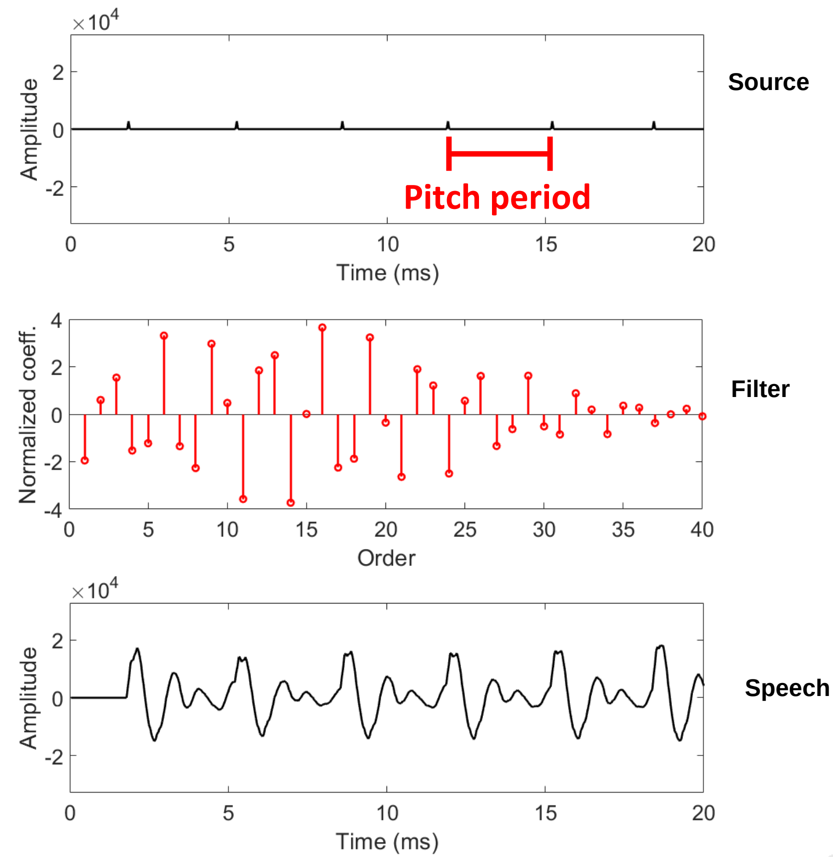
Spectral parameters

- LP coefficients

→ Composing Excitation → LPC Synthesis → Output speech

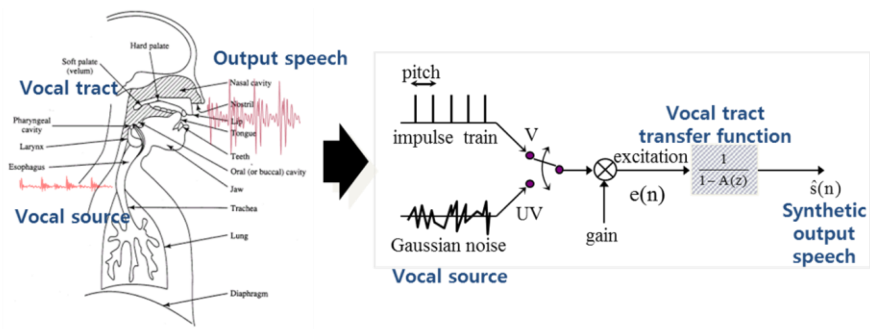
Acoustic parameters

Vocoder synthesis



Parametric LPC synthesis

LP coefficient 와 approximated excitation 을 이용해서 **음성**을 만들 수 있습니다.



Excitation parameters

- Pitch period (or F0)
- Voicing flag
- Gain

Spectral parameters

- LP coefficients

→ Composing Excitation → LPC Synthesis → Output speech

Acoustic parameters

Vocoder synthesis

Recorded speech

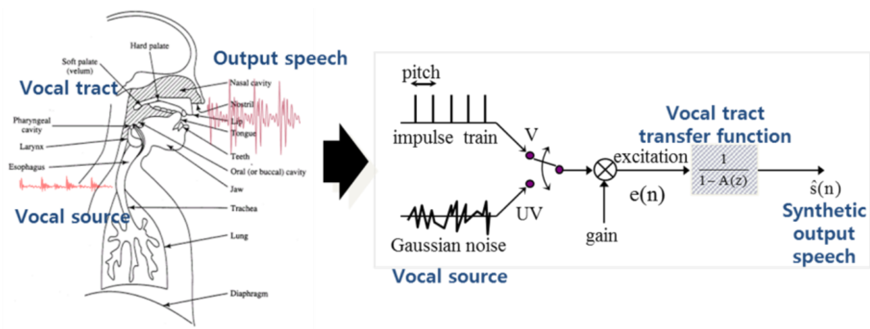
Generated speech

Low-tone

High-tone

Parametric LPC synthesis

LP coefficient 와 approximated excitation 을 이용해서 **음성**을 만들 수 있습니다.



Excitation parameters

- Pitch period (or F0)
- Voicing flag
- Gain

Spectral parameters

- LP coefficients

→ Composing Excitation → LPC Synthesis → Output speech

Acoustic parameters

Vocoder synthesis

Recorded speech

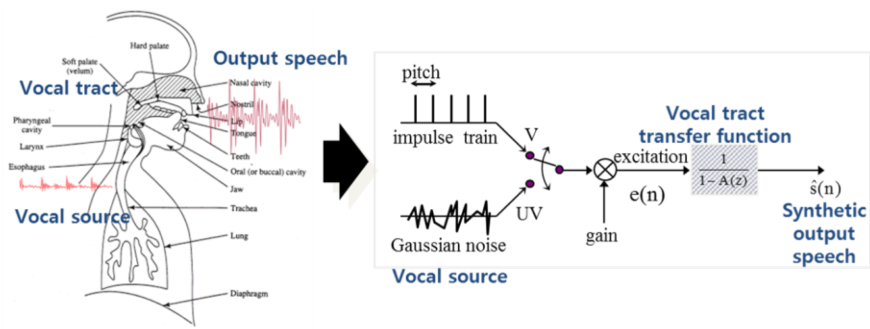
Generated speech

Low-tone

High-tone

Parametric LPC synthesis

LP coefficient 와 approximated excitation 을 이용해서 **음성**을 만들 수 있습니다.



Excitation parameters

- Pitch period (or F0)
- Voicing flag
- Gain

Spectral parameters

- LP coefficients

→ Composing Excitation → LPC Synthesis → Output speech

Acoustic parameters

Vocoder synthesis

Recorded speech

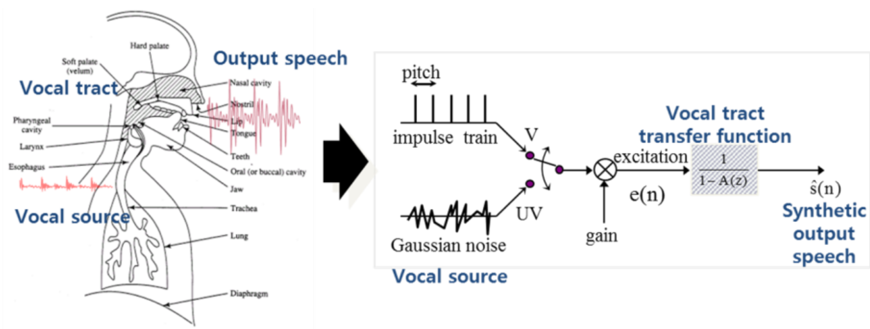
Generated speech

Low-tone

High-tone

Parametric LPC synthesis

LP coefficient 와 approximated excitation 을 이용해서 **음성**을 만들 수 있습니다.



Excitation parameters

- Pitch period (or F0)
- Voicing flag
- Gain

Spectral parameters

- LP coefficients

Composing Excitation → LPC Synthesis → Output speech

Acoustic parameters

Vocoder synthesis

Recorded speech

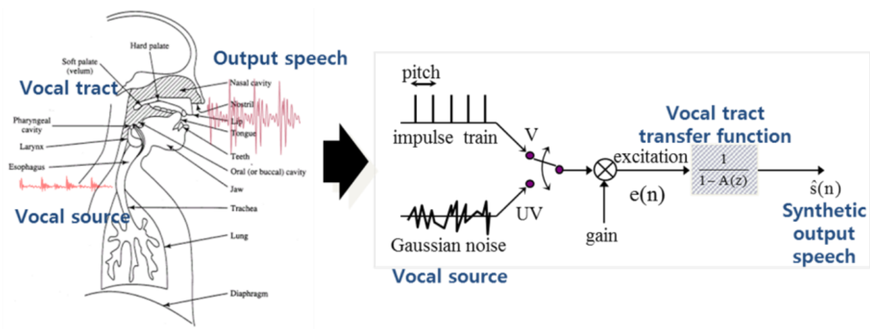
Generated speech

Low-tone

High-tone

Parametric LPC synthesis

LP coefficient 와 approximated excitation 을 이용해서 **음성**을 만들 수 있습니다.



Excitation parameters

- Pitch period (or F0)
- Voicing flag
- Gain

Spectral parameters

- LP coefficients

→ Composing Excitation → LPC Synthesis → Output speech

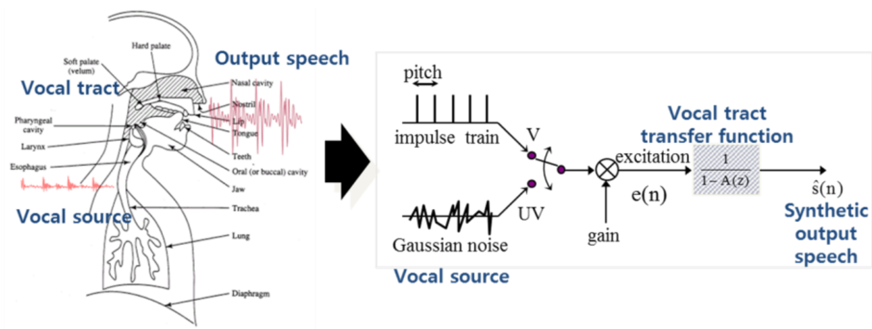
Acoustic parameters | **Vocoder synthesis**

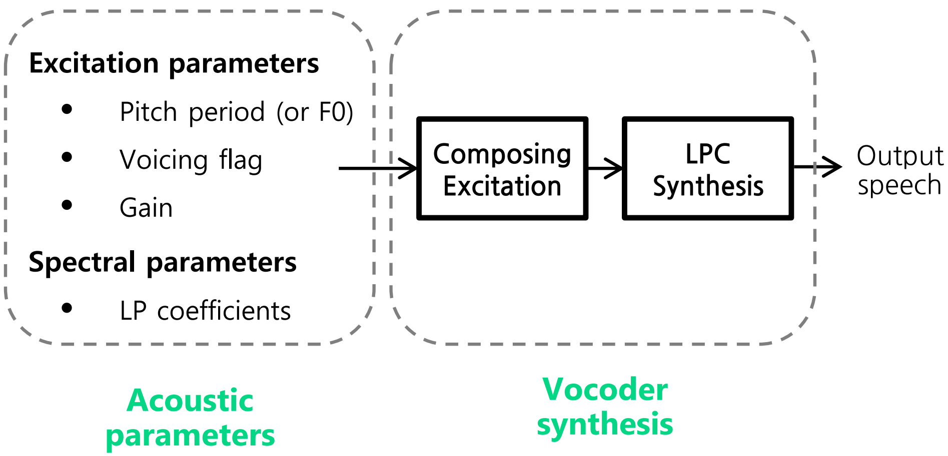
Spectral parameters

- How to extract LP coefficients ?
  - $\hat{s}(n) = \sum_{k=1}^{p} a(k)s(n-k)$
  - $e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a(k)s(n-k)$
- Minimizing mean square error
  - $\underset{a_k}{\operatorname{argmin}}\, E\left\{\left\| s(n) - \sum_{k=1}^{p} a(k)s(n-k) \right\|^2\right\}$
  - Levinson-Durbin recursion
- Parameterization
  - Line spectral frequency (LSF)
  - Mel-generalized cepstrum (MGC)
  - Mel-spectrum

Parametric LPC synthesis

LP coefficient 와 approximated excitation 을 이용해서 **음성**을 만들 수 있습니다.





Excitation parameters

- Approximation methods
  - Pulse or noise (PoN)
    - Pitch period, voicing flag, gain
  - Mixed excitation (STRAIGHT, WORLD)
    - Pitch period, voicing flag, gain
    - Band aperiodicity

Summary

음성 개념 1: Pitch period (or F0), formant

음성 개념 2: Speech production model, linear prediction

음성 개념 3: Parametric LPC vocoder

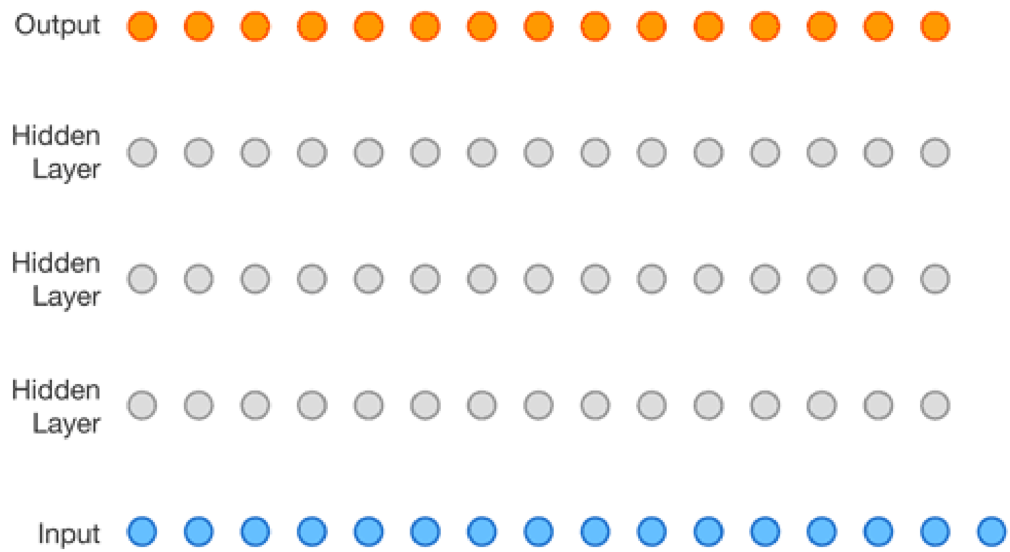
CLOVA

Vocoding model

Autoregressive WaveNet vocoder

WaveNet synthesis

Neural network 로 **sample** 단위의 **음성** 신호를 추정할 수 있습니다.



현재 음성 신호를 예측할 때 과거 음성 신호를 함께 사용합니다.
이러한 방법을 **Autoregressive Model** 라고 정의합니다.

출처: https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

WaveNet synthesis

중요하니... 이론을 좀 ..

