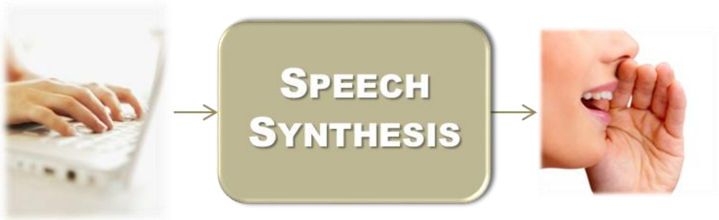
Speech synthesis

- Improved time-frequency trajectory excitation

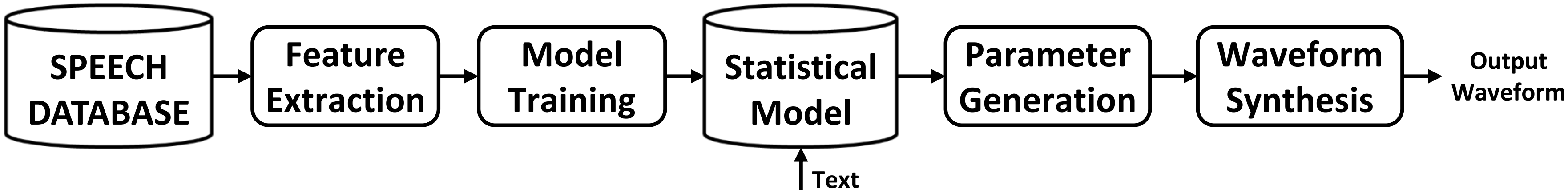
2015. 12. 10

EUNWOO SONG (宋銀宇)

Speech Synthesis



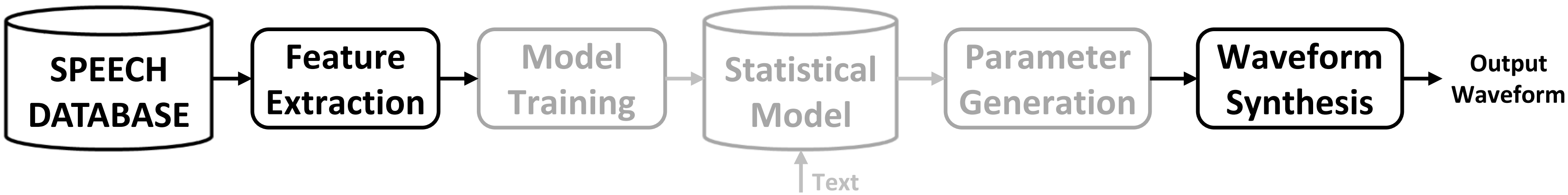
Major issues on speech synthesis [Zen' 09]



- Limitations in vocoding
    - How to design excitation & spectral parameters ?
- Inaccuracies of acoustic models
    - How to model the acoustic parameters ?
    - How to estimate the acoustic model parameters accurately ?
- Over-smoothed outputs
    - How to lively generate speech parameters ?

Today's talk

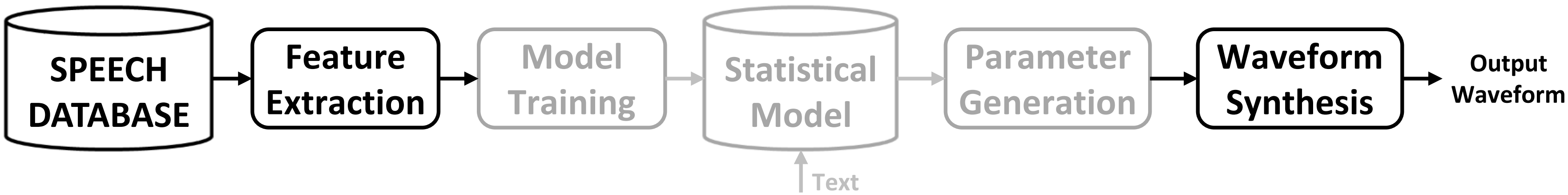
Limitations in vocoding



- Improved time-frequency trajectory excitation
    - E. Song, Y.S. Joo, and H.G. Kang, “Improved time-frequency trajectory excitation modeling for a statistical parametric speech synthesis system,” in *proc. of ICASSP*, 2015.

TODAY'S TALK

Limitations in vocoding



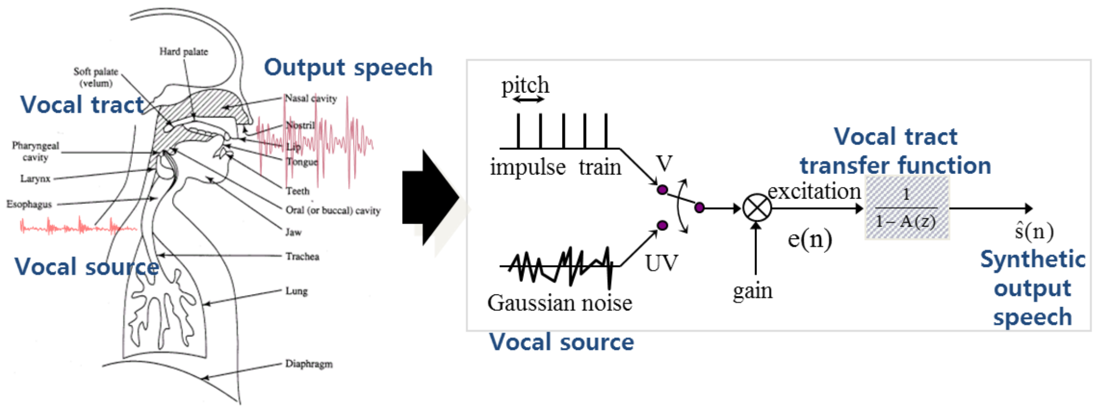
- Improved time-frequency trajectory excitation
  - E. Song, Y.S. Joo, and H.G. Kang, "Improved time-frequency trajectory excitation modeling for a statistical parametric speech synthesis system," in *proc. of ICASSP*, 2015.
- And its application to DNN-based speech synthesis
  - E. Song and H.G. Kang, "Deep neural network-based statistical parametric speech synthesis system using improved time frequency trajectory excitation modeling," in *proc. of INTERSPEECH*, 2015.

Speech synthesis

- Conventional vocoding techniques

VOCODING TECHNIQUES (1/5)

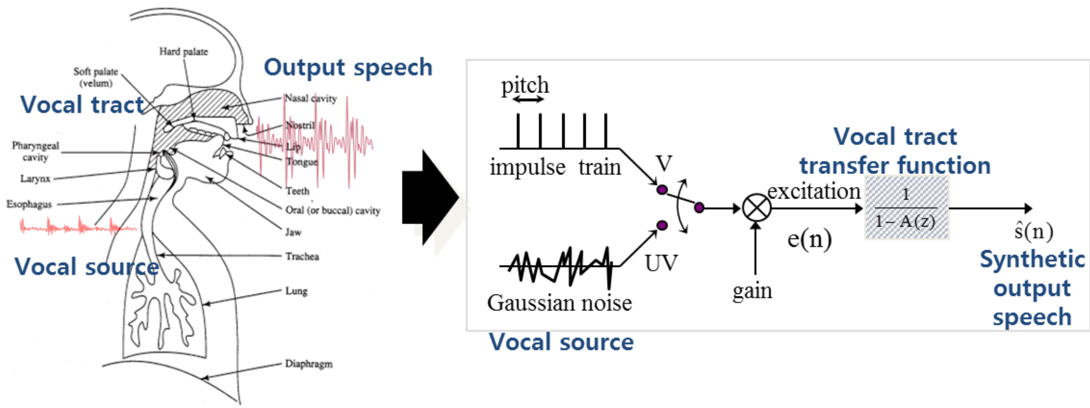
Pulse-or-noise (PoN) based on speech production model [Atal' 82]