WaveNet

- A. Van den Oord, et. al., “WaveNet; a generative model for raw audio,” CoRR abs/1609.03499, 2016.
- The first TTS algorithm that generates signal with a sample-by-sample manner

Properties

- Turn regression task into classification task (Speech is quantized to 8 bits (256 classes))
- Directly predicts the distribution of next sample, given condition and previous samples
- Maximize likelihood
  - $p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t | x_1, \cdots, x_{t-1})$

Key features

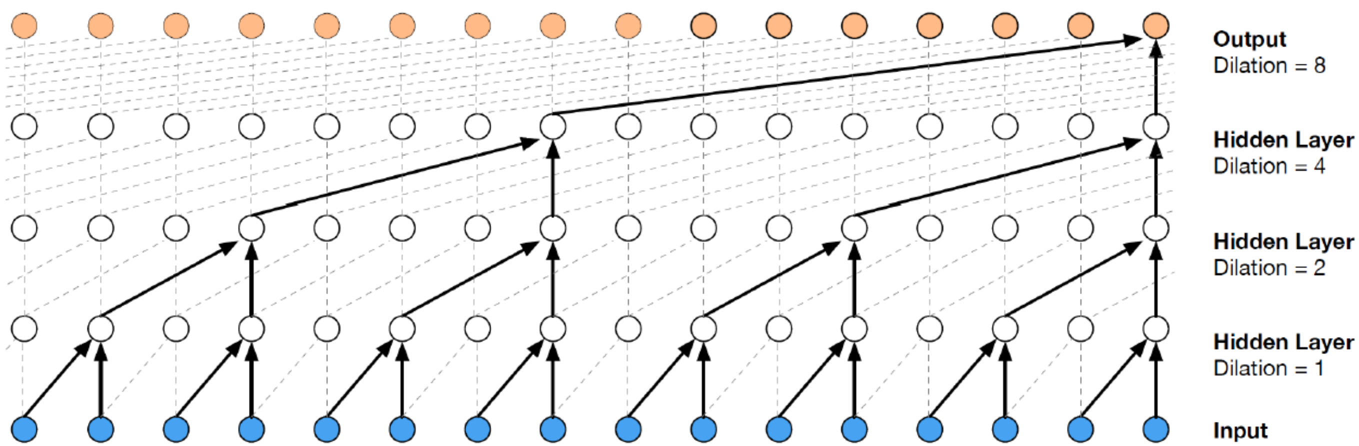
- Dilated causal convolutions
- Softmax distribution
- Gated activation units
- Residual and skip connections
- Conditional WaveNets

WaveNet synthesis

중요하니... 이론을 좀 ..

Dilated causal convolution

- Stacked dilated convolution: 1, 2, 4, 8, 16, ...



Softmax distributions

- 8 bit (256 level) mu-law companding transformation
  - $f(x_t) = sign(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}$

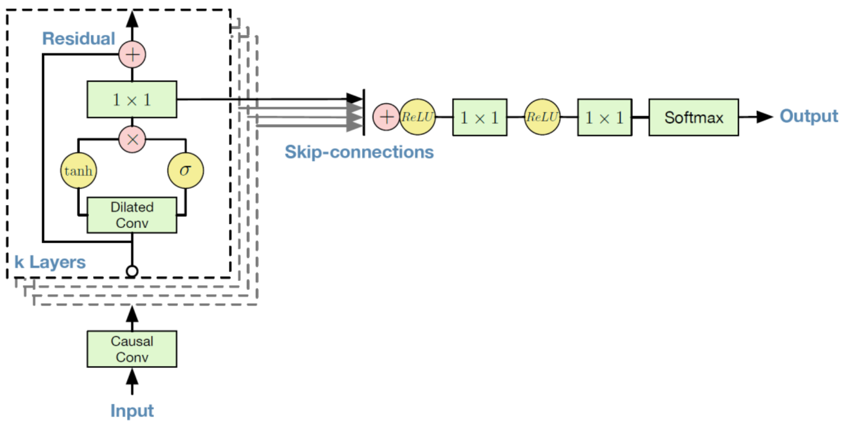
WaveNet synthesis

중요하니... 이론을 좀 ..

Gated activation units

- $\mathbf{z} = \tanh(W_{f,k} * \mathbf{x}) \odot \delta(W_{g,k} * \mathbf{x})$

Residual and skip connections



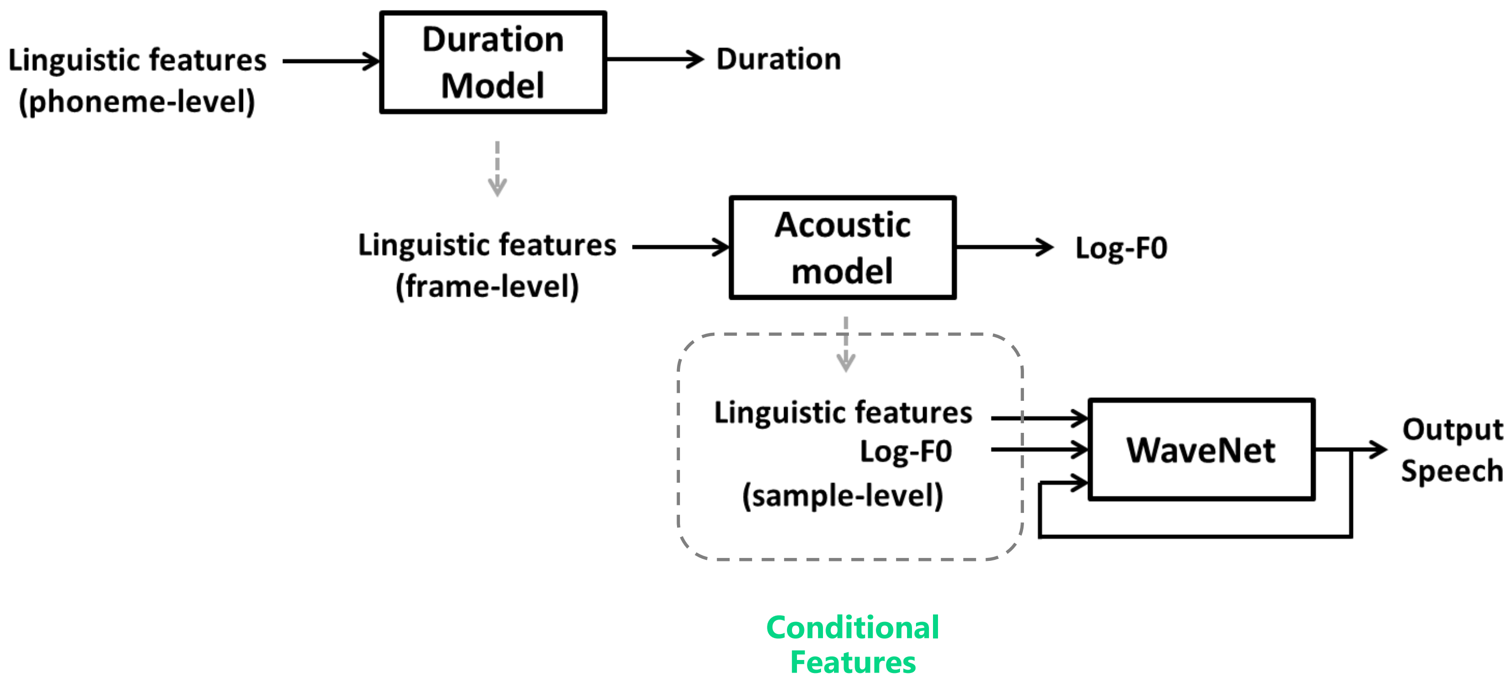
Conditional WaveNets

- $p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^{T} p(x_t|x_1, \cdots, x_{t-1}, \mathbf{h})$
- $\mathbf{z} = \tanh(W_{f,k} * \mathbf{x} + \boldsymbol{V}_{f,k}^{T}\mathbf{h}) \odot \delta(W_{g,k} * \mathbf{x} + \boldsymbol{V}_{g,k}^{T}\mathbf{h})$

WaveNet synthesis

End-to-end 는 아닙니다만 ..

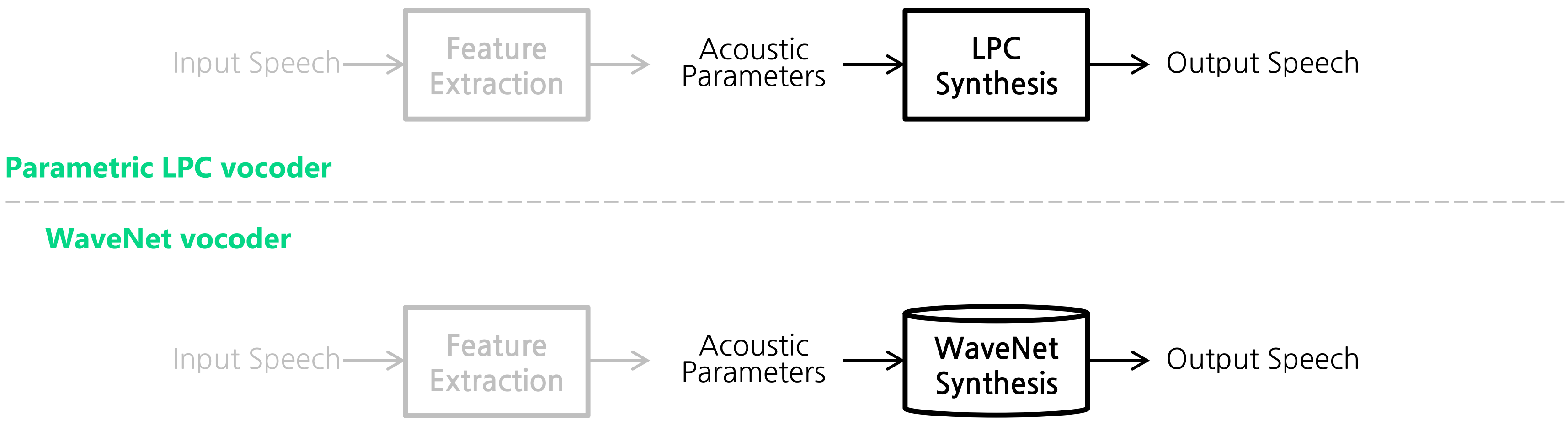
처음에는 Vocoder 모델이 아니라 **End-to-end TTS 모델**로 사용되었습니다.



출처: A. Van Den Oord, et al., “WaveNet: A generative model for raw audio,” *CoRR abs/1609.03499*, 2016.

WaveNet synthesis

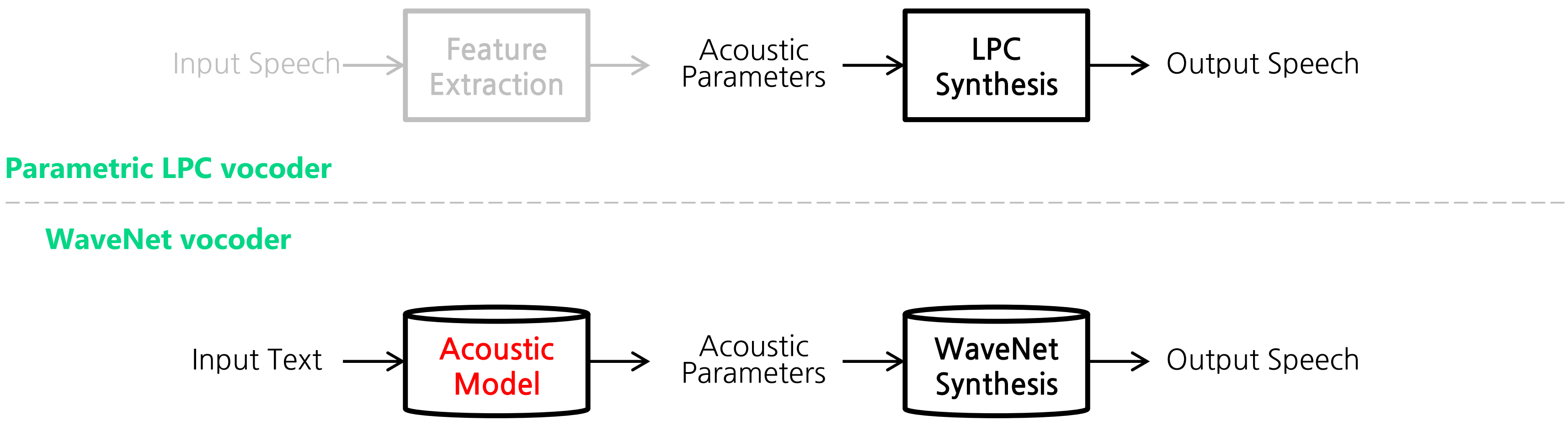
Input Condition 으로 **Acoustic Parameter** 를 넣어줘야 비로소 **Vocoder** 가 됩니다.



출처: A. Tamamory, et. al., "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, 2017.

WaveNet synthesis

Input Condition 으로 **Acoustic Parameter** 를 넣어줘야 비로소 **Vocoder** 가 됩니다.



Tacotron 2

출처: J. Shen, et. al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, 2018.

WaveNet synthesis

Parametric LPC Vocoder 보다 월등히 좋은 성능을 보여줍니다.

Table 1: *Comparative methods of waveform synthesis; spectrum envelop was extracted by STRAIGHT analysis.*

Comparative Method	**Source of mel-cepstrum**	**Waveform Synthesis**
Plain-MLSA	STFT	MLSA filter
STRAIGHT-MLSA	Spectrum envelop	MLSA filter
Plain-WaveNet	STFT	WaveNet
STRAIGHT-WaveNet	Spectrum envelop	WaveNet

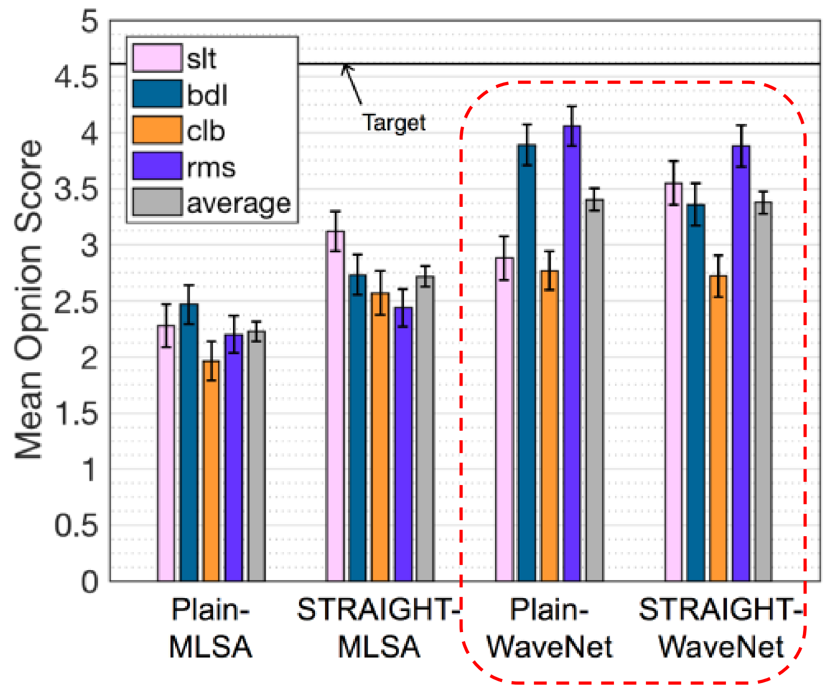


Figure 3: *Sound quality of synthesized speech*

Training data: 1 hour per each speaker

출처: A. Tamamory, et. al., “Speaker-dependent WaveNet vocoder,” in *Proc. Interspeech*, 2017.