- Spectral part (vocal tract-related)
  - Spectral parameter : linear prediction coefficient (LPC), cepstral coefficient
- Excitation part (vocal source-related)
  - Voiced frame : periodic pulse
  - Unvoiced frame : Gaussian noise
  - Fundamental frequency (F0)

VOCODING TECHNIQUES (1/5)

Pulse-or-noise (PoN) based on speech production model [Atal’ 82]



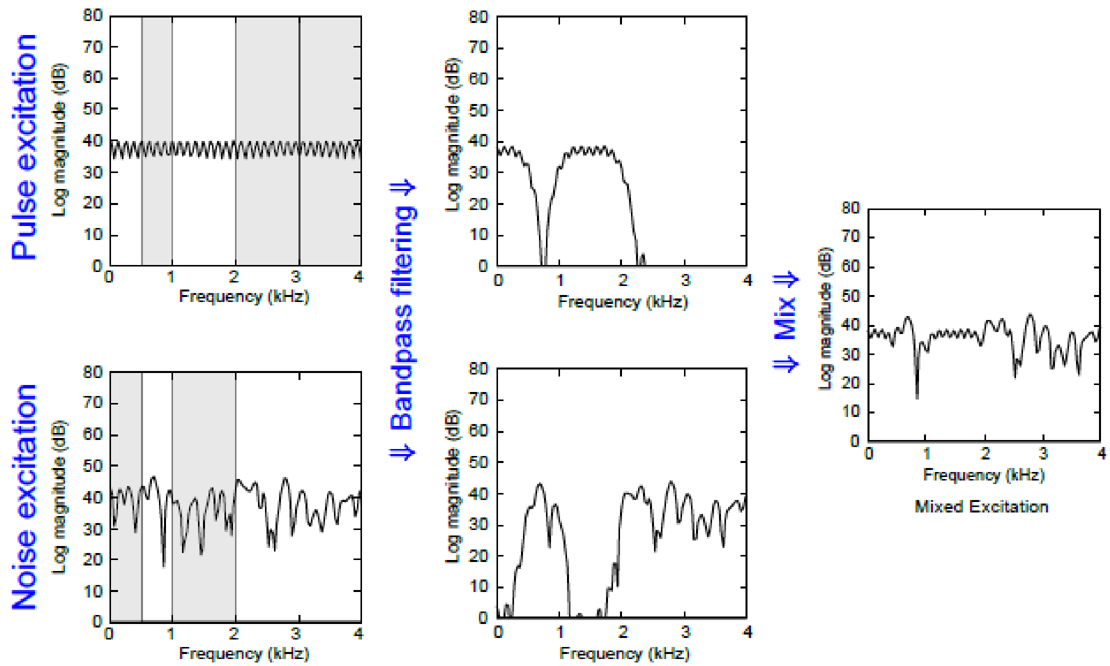
- Spectral part (vocal tract-related)
  - Spectral parameter : linear prediction coefficient (LPC), cepstral coefficient
- Excitation part (vocal source-related)
  - Voiced frame : periodic pulse
  - Unvoiced frame : Gaussian noise
  - Fundamental frequency (F0)

→ Mechanical sound

VOCODING TECHNIQUES (2/5)

Mixed excitation linear prediction (MELP) [McCree' 95]

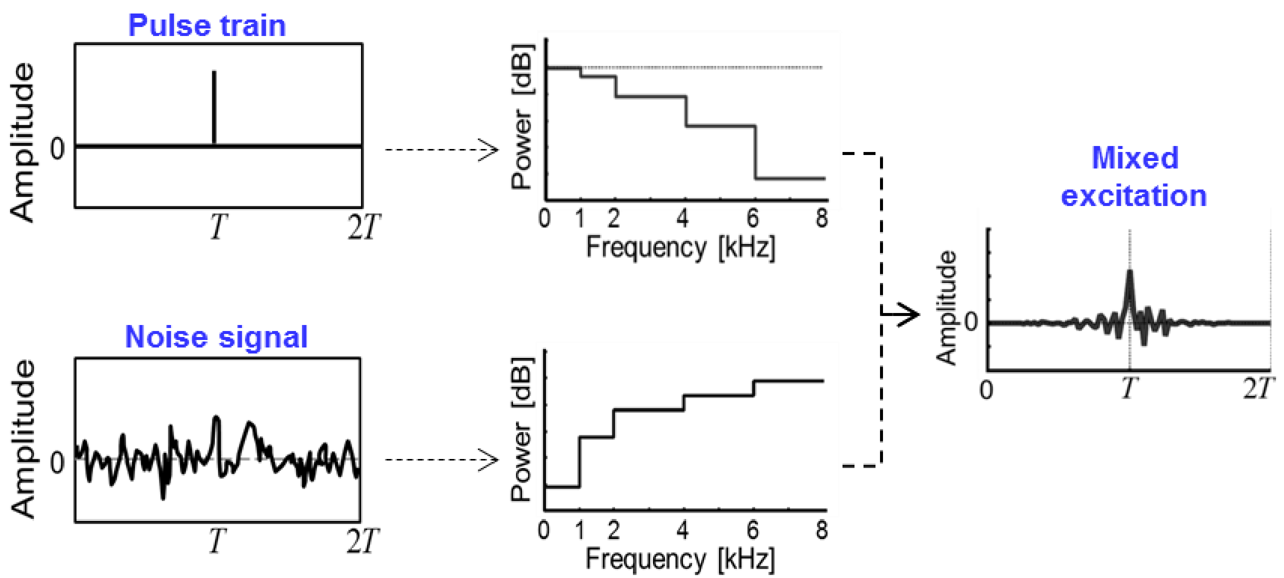
- Excitation signal is divided into fixed number of frequency bands
- Each frequency ***band*** is modeled by ***either pulse or noise***



Vocoding techniques (3/5)

STRAIGHT [Kawahara' 97]

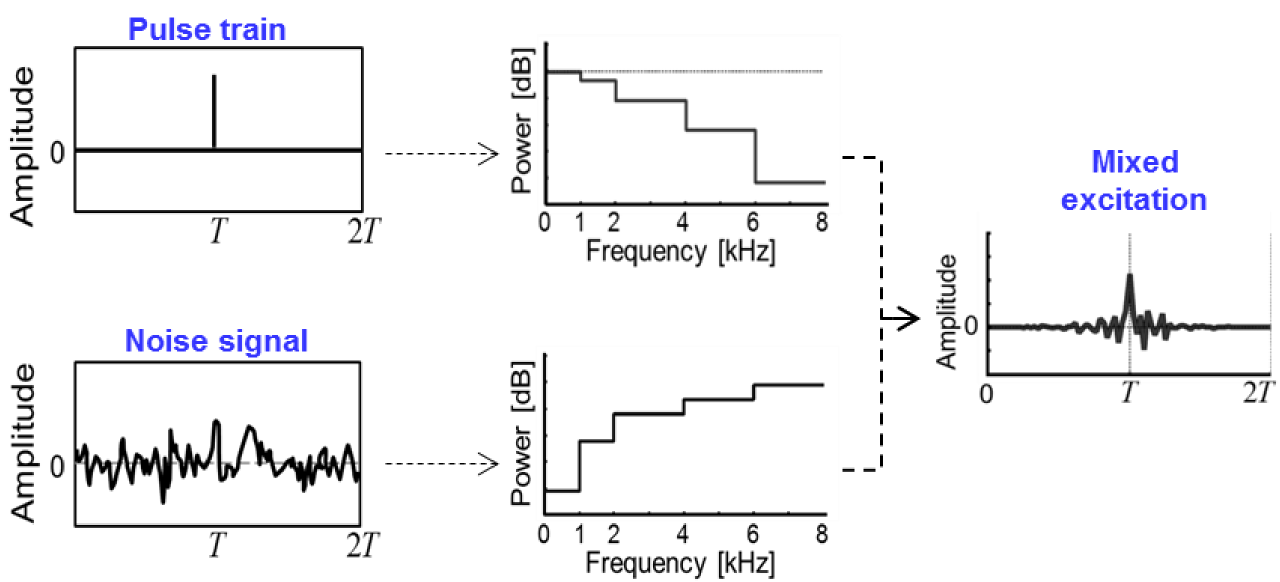
- Excitation signal is divided into fixed number of frequency bands
- Each frequency band is modeled by weight (***band aperiodicity; BAP***)



VOCODING TECHNIQUES (3/5)

STRAIGHT [Kawahara' 97]

- Excitation signal is divided into fixed number of frequency bands
- Each frequency band is modeled by weight (***band aperiodicity; BAP***)



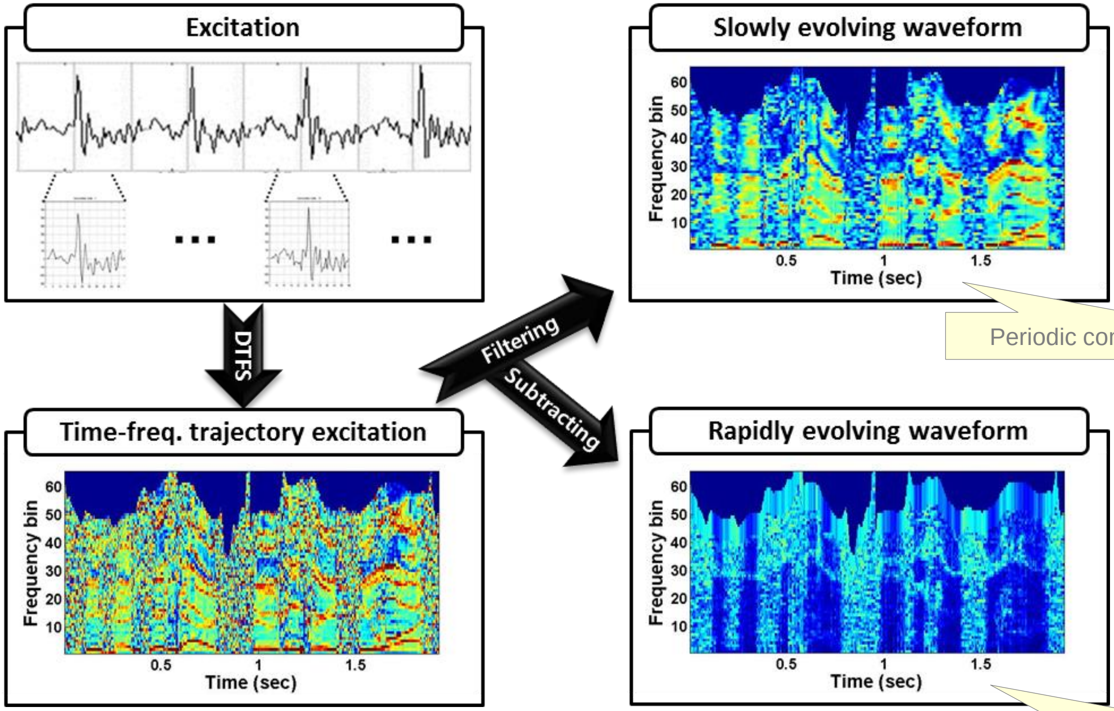
Fixing the boundary *of each frequency band* ***cannot fully represent*** *the* ***time-varying*** *characteristics of various types of phonetic information*

VOCODING TECHNIQUES (4/5)

Waveform interpolation (WI)

Time-frequency trajectory excitation (TFTE) [Choy' 98]

- Pitch-dependent **excitation** signal is transformed into discrete Fourier transform (DFT) domain
- Each frequency ***bin*** is modeled by ***slowly evolving waveform (SEW)*** and ***rapidly evolving waveform (REW)***



Periodic component of excitation

Remaining noisy component

VOCODING TECHNIQUES (5/5)

Examples of vocoded speech

PoN: pulse or noise

TFTE: time-frequency trajectory excitation

Original	PoN	STRAIGHT	TFTE

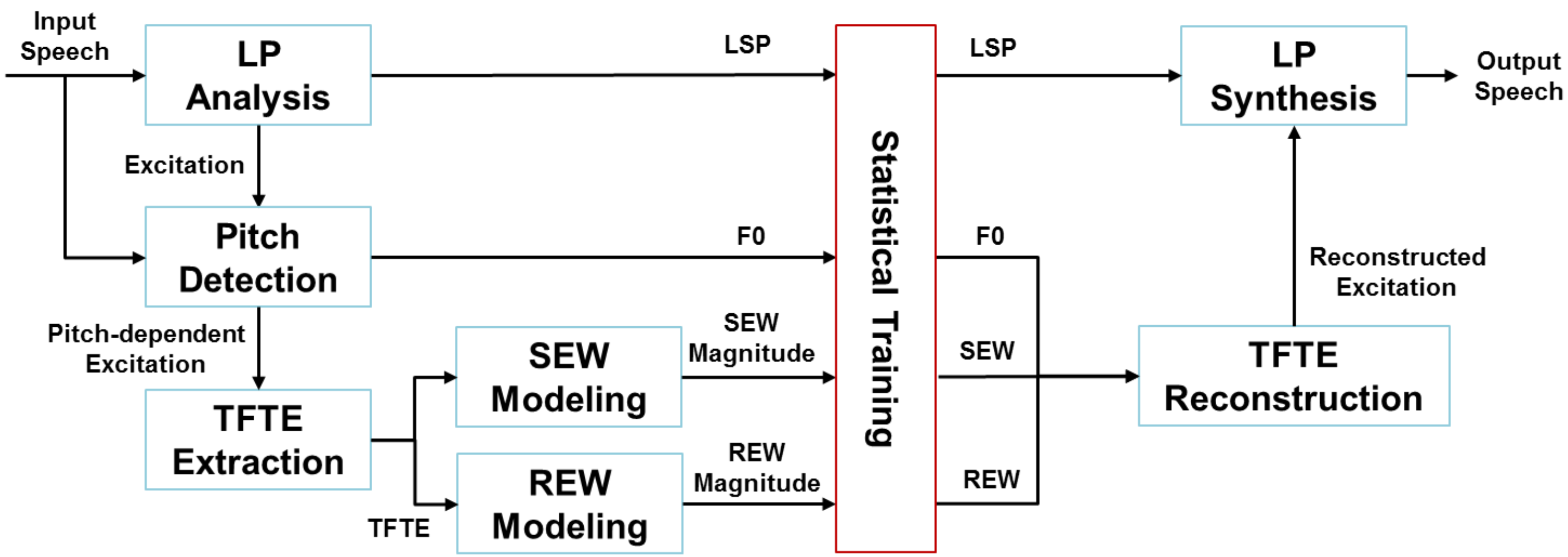
Speech synthesis

- Improved time-frequency trajectory excitation

E. Song, Y.S. Joo, and H.G. Kang, "Improved time-frequency trajectory excitation modeling for a statistical parametric speech synthesis system," in *proc. of ICASSP*, 2015.