WaveNet synthesis

WaveNet 모델의 성능을 더 높일 수 있는 방법

Table 1: *Comparative methods of waveform synthesis; spectrum envelop was extracted by STRAIGHT analysis.*

Comparative Method	Source of mel-cepstrum	Waveform Synthesis
Plain-MLSA	STFT	MLSA filter
STRAIGHT-MLSA	Spectrum envelop	MLSA filter
Plain-WaveNet	STFT	WaveNet
STRAIGHT-WaveNet	Spectrum envelop	WaveNet

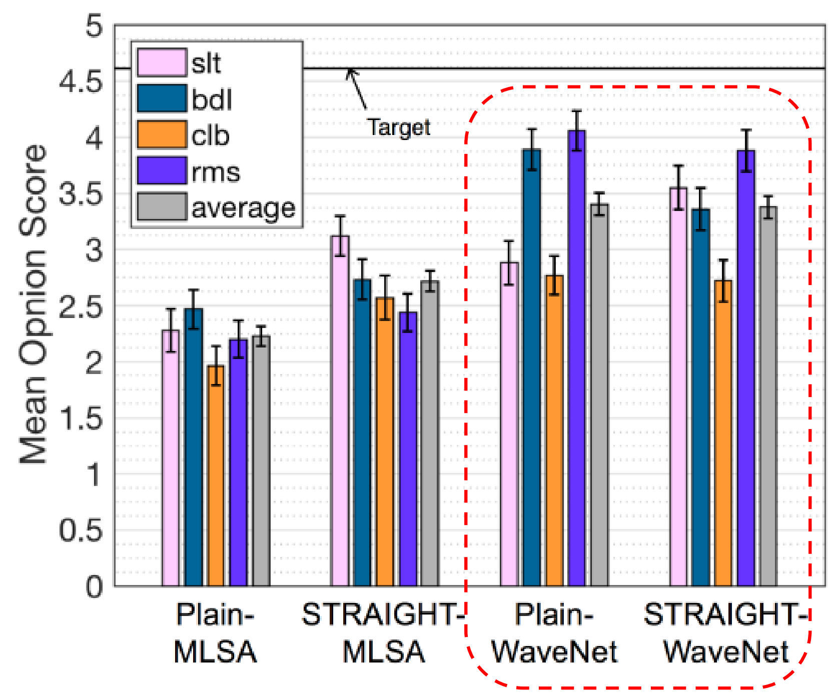


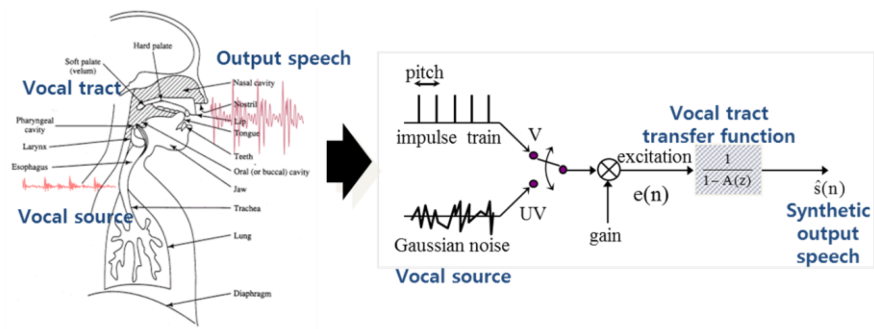
Figure 3: *Sound quality of synthesized speech*

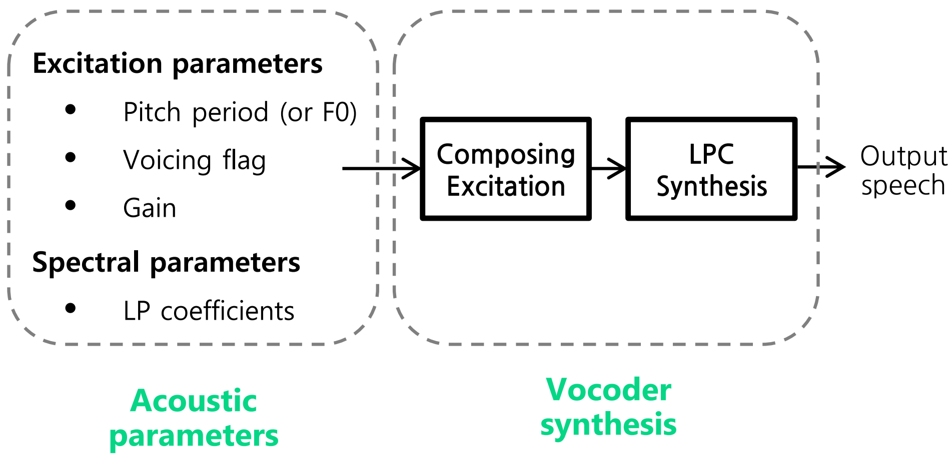
Training data: 1 hour per each speaker

출처: A. Tamamory, et. al., “Speaker-dependent WaveNet vocoder,” in *Proc. Interspeech*, 2017.

Recall: Parametric LPC vocoder

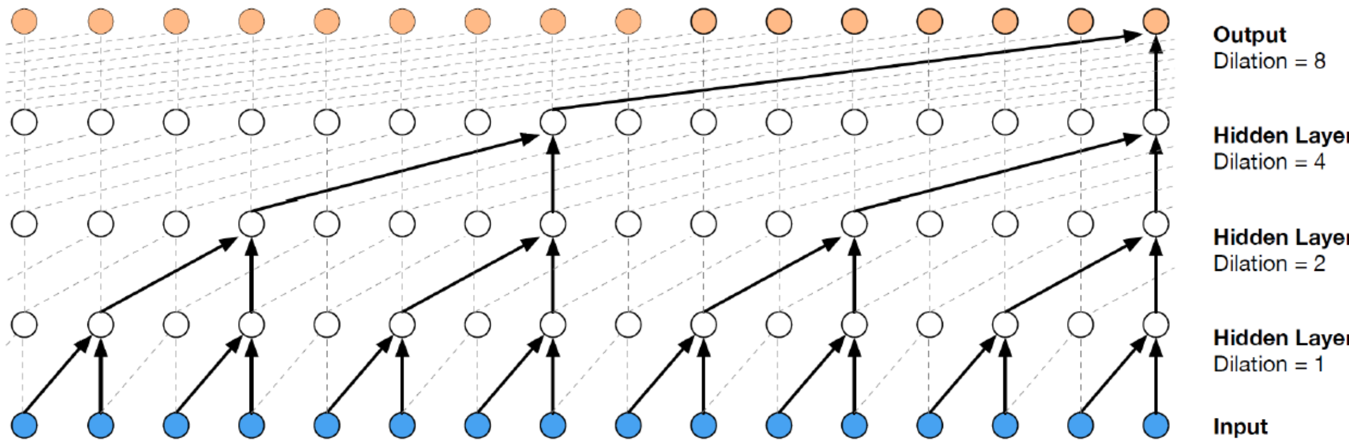
Excitation 신호를 추정하고 LPC Synthesis Filter를 이용해 음성을 만드는 방법





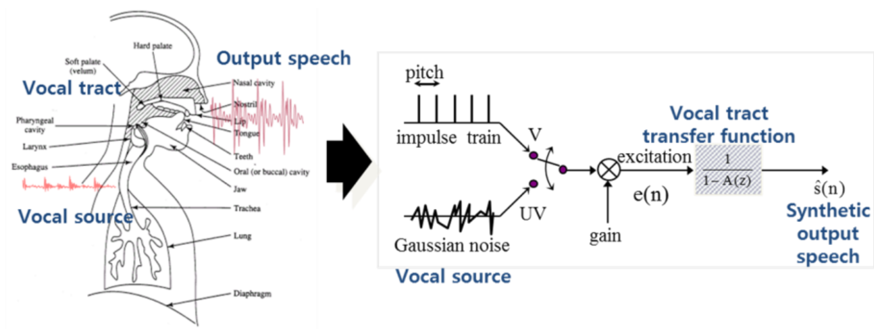
Recall: WaveNet vocoder

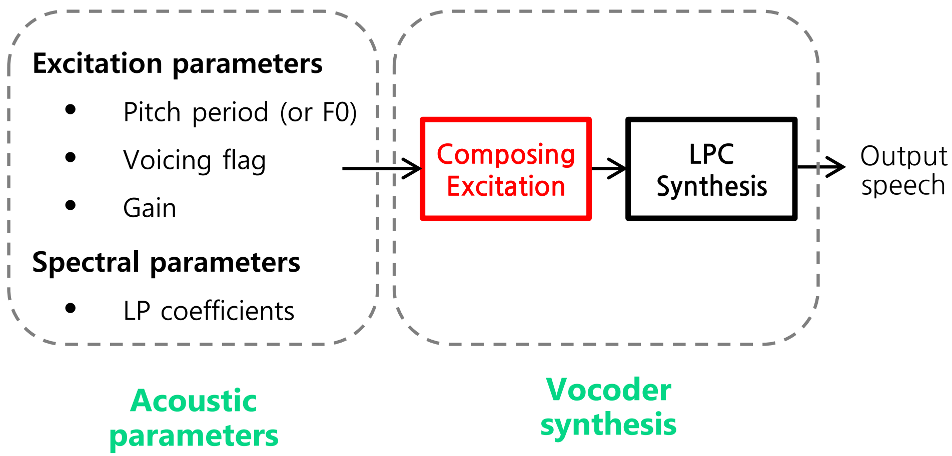
Time-domain 의 **음성** 샘플을 직접 추정하는 방법



Recall: WaveNet vocoder

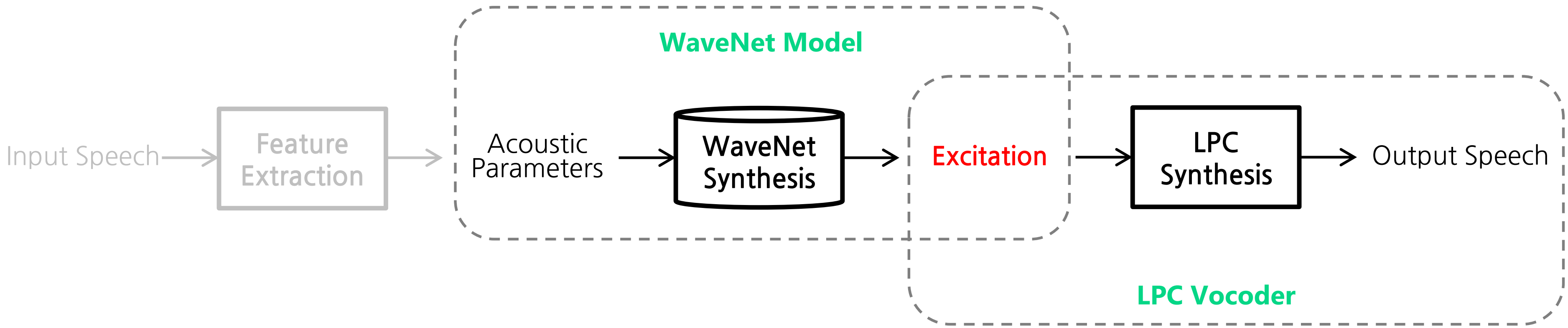
WaveNet 모델로 Time-domain 의 **Excitation** 샘플을 직접 추정한다면?





Neural excitation vocoder

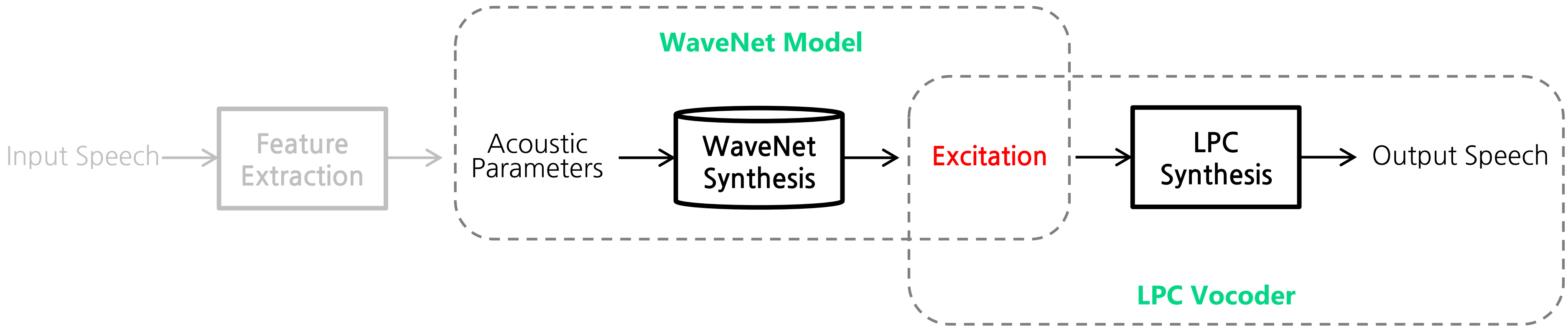
합성음 품질을 더욱 높힐 수 있다 !



Recorded speech | TTS + LPC vocoder | TTS + WaveNet vocoder | TTS + LP-WaveNet vocoder

Neural excitation vocoder

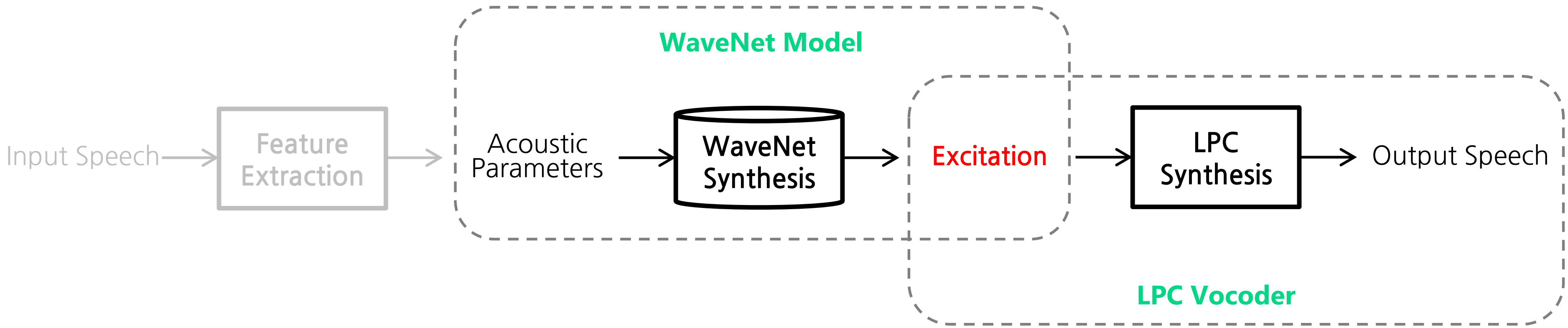
합성음 품질을 더욱 높힐 수 있다 !