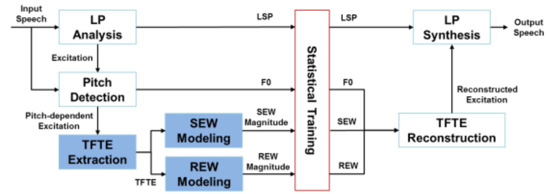
HIGH QUALITY VOCODER: TFTE (1/3)

Framework of TFTE-based speech synthesis



[Analysis]

[Synthesis]



HIGH QUALITY VOCODER: TFTE (2/3)

Time-frequency trajectory excitation (TFTE) [Choy' 98]

- TFTE has a length of one pitch period

$$u(n,\phi) = \sum_{k=1}^{P(n)/2} \left[A_k(n)\cos(k\phi) + B_k(n)\sin(k\phi) \right]$$

n : *frame index*
ϕ : *freq. index*
$P(n)$: *pitch period*
$A_k(n),\ B_k(n)$: *DFT coef.*

Decomposition of TFTE

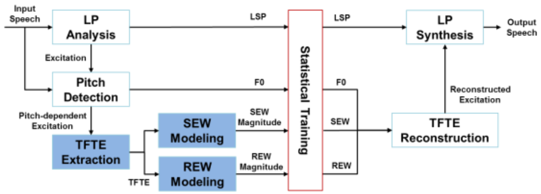
- SEW: periodic components of excitation

$$u_{SEW}(n,\phi) = \sum_{m=1}^{M} h(m)u(n-m,\phi)$$

$h(l)$: *low-pass filter in the time axis*

- REW: aperiodic components of excitation

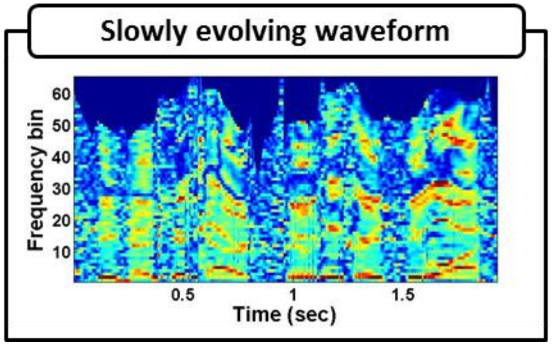
$$u_{REW}(n,\phi) = u(n,\phi) - u_{SEW}(n,\phi)$$

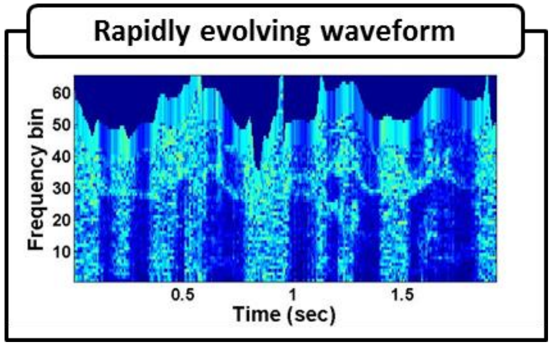


HIGH QUALITY VOCODER: TFTE (3/3)

Advantage of TFTE

- Efficiency of extracting time-varying periodicity in a unit of individual frequency bin



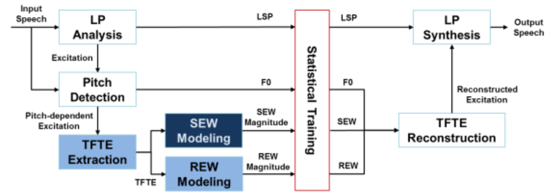


Limitation of TFTE

- Difficulties in modeling of TFTE parameters due to pitch-dependent dimension of TFTE

Parameterization method of TFTE: Improved TFTE (ITFTE)

MODELING OF TFTE: ITFTE (1/5)

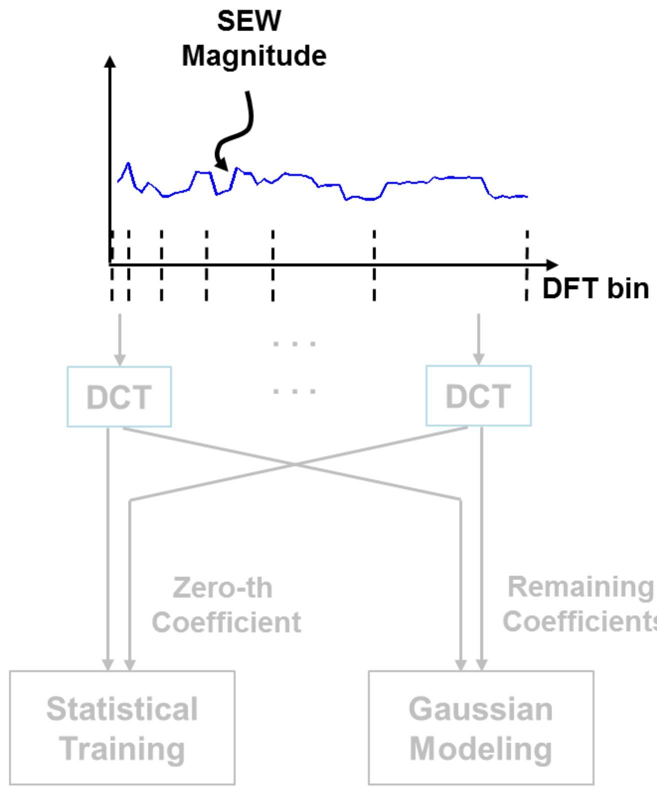


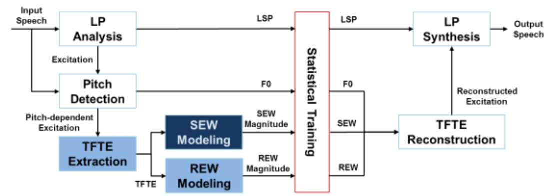
Modeling of SEW [Song' 15-1]

- SEW magnitude is first divided into K number of frequency sub-block

$$\begin{bmatrix} c_{k,1} \\ \vdots \\ c_{k,J_k} \end{bmatrix}^T = \begin{bmatrix} u_{SEW}(n, J_{k-1}+1) \\ \vdots \\ u_{SEW}(n, J_{k-1}+J_k) \end{bmatrix}^T, \quad \sum_{k=1}^{K} J_k = P(n)/2$$

$$\boxed{\begin{array}{l} C_{k,j} : j^{th}\ SEW\ magnitude\ of \\ \qquad k^{th}\ sub-block \\ J_k : length\ of\ k^{th}\ sub-block \end{array}}$$



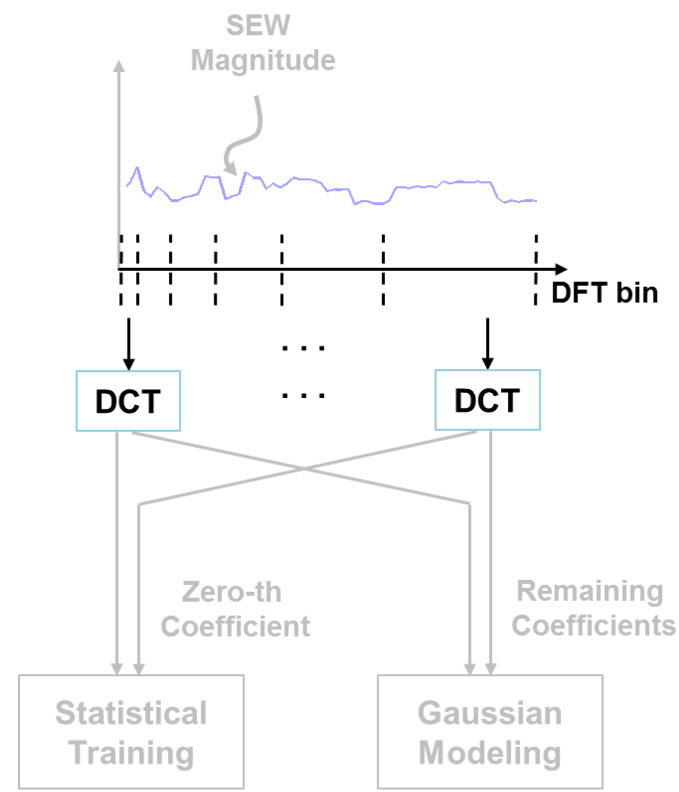


Modeling of TFTE: ITFTE (2/5)

Modeling of SEW [Song' 15-1]

- Then, each sub-block is transformed with discrete cosine transform (DCT)