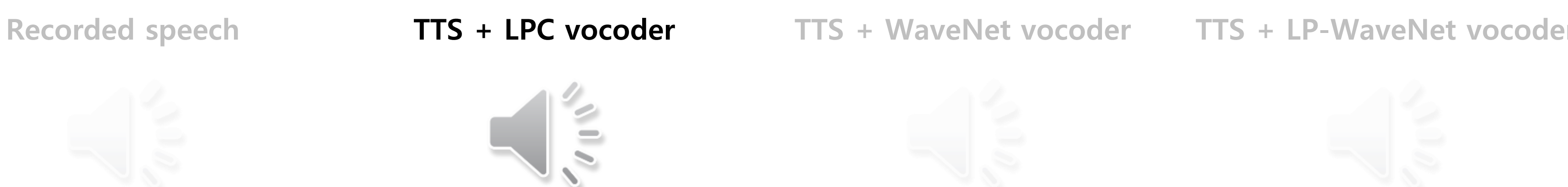


Recorded speech TTS + LPC vocoder TTS + WaveNet vocoder TTS + LP-WaveNet vocoder

Neural excitation vocoder

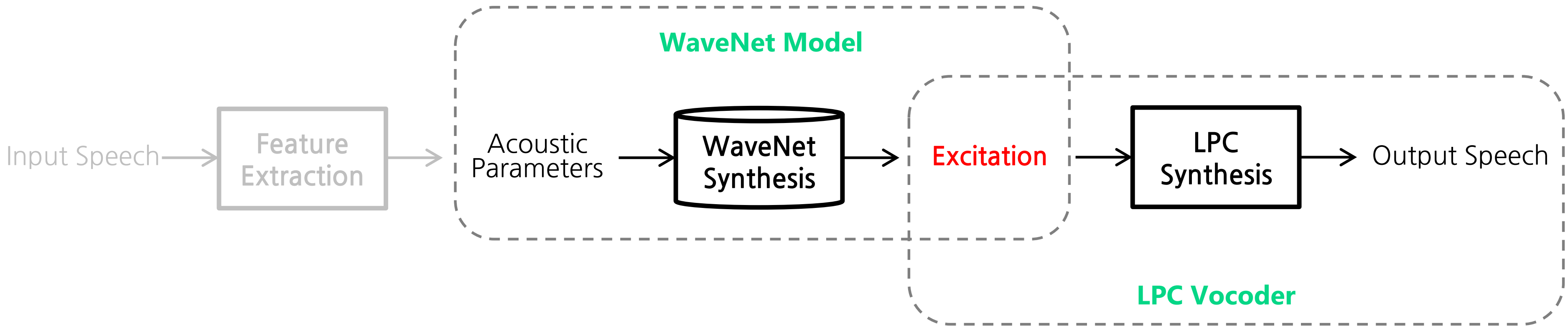
합성음 품질을 더욱 높힐 수 있다 !





Neural excitation vocoder

합성음 품질을 더욱 높힐 수 있다 !



Recorded speech

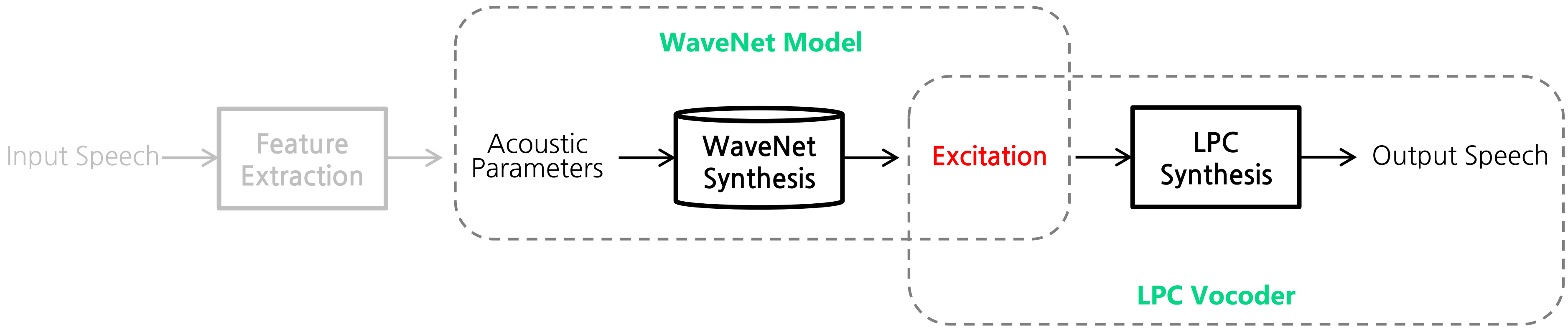
TTS + LPC vocoder

TTS + WaveNet vocoder

TTS + LP-WaveNet vocoder

Neural excitation vocoder

합성음 품질을 더욱 높힐 수 있다 !



Recorded speech

TTS + LPC vocoder

TTS + WaveNet vocoder

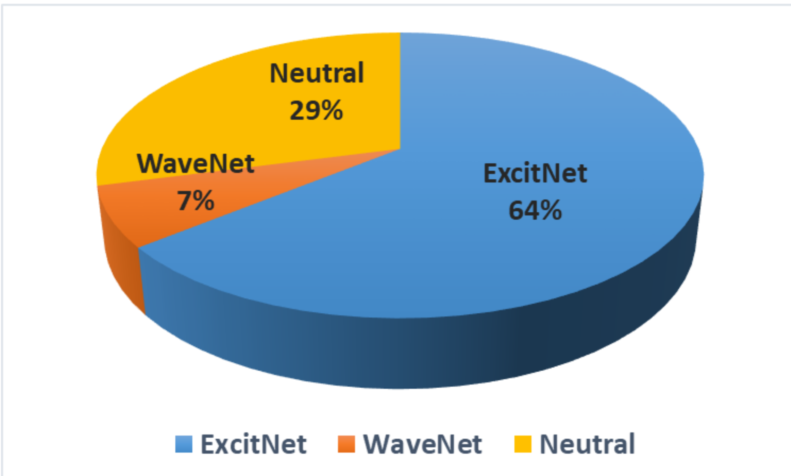
TTS + LP-WaveNet vocoder

Neural excitation vocoder

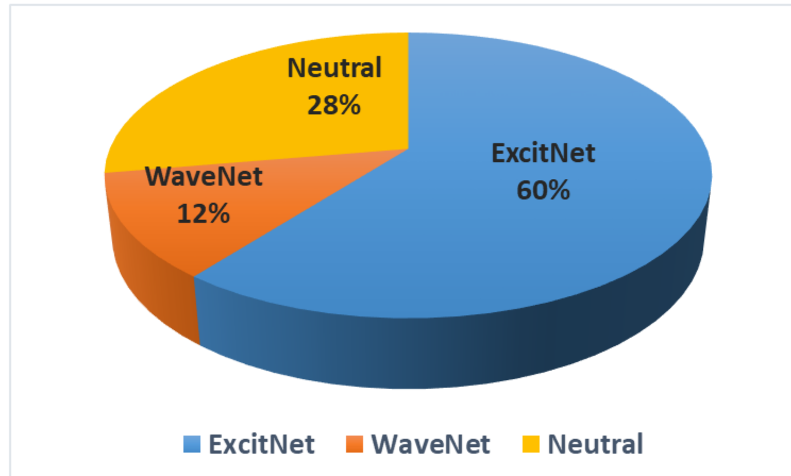
Engineering Day 2019

합성음 품질을 더욱 높힐 수 있다 !

Korean female speaker



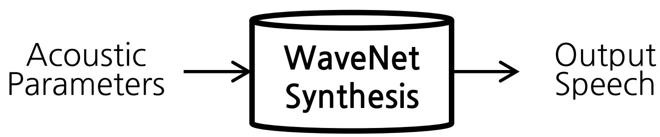
Korean male speaker

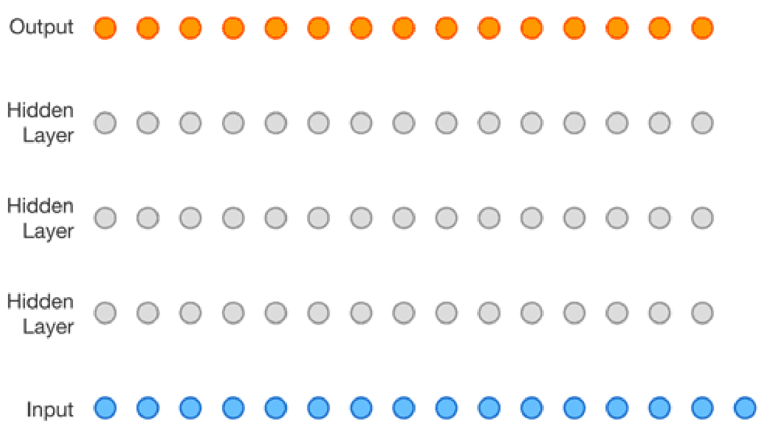


Recorded speech | TTS + LPC vocoder | TTS + WaveNet vocoder | TTS + LP-WaveNet vocoder

Summary

WaveNet Vocoder 를 꼭 기억해 주세요!





Autoregressive WaveNet vocoder

- Sample-by-sample generation
  - $p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^{T} p(x_t | x_1, \cdots, x_{t-1}, \mathbf{h})$
  - $\mathbf{h}$: Conditional acoustic parameter

Neural excitation vocoder

- WaveNet + LPC synthesis
  - GlottNet, ExcitNet, LP-WaveNet …

Similar approaches

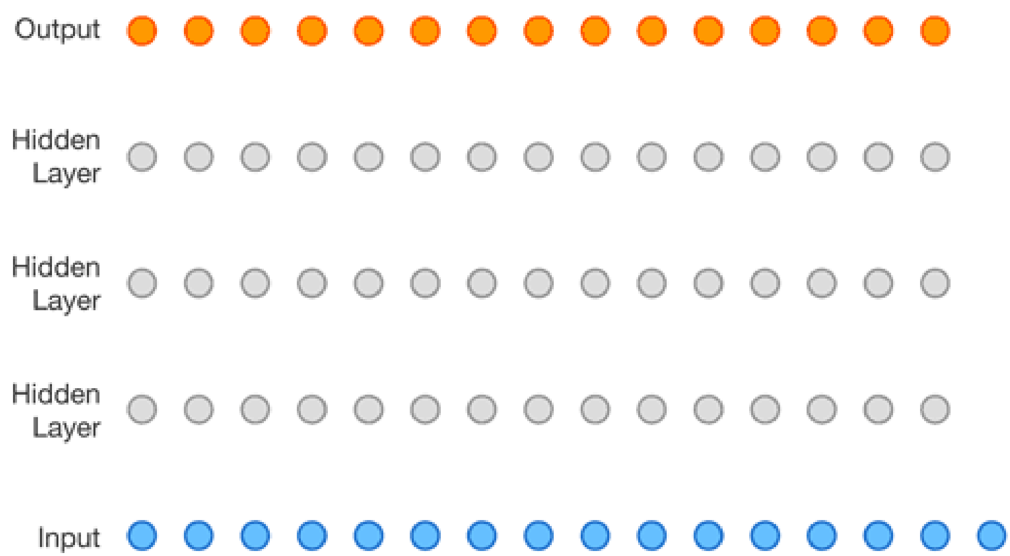
- WaveRNN, SampleRNN vocoder
  - RNN-based generation (cf. WaveNet: CNN)
  - LPCNet: WaveRNN + LPC synthesis

Vocoding model

Non-autoregressive WaveNet synthesis

Recall

현재 음성 신호를 예측할 때 **과거 음성** 신호를 함께 사용하는 방법: **Autoregressive** Model

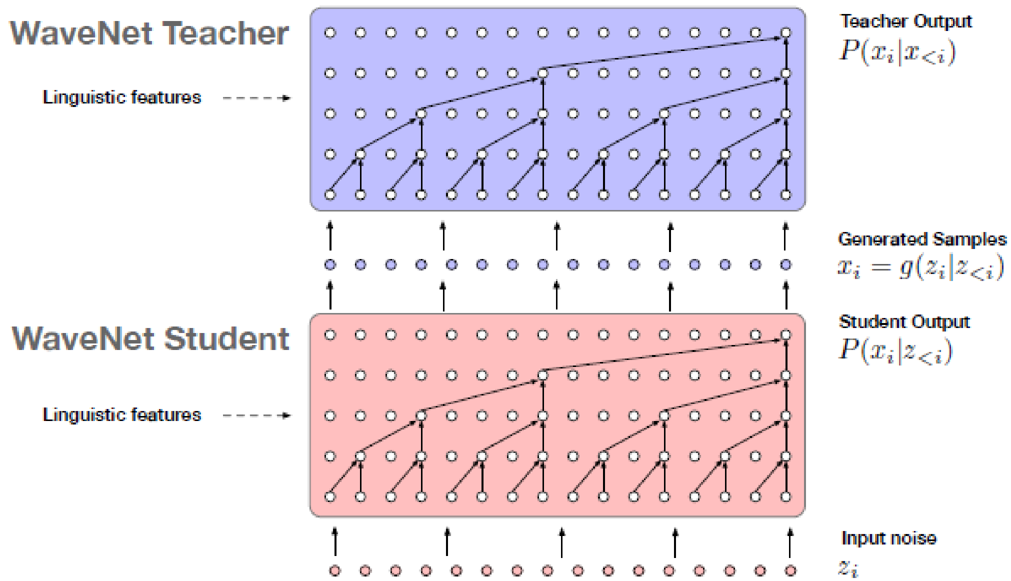


Autoregressive Model 은 고품질의 음성을 생성할 수 있으나,
1초 음성을 만들 때 약 5분 정도의 시간이 소요된다는 치명적인 문제가 있습니다.

출처: https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

Parallel WaveNet

음성 신호를 **Parallel** 방식으로 예측하는 방법: **Non-autoregressive** Model

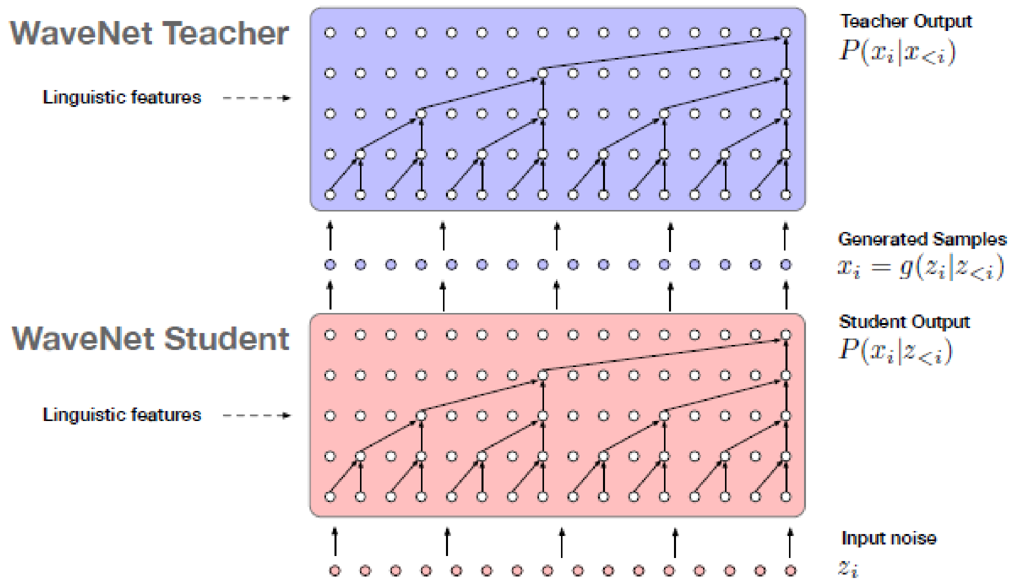


WaveNet 의 속도 문제를 해결하기 위해 제안된 방법이
Non-autoregressive 구조의 **Parallel WaveNet** 입니다.

출처: A. van den Oord, et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. ICML*, 2018.

Parallel WaveNet

음성 신호를 **Parallel** 방식으로 예측하는 방법: **Non-autoregressive** Model

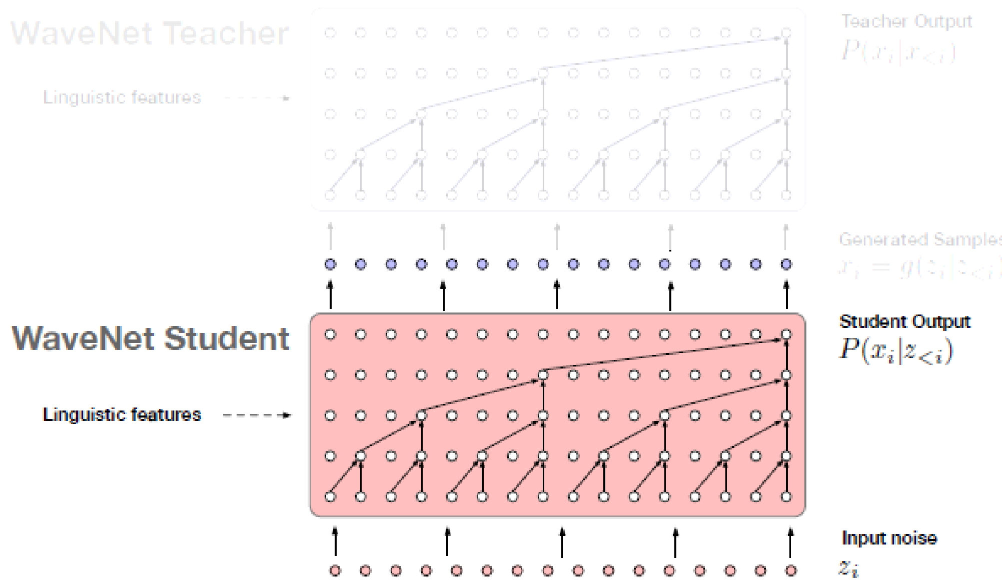


Autoregressive WaveNet (=Teacher) 모델의 확률 분포를
Non-autoregressive Parallel WaveNet (=Student) 모델이 배우도록 훈련합니다.

출처: A. van den Oord, et al., “Parallel WaveNet: Fast high-fidelity speech synthesis,” in *Proc. ICML*, 2018.

Parallel WaveNet

음성 신호를 **Parallel** 방식으로 예측하는 방법: **Non-autoregressive** Model

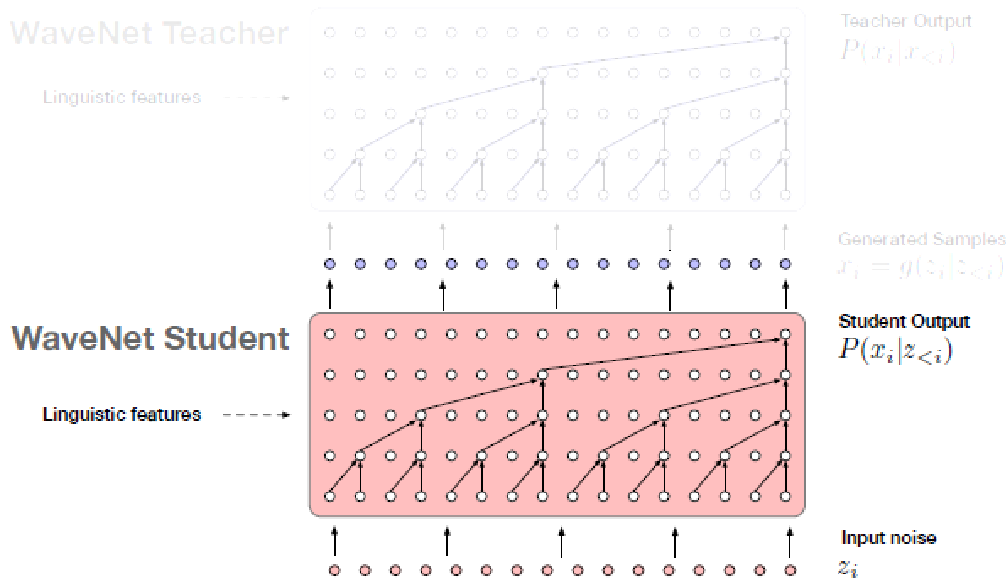


Non-autoregressive Parallel WaveNet 모델은
과거 음성을 사용하지 않으므로, 생성 속도에 제한이 없습니다.
(1초 음성을 약 0.02초 만에 생성 가능)

출처: A. van den Oord, et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. ICML*, 2018.

Parallel WaveNet

음성 신호를 **Parallel** 방식으로 예측하는 방법: **Non-autoregressive** Model

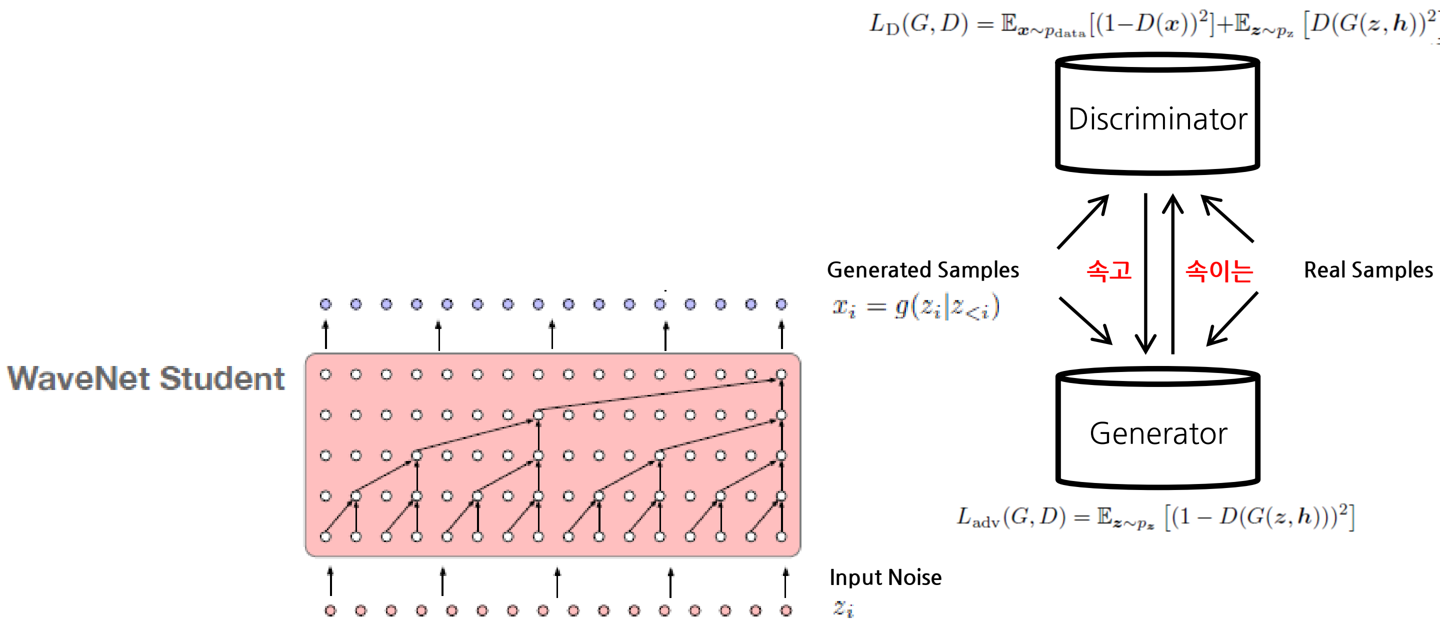


하지만 그만큼 모델 학습 방법이 어려워서

출처: A. van den Oord, et al., “Parallel WaveNet: Fast high-fidelity speech synthesis,” in *Proc. ICML*, 2018.

Parallel WaveGAN

음성 신호를 **Parallel** 방식으로 예측하는 방법: **Non-autoregressive** Model



GAN 을 이용해서 Non-autoregressive WaveNet 을 직접 학습합니다.

출처: R. Yamamoto, et al., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, 2020.

Parallel WaveGAN

Engineering Day 2020

음성 신호를 **Parallel** 방식으로 예측하는 방법: **Non-autoregressive** Model

Autoregressive WaveNet

합성음 품질이 좋지만
생성 속도가 느리다

300 RT

Parallel WaveGAN

학습도 쉽고
생성 속도도 빠르고
합성음 품질도 좋다

RT: 1초 음성을 생성할 때 걸리는 시간

Engineering Day 2020

Parallel WaveGAN

음성 신호를 **Parallel** 방식으로 예측하는 방법: **Non-autoregressive** Model

Autoregressive WaveNet

합성음 품질이 좋지만
생성 속도가 느리다

Parallel WaveGAN

학습도 쉽고
생성 속도도 빠르고
합성음 품질도 좋다

0.02 RT

RT: 1초 음성을 생성할 때 걸리는 시간

Summary

Autoregressive 생성 방법과 Non-autoregressive 생성 방법을 꼭 기억해 주세요!

Autoregressive vocoder

- Sample-by-sample generation
  - $p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^{T} p(x_t|x_1, \cdots, x_{t-1}, \mathbf{h})$
  - $\mathbf{h}$: Conditional acoustic parameter

Non-autoregressive vocoder

- Parallel generation
  - $p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^{T} p(x_t|z_1, \cdots, z_{t-1}, \mathbf{h})$
  - z_i: Random variable
  - $\mathbf{h}$: Conditional acoustic parameter

Teacher-student distillation

- Parallel WaveNet, ClariNet

GAN-based approaches

- Parallel WaveGAN
- MelGAN, VocGAN, Hi-Fi GAN

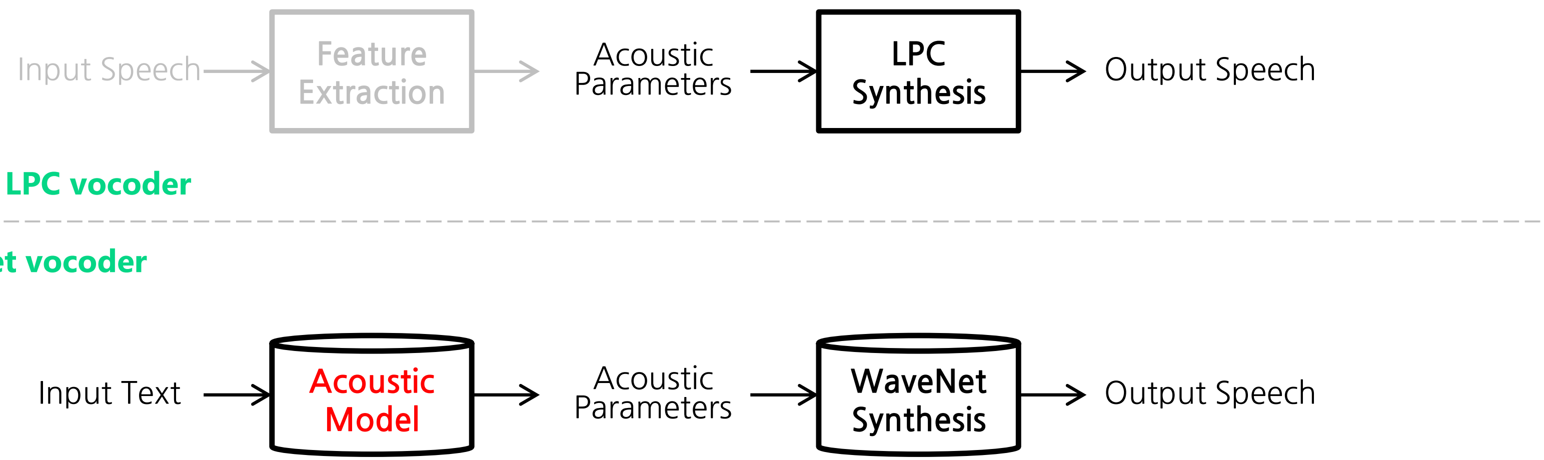
CLOVA

Acoustic model

Statistical parametric speech synthesis

Recall

Acoustic model 은 **Text** 로부터 **Acoustic Parameter** 를 추정하는 역할을 합니다.



Tacotron 2

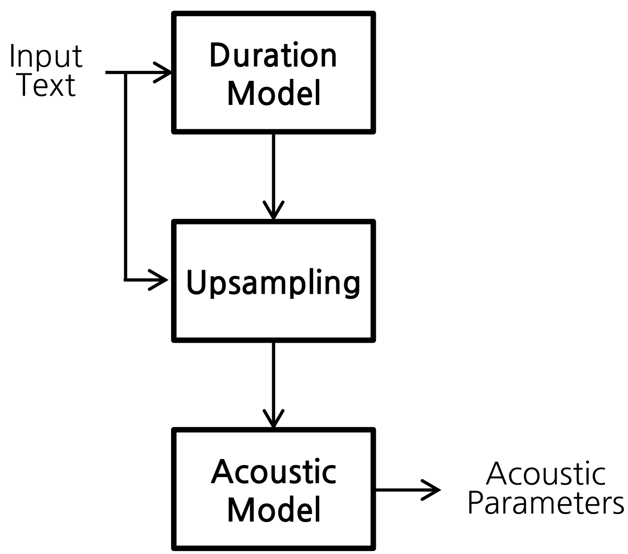
출처: J. Shen, et. al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018.

Overview

Acoustic model 은 **Text** 로부터 **Acoustic Parameter** 를 추정하는 역할을 합니다.