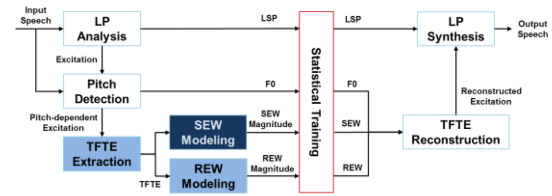
$$C_{k,m} = \frac{1}{J_k} \sum_{j=1}^{J_k} c_{k,j} \cos\left(\frac{\pi}{J_k} (j - 0.5)(m - 1) \right)$$



YONSEI UNIVERSITY

Microsoft Research

Modeling of TFTE: ITFTE (2/5)

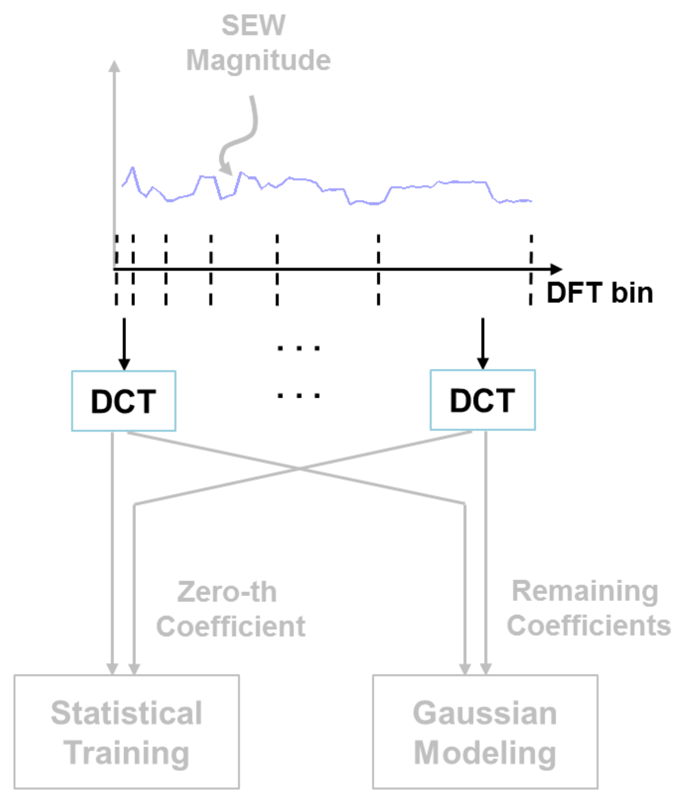


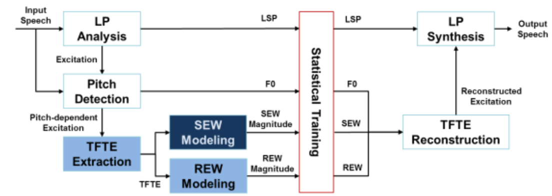
Modeling of SEW [Song' 15-1]

- Then, each sub-block is transformed with discrete cosine transform (DCT)

$$C_{k,m} = \frac{1}{J_k} \sum_{j=1}^{J_k} c_{k,j} \cos\left(\frac{\pi}{J_k} (j-0.5)(m-1) \right)$$

*Since the DCT is a good **decorrelator**, most energy of SEW magnitude is concentrated within the **first few** coefficients*





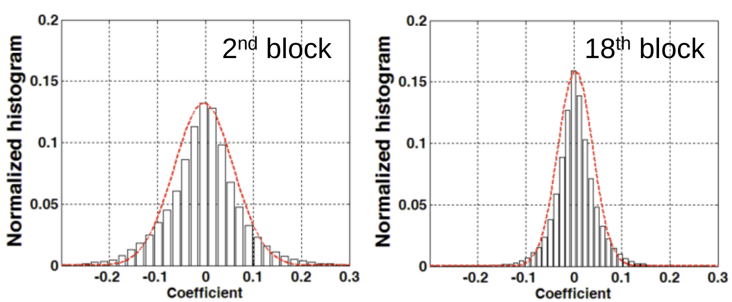
MODELING OF TFTE: ITFTE (3/5)

Modeling of SEW [Song' 15-1]

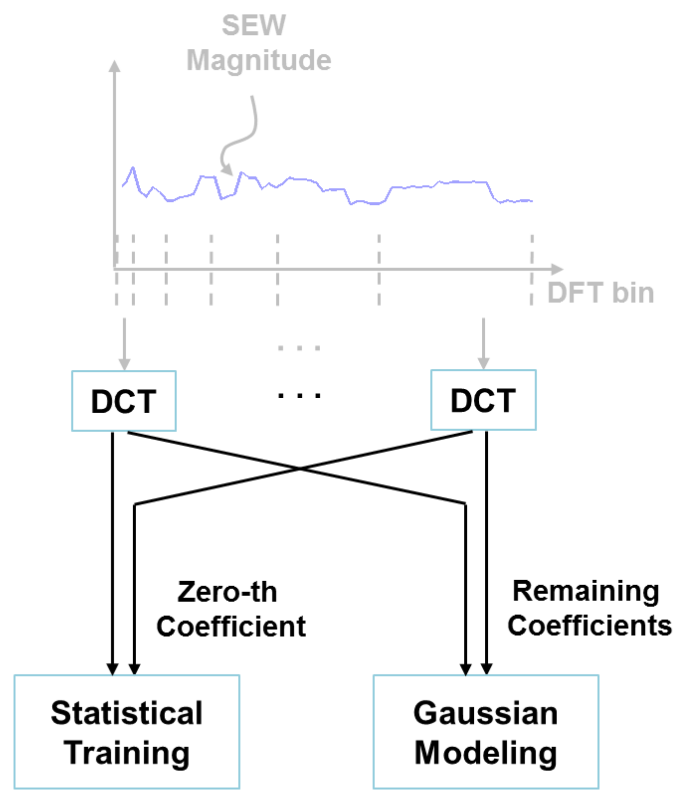
- 0-th coefficient of each sub-block is used for the HMM/DNN training

$$(\#\ of\ parameter) = (\#\ of\ sub\text{-}block)$$

- Remaining coefficients are stochastically generated by Gaussian random variables in the synthesis step



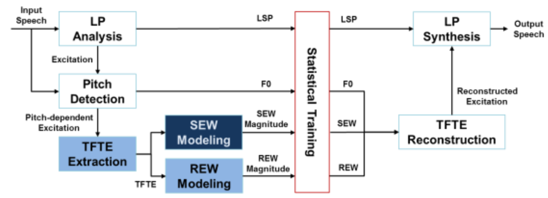
[Normalized histogram of remaining coefficients]



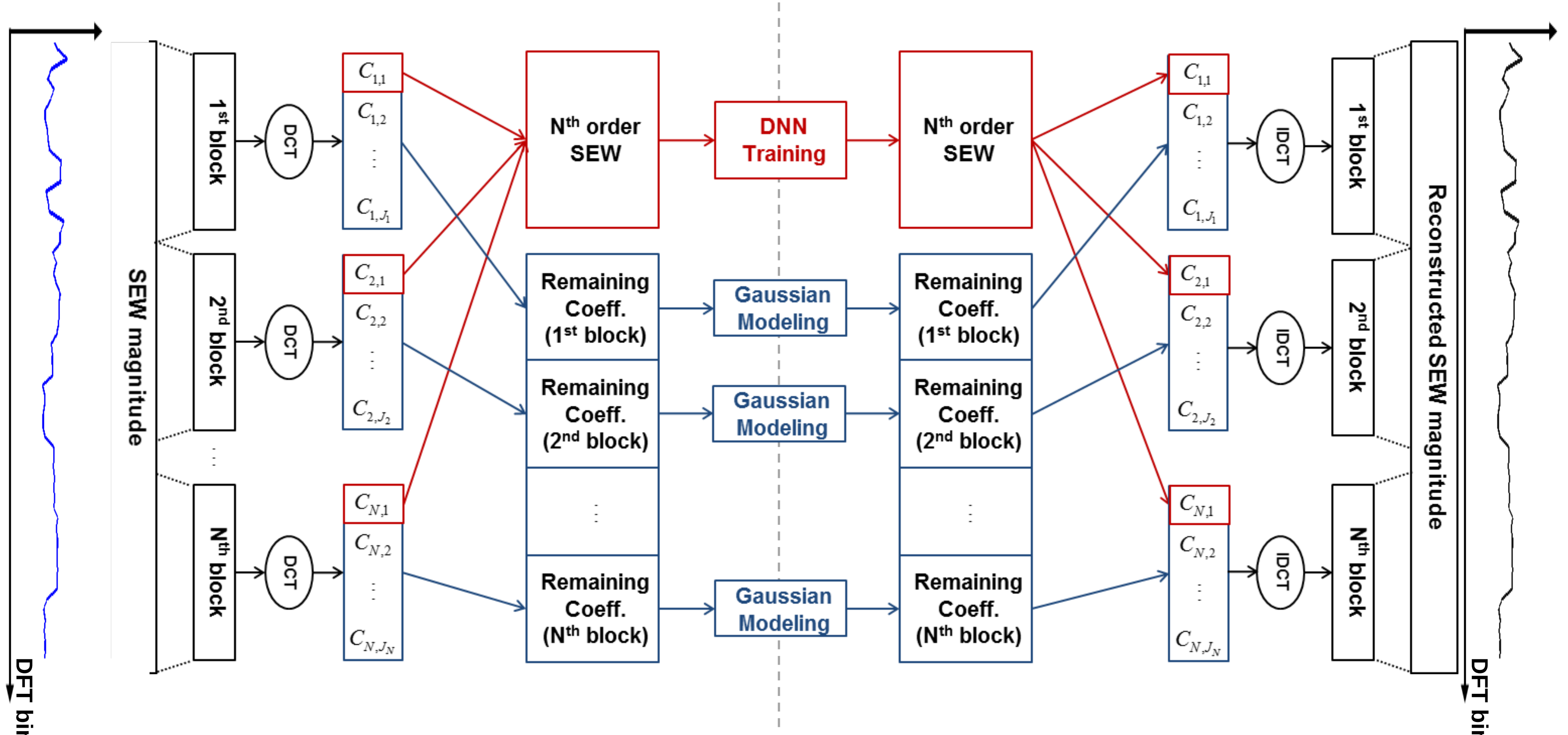
YONSEI UNIVERSITY

Microsoft Research

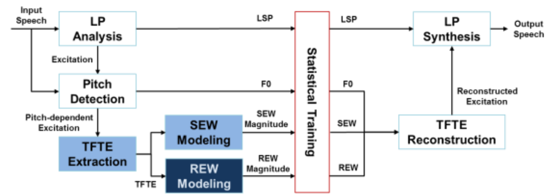
Modeling of TFTE: ITFTE (4/5)



Modeling of SEW [Song' 15-1]



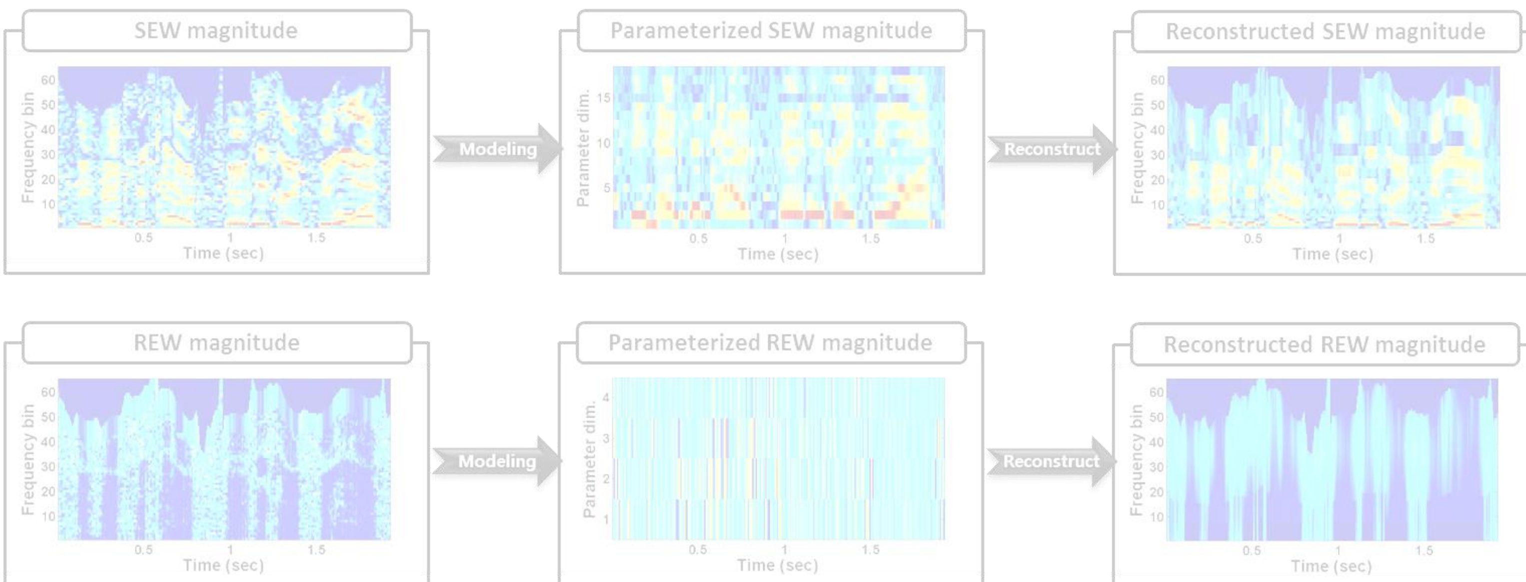
[Modeling of SEW]



Modeling of TFTE: ITFTE (5/5)

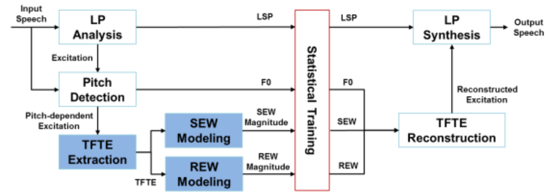
Modeling of REW [Song' 15-1]

- REW magnitude is modeled by power contour estimation method
- Typically, Legendre orthonormal polynomial coefficients are used for the modeling



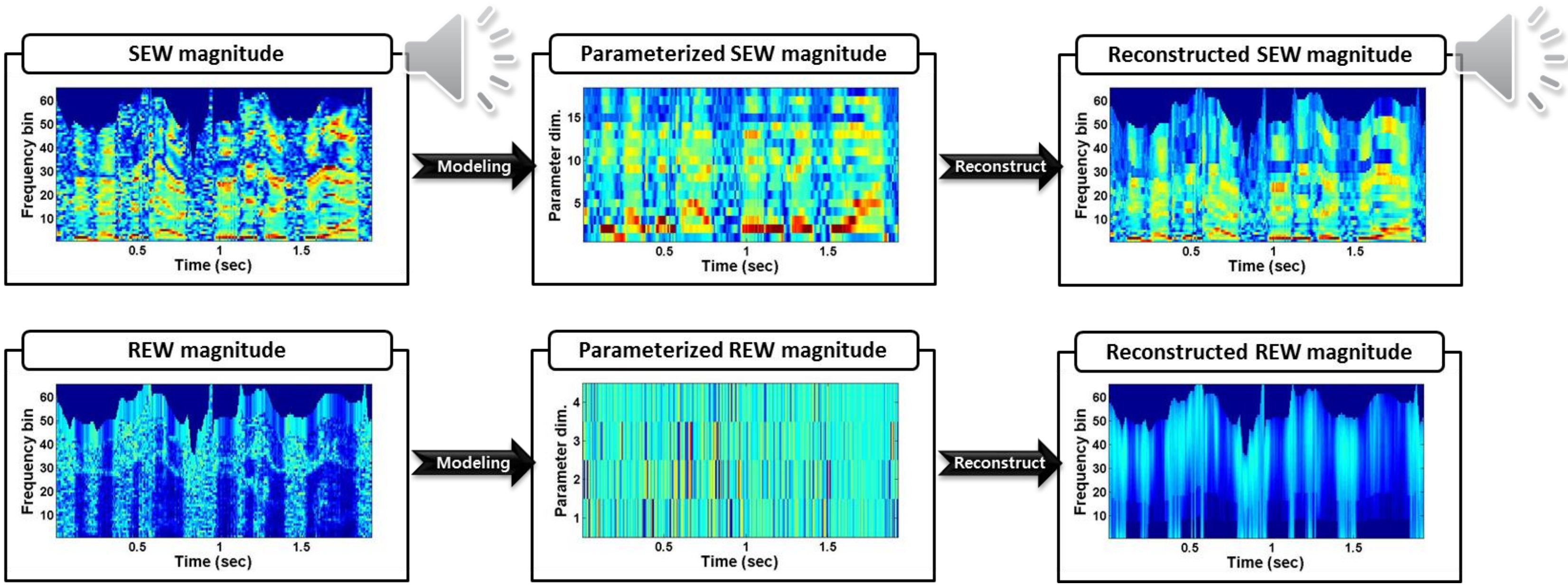
***Full frequency band** information of the SEW and REW*
*can be reconstructed by **fixed number of model coefficients***