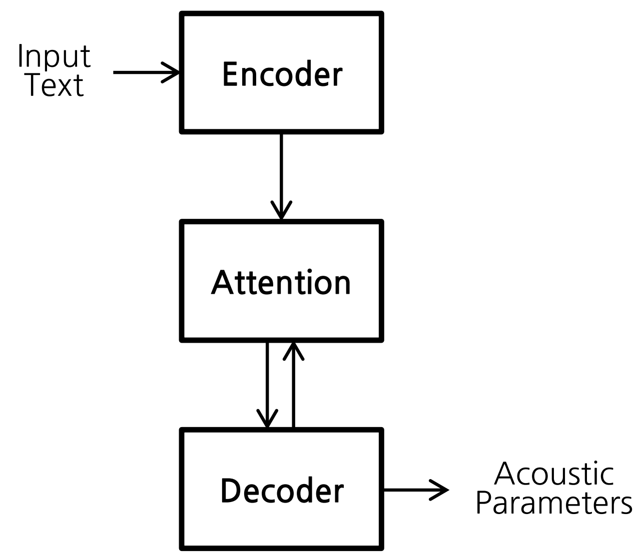
Statistical parametric speech synthesis

- Simple deep learning model (FF+LSTM)

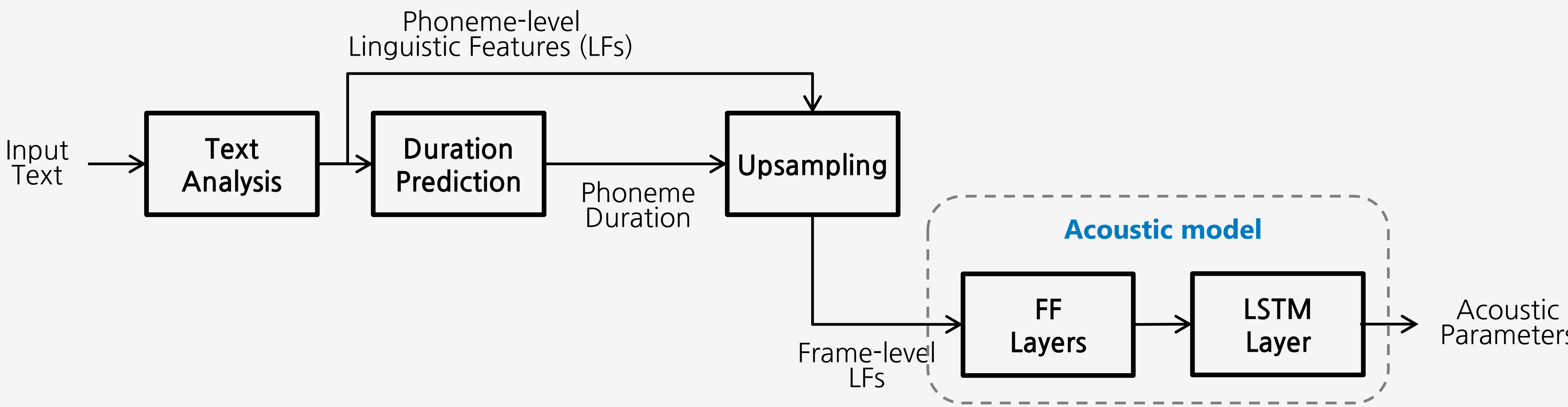


End-to-end speech synthesis

- Seq2seq model

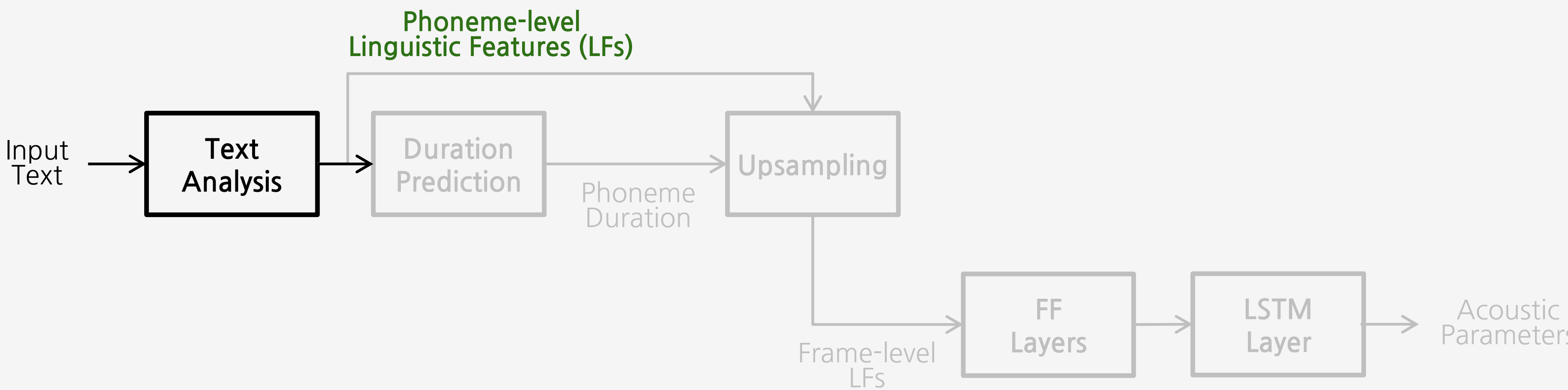


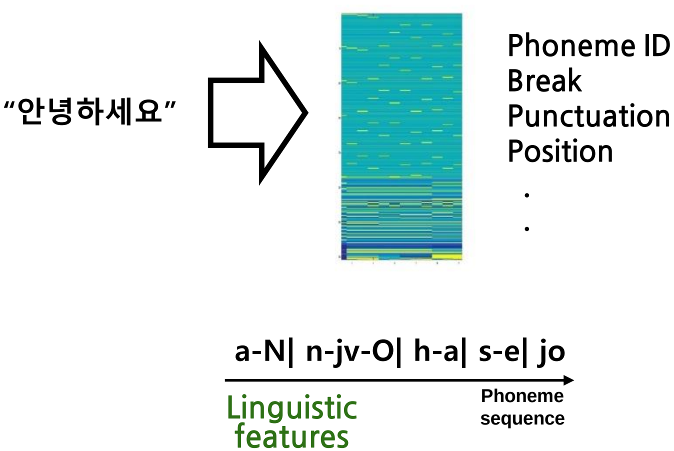
Statistical parametric speech synthesis (SPSS)



Statistical parametric speech synthesis (SPSS)

Text analyzer: Generates phoneme-level linguistic features (Phoneme: 음운론상의 최소 단위)

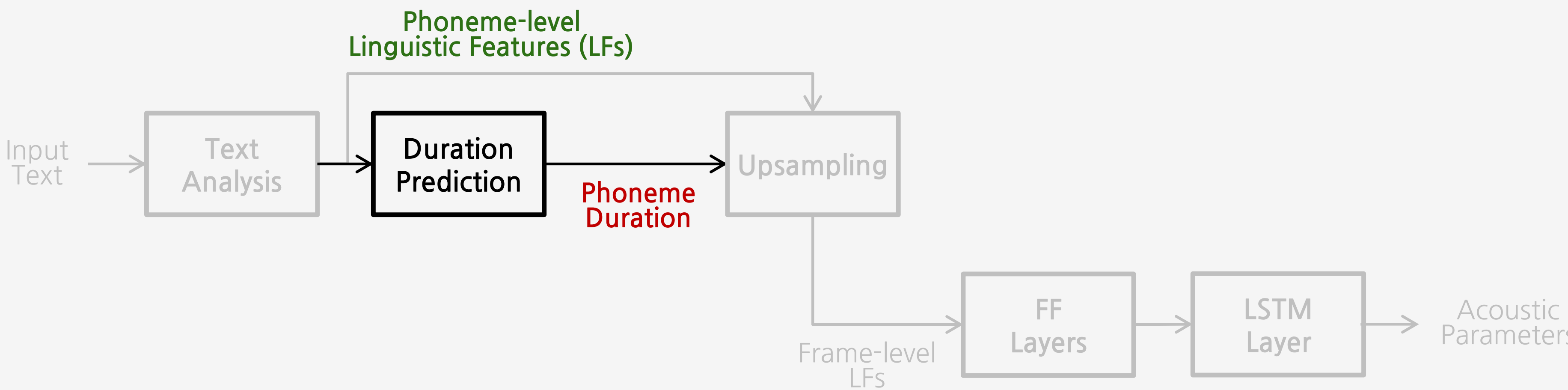


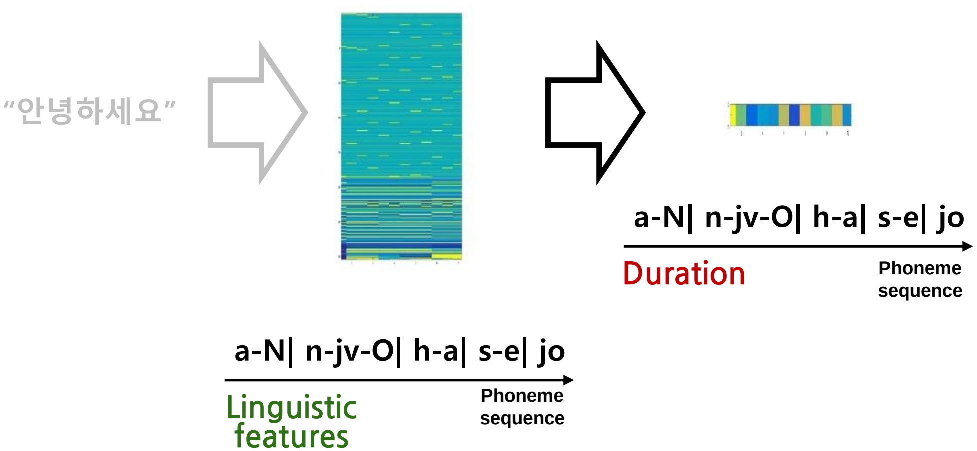



```
WD=[안녕하세요] PR=[a00 NX13 n00 jv00 OX13 h00 a03 s00 e03 jo04] BR=[6] OWD
WD=[눈이] PR=[n00 u03 n00 i04] OWD=[눈이] OPR=[누니] ONPR=[누니] DOM=[0] E
WD=[마주치자] PR=[m00 a03 z00 u03 c00 i03 z00 a04] BR=[6] OWD=[마주치자] OPR
WD=[가쁜] PR=[g00 a03 B00 U00 NX14] OWD=[가쁜] OPR=[가쁜] ONPR=[가쁜] DOM
WD=[숨] PR=[s00 u00 MX14] BR=[3] OWD=[숨] OPR=[숨] ONPR=[숨] DOM=[0] EM
WD=[사이로] PR=[s00 a03 i03 r00 o04] OWD=[사이로] OPR=[사이로] ONPR=[사이로]
WD=[미소] PR=[m00 i03 s00 o04] OWD=[미소] OPR=[미소] ONPR=[미소] DOM=[0] E
WD=[섞인] PR=[s00 v03 G00 i04] BR=[3] OWD=[섞인] OPR=[서끼] ONPR=[서끼] DOM
WD=[인사가] PR=[n00 i00 NX13 s00 a03 g00 a04] OWD=[인사가] OPR=[닌사가] ONP
WD=[배어] PR=[b00 e03 v04] OWD=[배어] OPR=[배어] ONPR=[베어] DOM=[0] EMO=
WD=[나온다] PR=[n00 a03 o00 NX13 d00 a04] PUNCT=[.] BR=[7] OWD=[나온다.] OP
```

Statistical parametric speech synthesis (SPSS)

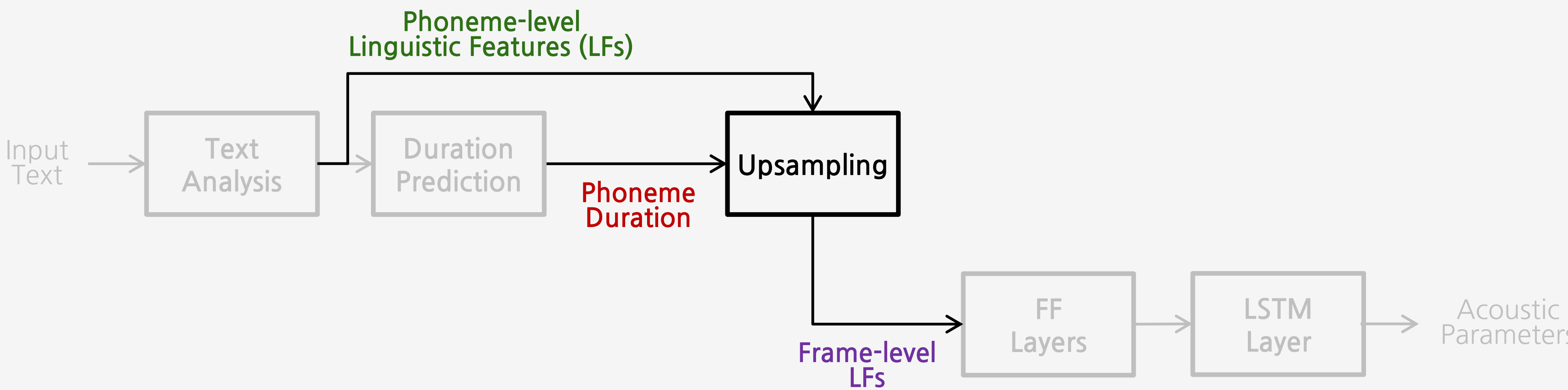
Duration model: Predicts phoneme duration

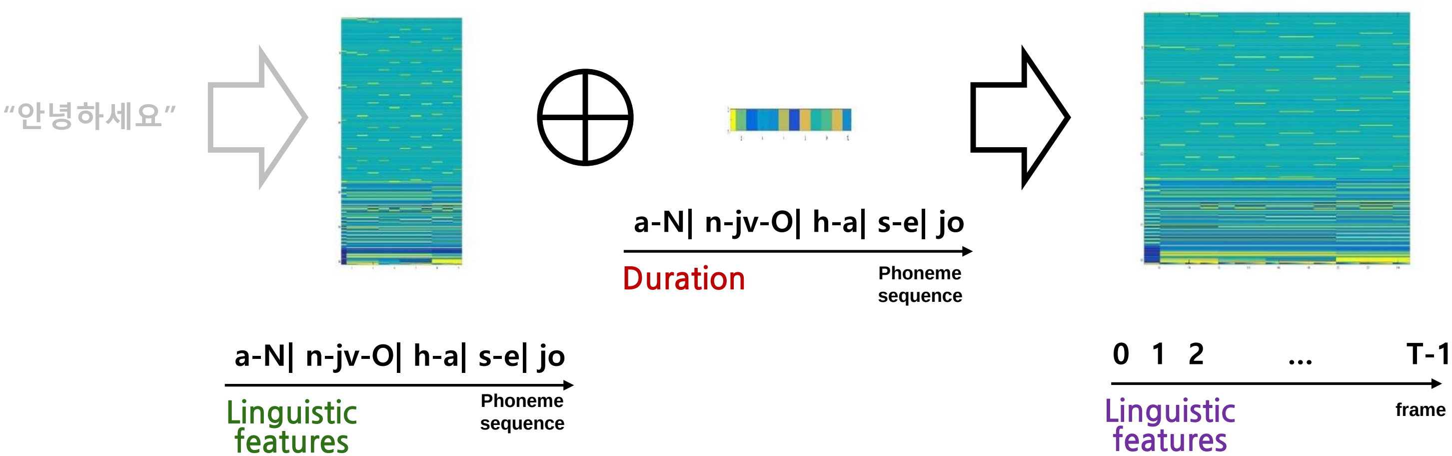




Statistical parametric speech synthesis (SPSS)