MODELING OF TFTE: ITFTE (5/5)



Modeling of REW [Song' 15-1]

- REW magnitude is modeled by power contour estimation method
- Typically, Legendre orthonormal polynomial coefficients are used for the modeling



***Full frequency band** information of the SEW and REW*
*can be reconstructed by **fixed number of model coefficients***

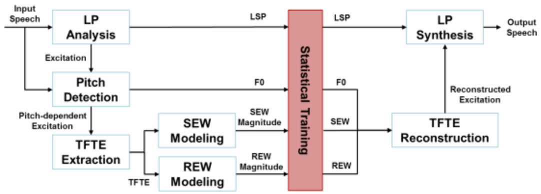
YONSEI UNIVERSITY

Microsoft Research

Speech synthesis

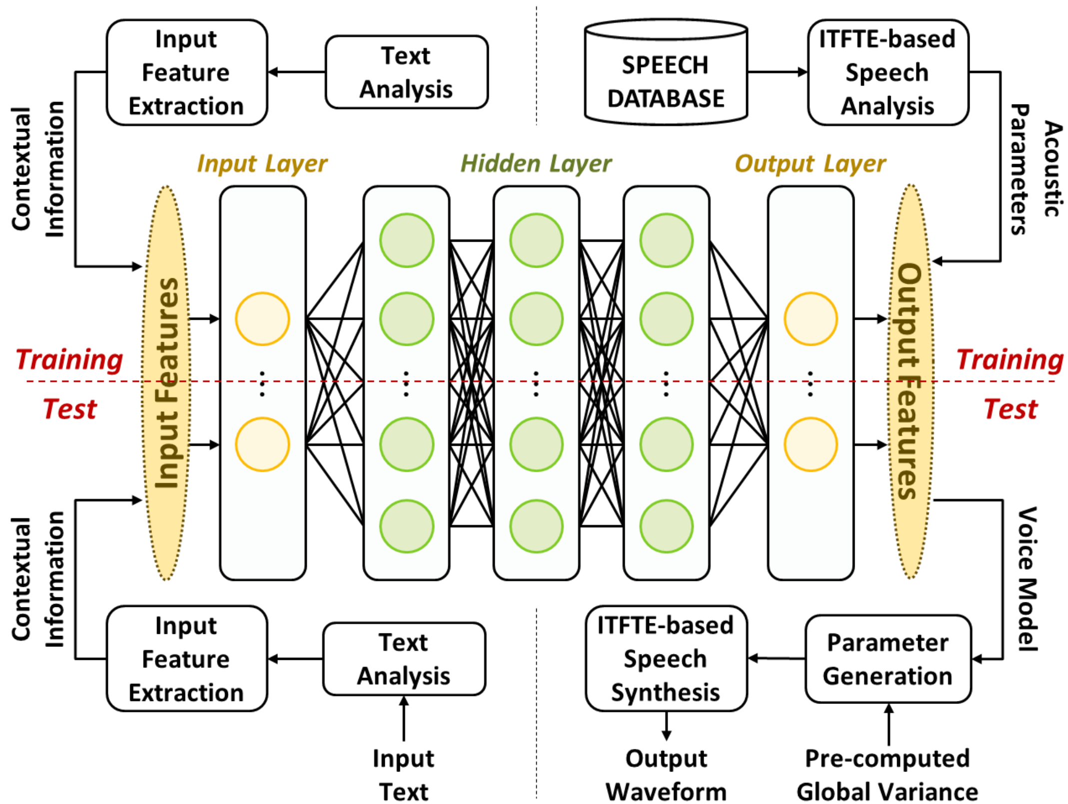
- ITFTE modeling for DNN-based speech synthesis

E. Song and H.G. Kang, “Deep neural network-based statistical parametric speech synthesis system using improved time frequency trajectory excitation modeling,” in proc. of INTERSPEECH, 2015.



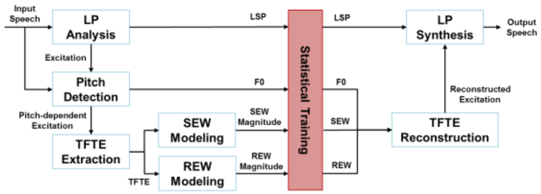
DNN-BASED SPEECH SYNTHESIS (1/8)

Modeling of non-linear mapping function between contextual information and acoustic parameters [Zen' 13]



YONSEI UNIVERSITY

Microsoft Research



DNN-BASED SPEECH SYNTHESIS (2/8)

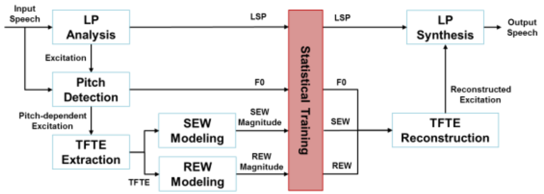
Analysis on trainability of ITFTE model [Song' 15-2]

- Trainability is measured by normalized root mean square error (NMSE) between original and generated ITFTE parameters

$$NMSE = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{\sum_{k=1}^{K}\left(x_{ori}(n,k)-x_{gen}(n,k)\right)^2}{\sum_{k=1}^{K}\left(x_{ori}(n,k)\right)^2}}$$

$$\boxed{\begin{array}{l} N\ :\ number\ of\ frame \\ K\ :\ dimension\ of\ parameter \end{array}}$$

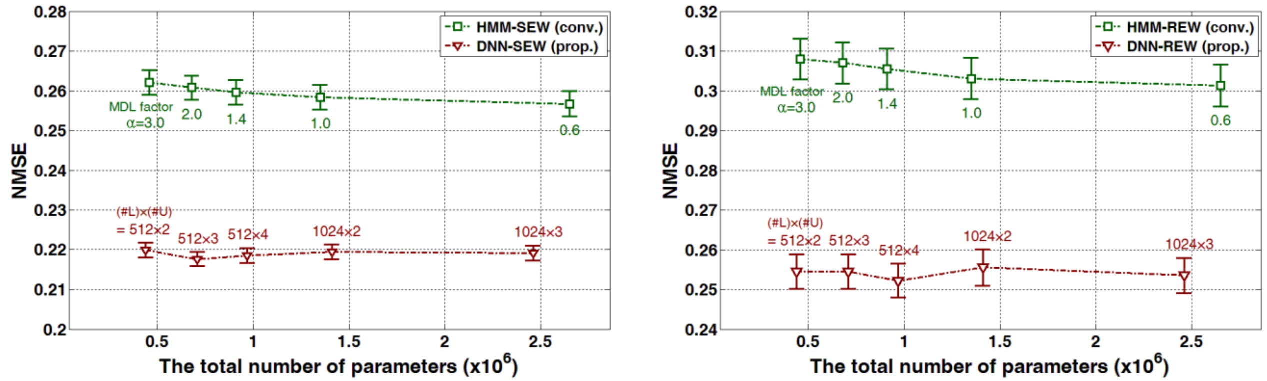
- Model size is controlled by
  - Conventional HMM-based system
    - Scale factor of the minimum description length (MDL) criteria [Shinoda' 00]
  - Proposed DNN-based system
    - Number of layers (#L) and number of units (#U)