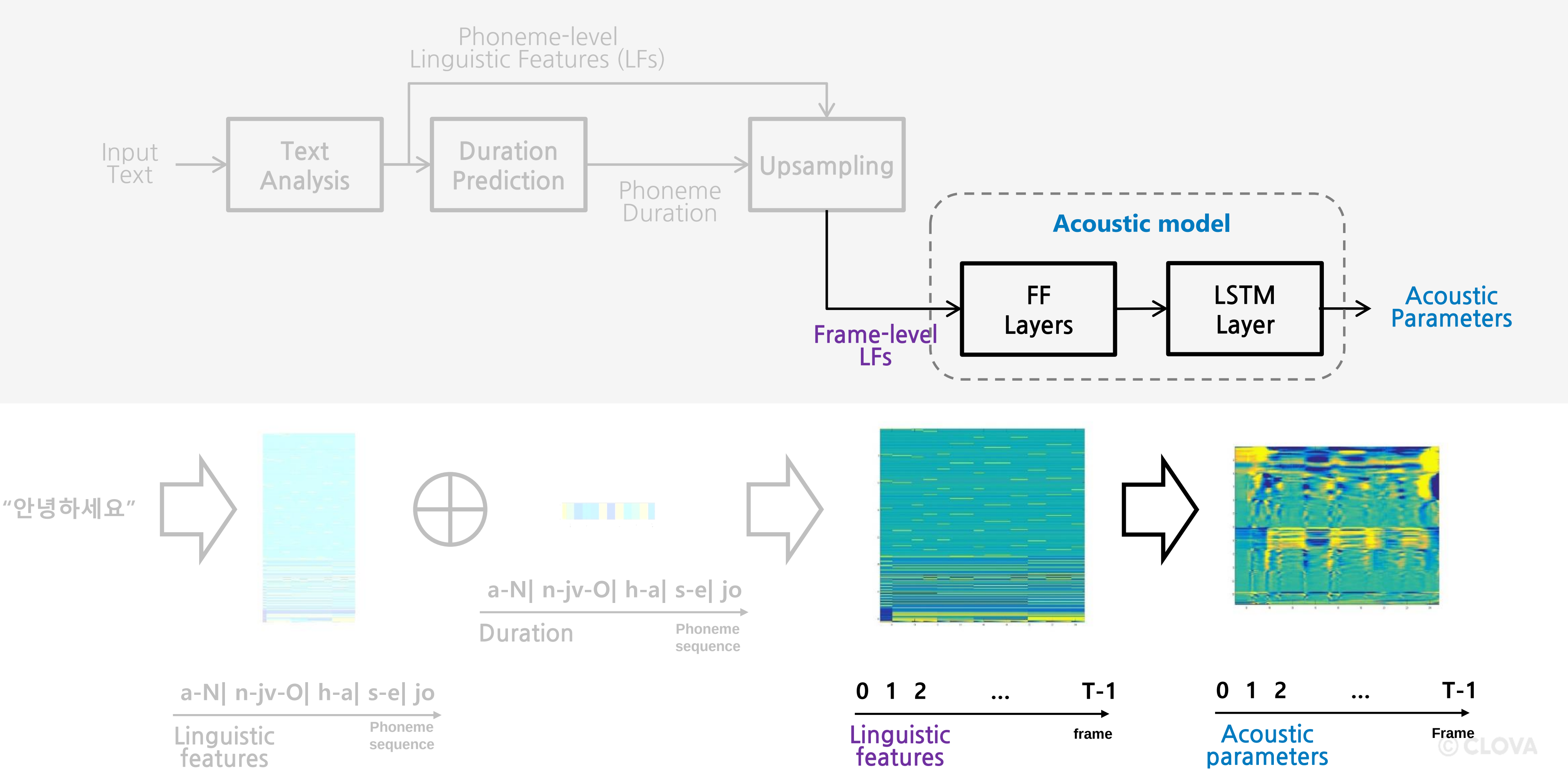
Linguistic upsampler: Generates frame-level linguistic features





Statistical parametric speech synthesis (SPSS)

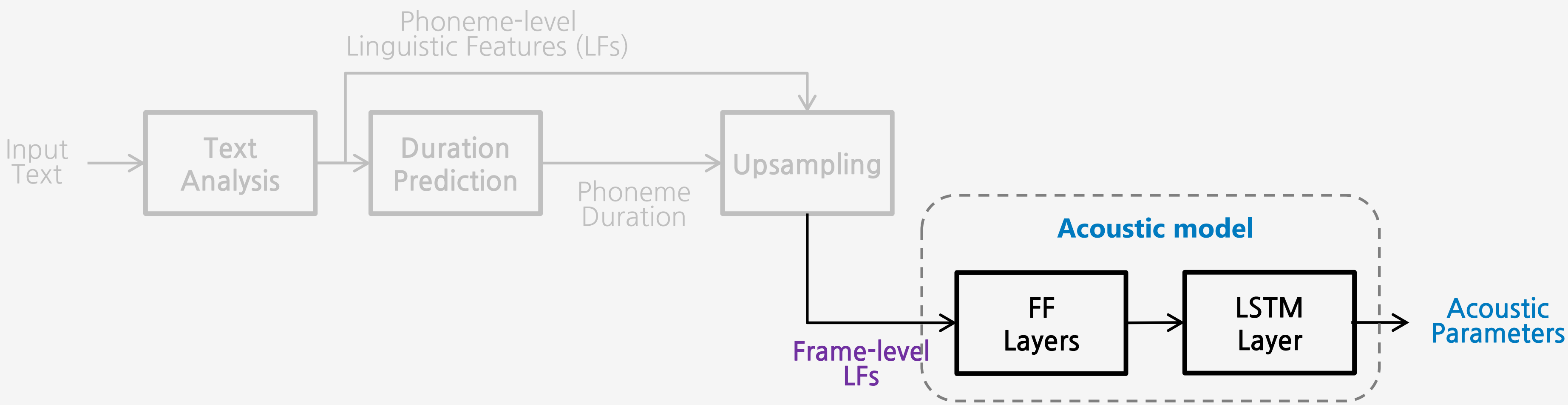
Acoustic model: Predicts frame-level acoustic parameters

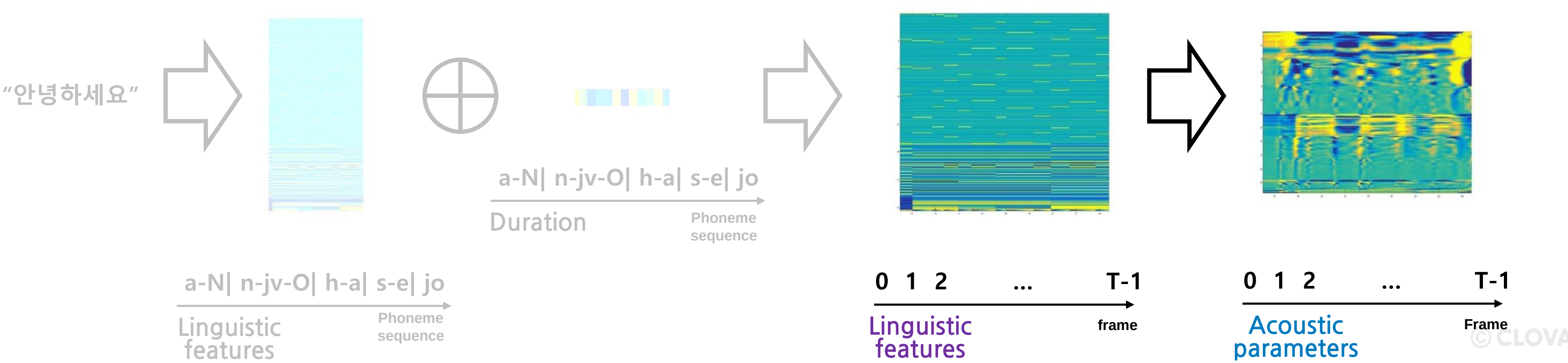


Statistical parametric speech synthesis (SPSS)

N Innovation Award 2020

Acoustic model: Predicts frame-level acoustic parameters





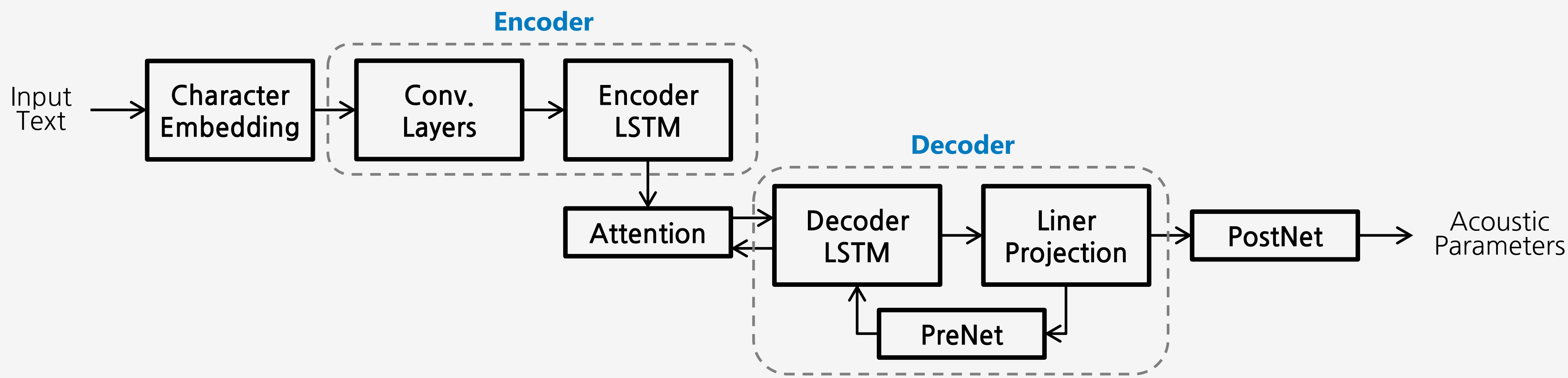
CLOVA

Acoustic model

End-to-end speech synthesis

End-to-end speech synthesis

(Text) **Encoder** 와 (Acoustic Parameter) **Decoder** 를 만들고, **Attention** 으로 Alignment 를 잡아주면 됩니다.

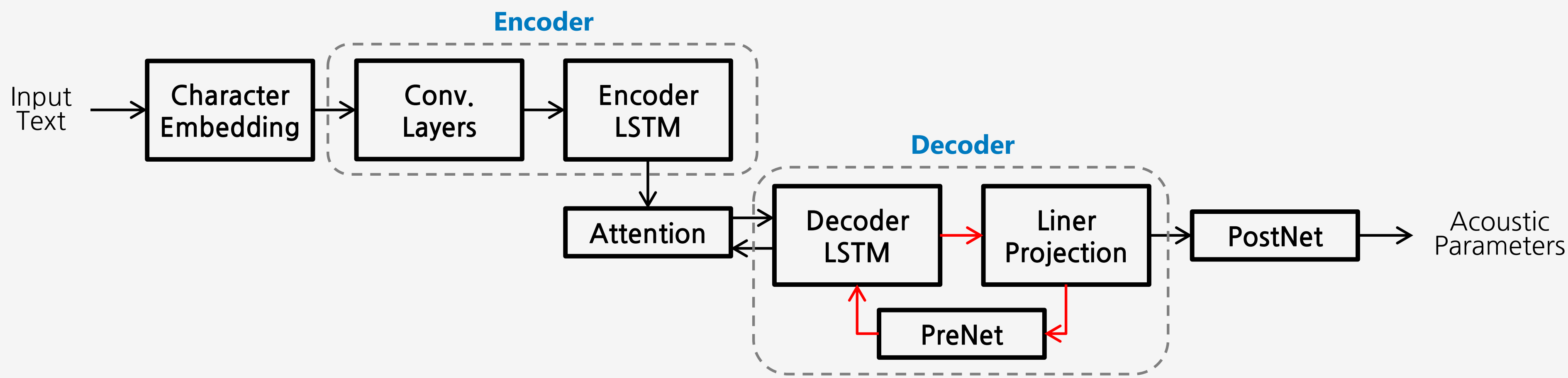


Tacotron 2

출처: J. Shen, et. al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, 2018.

End-to-end speech synthesis

(Text) **Encoder** 와 (Acoustic Parameter) **Decoder** 를 만들고, **Attention** 으로 Alignment 를 잡아주면 됩니다.



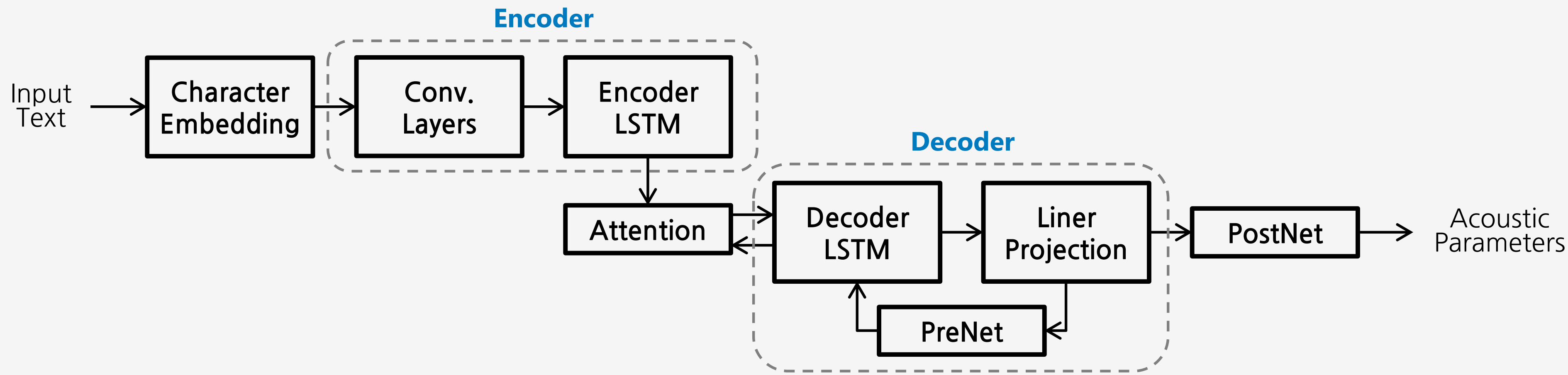
Seq2seq model with attention Phoneme Duration 없어도됨

Autoregressive acoustic model Acoustic Parameter 추정 정확도가 높아짐

출처: J. Shen, et. al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, 2018.

End-to-end speech synthesis

(Text) **Encoder** 와 (Acoustic Parameter) **Decoder** 를 만들고, **Attention** 으로 Alignment 를 잡아주면 됩니다.



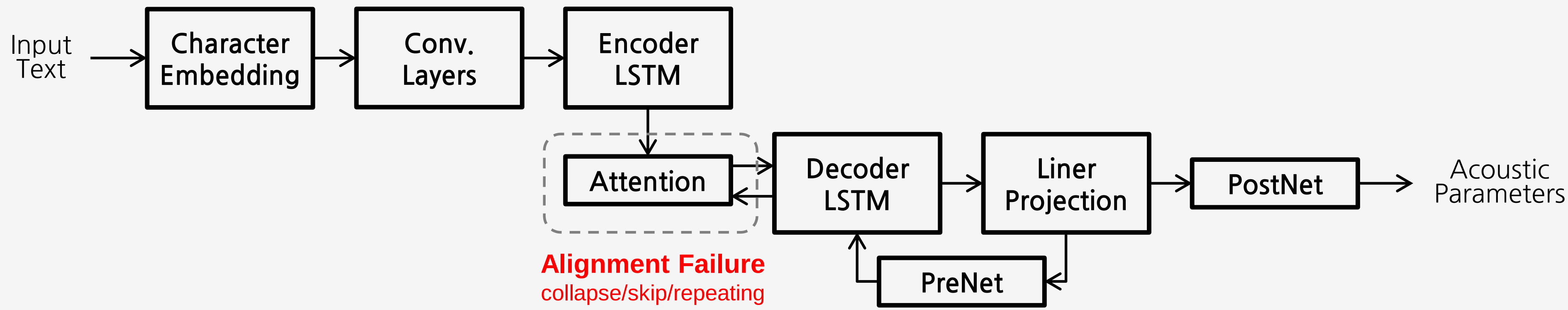
System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

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

출처: J. Shen, et. al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018.

End-to-end speech synthesis

(Text) **Encoder** 와 (Acoustic Parameter) **Decoder** 를 만들고, **Attention** 으로 Alignment 를 잡아주면 됩니다.



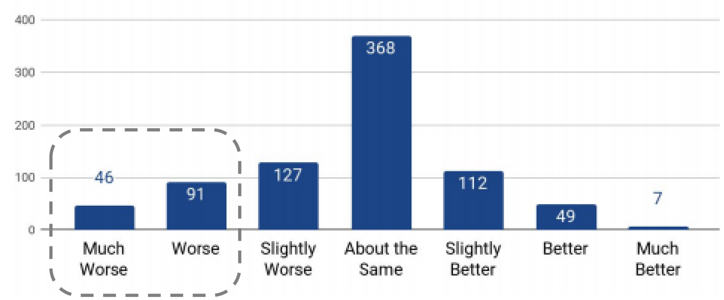
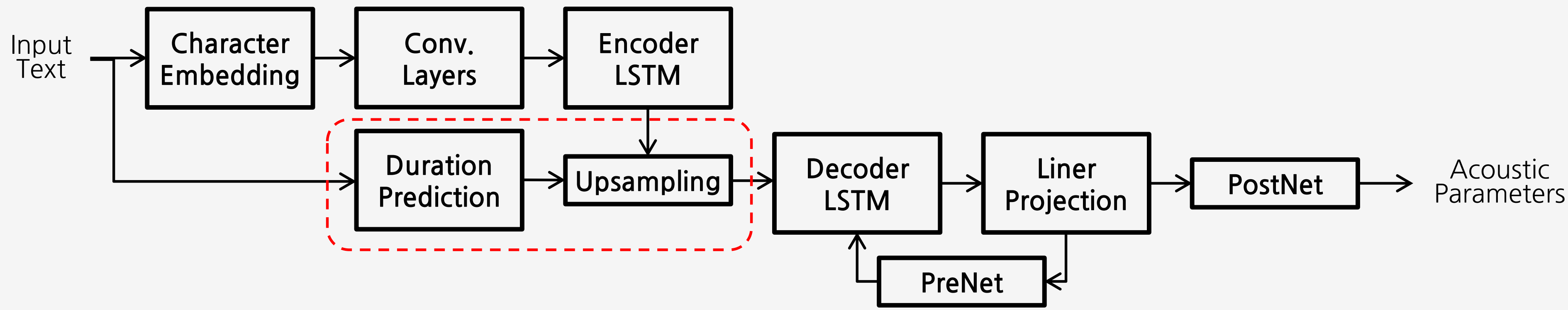


Fig. 2. Synthesized vs. ground truth: 800 ratings on 100 items.

출처: J. Shen, et. al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018.

End-to-end speech synthesis

(Text) **Encoder** 와 (Acoustic Parameter) **Decoder** 를 만들고, **Duraion Model** 로 Alignment 를 잡아주면 됩니다.



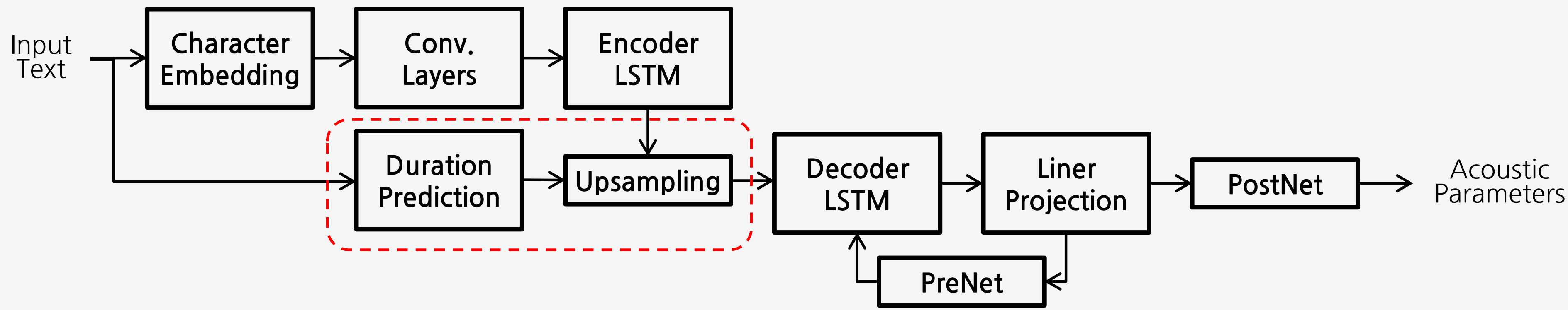
Alignment failure

w/ duration model

출처: T. Okamoto, et al., “Tacotron-based acoustic model using phoneme alignment for practical neural text-to-speech systems,” in *Proc. ASRU*, 2019.

End-to-end speech synthesis

(Text) **Encoder** 와 (Acoustic Parameter) **Decoder** 를 만들고, **Duraion Model** 로 Alignment 를 잡아주면 됩니다.



Alignment failure

w/ duration model

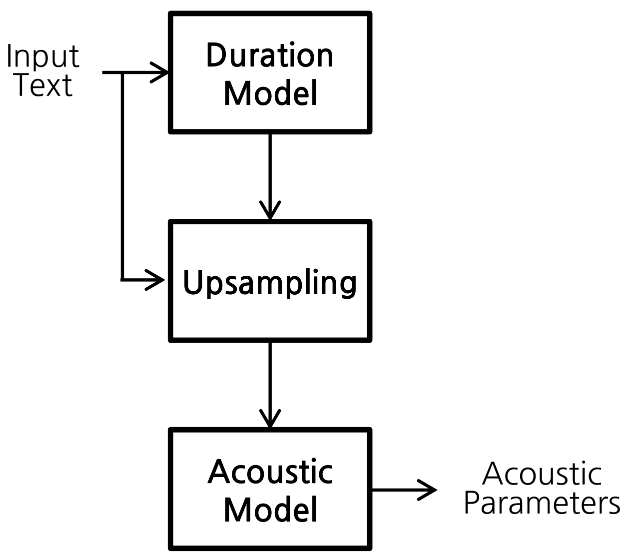
출처: T. Okamoto, et al., “Tacotron-based acoustic model using phoneme alignment for practical neural text-to-speech systems,” in *Proc. ASRU*, 2019.

Summary

Acoustic model 은 **Text** 로부터 **Acoustic Parameter** 를 추정하는 역할을 합니다.

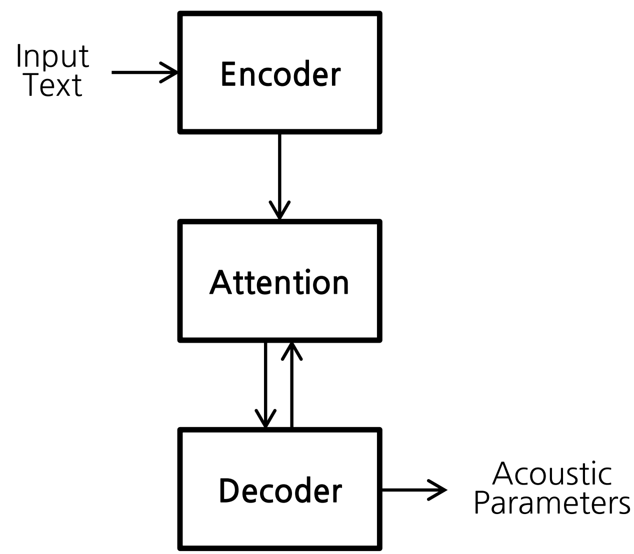
Statistical parametric speech synthesis

- Simple deep learning model (FF+LSTM)



End-to-end speech synthesis

- Seq2seq model



Summary

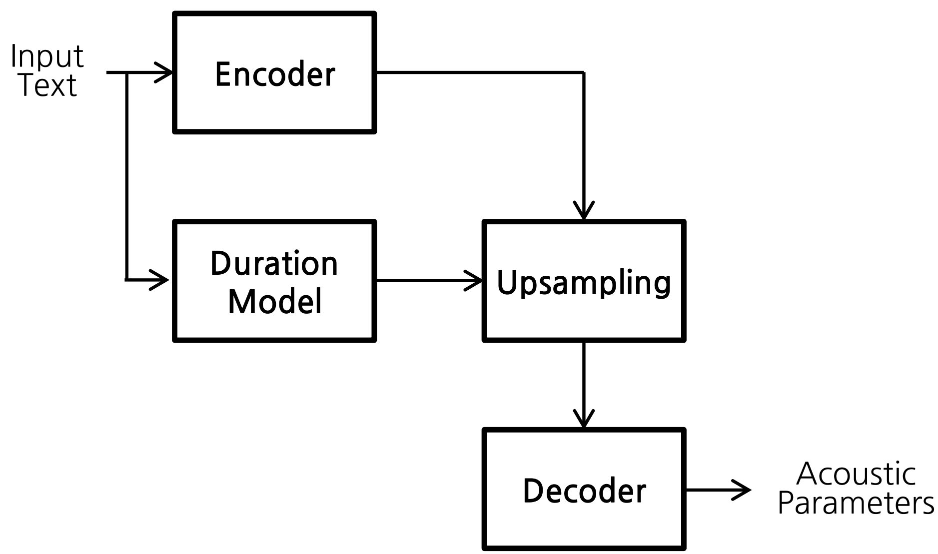
Acoustic model 은 **Text** 로부터 **Acoustic Parameter** 를 추정하는 역할을 합니다.

Statistical parametric speech synthesis

- Simple deep learning model (FF+LSTM)

End-to-end speech synthesis

- Seq2seq model



Summary

Acoustic model 은 **Text** 로부터 **Acoustic Parameter** 를 추정하는 역할을 합니다.

Statistical parametric speech synthesis

- Simple deep learning model (FF+LSTM)

End-to-end speech synthesis

- Autoregressive models
  - Tacotron 1, 2
  - Transformer
- Non-autoregressive model
  - FastSpeech 2, Parallel Tacotron

Summary

Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.

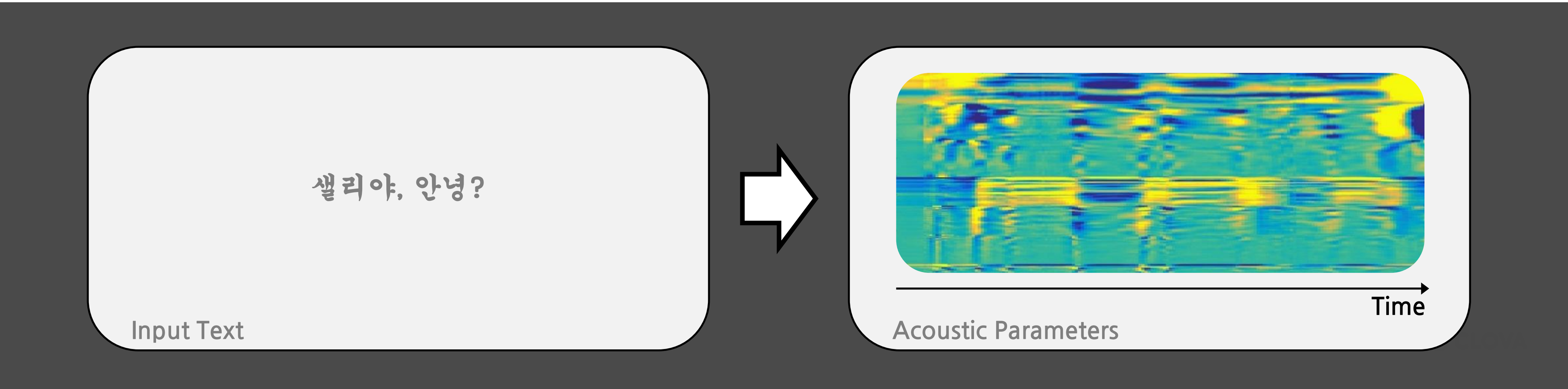
DNN TTS = Acoustic model + Vocoder

Summary

Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.

톤의 높낮이, 음색, 어조, 강세 등

텍스트에서 **Acoustic Parameter** 를 추정

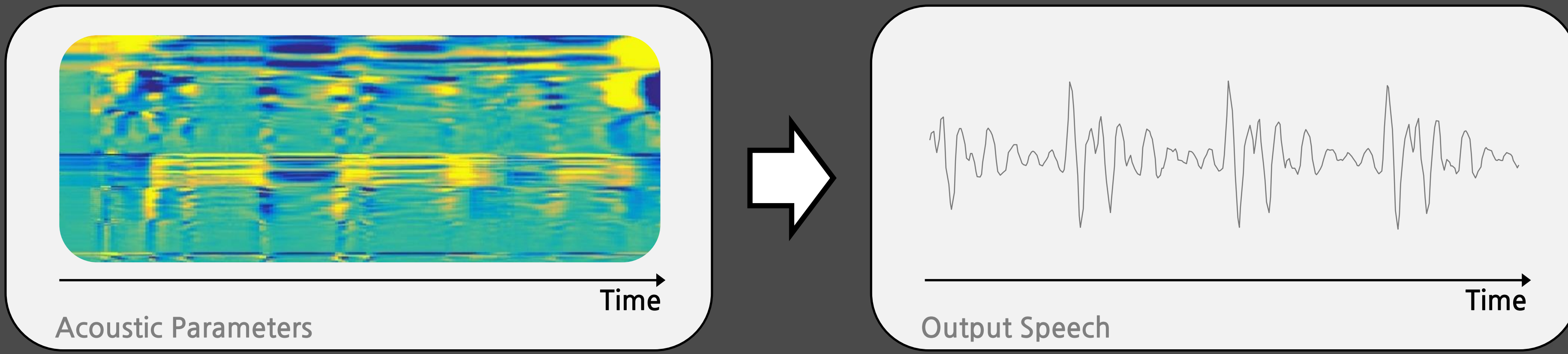


Summary

Text-to-speech (TTS) 란 기계가 사람처럼 **텍스트를 읽어주는** 기술입니다.

Acoustic Parameter 에서 음성 신호를 추정

Input Text → Acoustic Model → Acoustic Parameters → Vocoder → Output Speech



Q / A