DNN-BASED SPEECH SYNTHESIS (3/8)

Analysis on trainability of ITFTE model [Song' 15-2]

- Average NMSE with 95% confidence interval (CI)



- NMSE of HMM-based system is larger than that of DNN-based one
- 95% CI of HMM-based system is wider than that of DNN-based one

It implies that excitation signal contains many frames with large errors, which would be expected to degrade naturalness of synthesized speech

DNN-BASED SPEECH SYNTHESIS (4/8)

Experiment setup [Song' 15-2]

Database	Korean male speaker
Training/ validation/ test	2700(3.5 hour)/ 100/ 100 utterances
Sampling rate	16 kHz
Analysis window	20ms width, 5ms shift
Linguistic feature	8 categorical features + 7 numerical features
Acoustic feature	(see table below)

Acoustic feature:

Feature	dimension
Line spectral pairs	$24 + \Delta + \Delta\Delta$
SEW magnitude	$18 + \Delta + \Delta\Delta$
REW magnitude	$4 + \Delta + \Delta\Delta$
log-F0	$1 + \Delta + \Delta\Delta$
energy	$1 + \Delta + \Delta\Delta$

DNN-BASED SPEECH SYNTHESIS (5/8)

Experiment setup [Song' 15-2]

HMM topology	5-state, left-to-right HMM Model size is controlled by MDL factor	
DNN Architecture	Layer dimension	
	Input	203-dim. binary features 7-dim numerical features
	Output	144-dim. ITFTE parameters
	Hidden	512x512, 512x512x512 1024x1024, 1024x1024x1024
	Activation/ output function	
	sigmoid	
	Normalization	
	Input	Zero-mean, unity-variance
	Output	min.-max. (0.01 to 0.99)

DNN-BASED SPEECH SYNTHESIS (6/8)

Objective test results [Song' 15-2]

- Test results for different MDL factors (α) of HMM-ITFTE system

HMM-ITFTE	LSD (dB)	F0 RMSE (Hz)	SEW NMSE	REW NMSE
α=3.0 (0.44M)	3.227	16.788	0.262	0.308
α=2.0 (0.68M)	3.188	16.732	0.261	0.307
α=1.0 (1.35M)	3.129	16.401	0.258	0.303
α=0.6 (2.65M)	3.098	16.759	0.257	0.301

- Test results for different architectures of DNN-ITFTE system

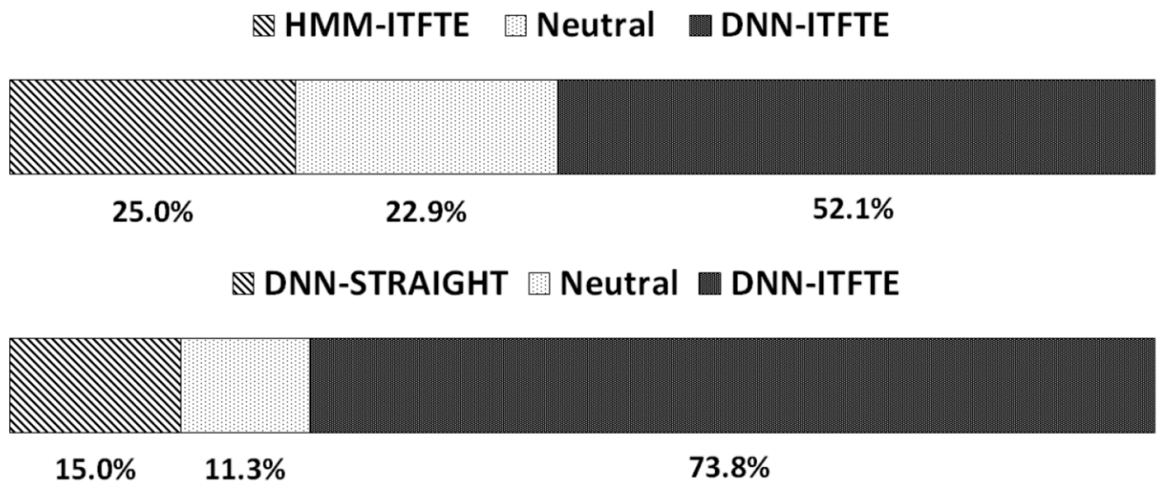
DNN-ITFTE	LSD (dB)	F0 RMSE (Hz)	SEW NMSE	REW NMSE
512$\times$2 (0.46M)	3.240	14.748	0.220	0.255
512$\times$3 (0.71M)	3.192	13.218	0.218	0.254
1024$\times$2 (1.41M)	3.207	14.477	0.219	0.256
1024$\times$3 (2.46M)	3.189	15.766	0.219	0.254

ITFTE parameters generated by the DNN-based system contain smaller estimation errors than those generated by the HMM-based system

DNN-BASED SPEECH SYNTHESIS (7/8)

Subjective test results (A/B preference test) [Song' 15-2]

- 20 utterances are randomly selected
- 12 listeners are asked to provide quality judgment



DNN-ITFTE system provides much higher perceptual quality than that of DNN-STRAIGHT and HMM-ITFTE

DNN-BASED SPEECH SYNTHESIS (8/8)

Examples of synthesized speech

Original				
HMM-ITFTE				
DNN-ITFTE				
DNN-STR.				

SUMMARY

Major issues on speech synthesis

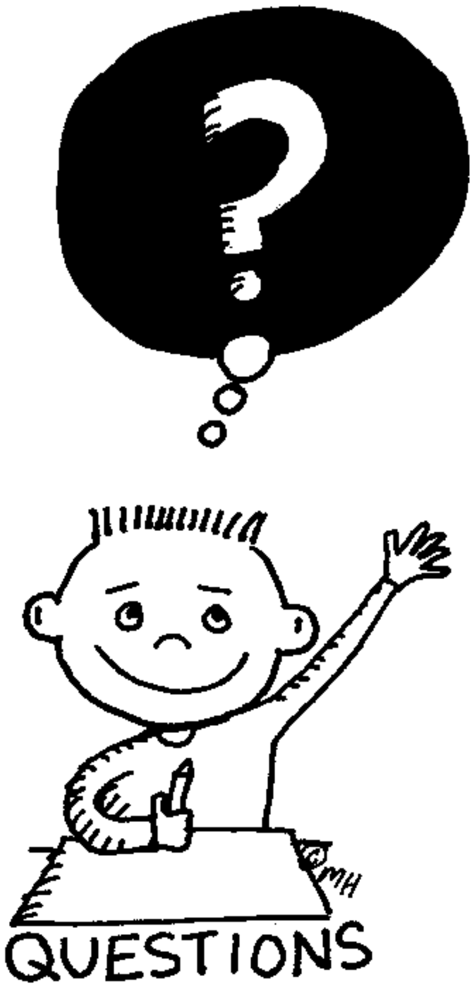
- Limitations in vocoding
- Inaccuracies of acoustic models
- Over-smoothed outputs

Improved time-frequency trajectory excitation (ITFTE)

- Parameterization method of TFTE vocoder for the HMM/DNN training

DNN-based speech synthesis using ITFTE method

- Improvement of modeling accuracy



E-mail: sewplay@dsp.yonsei.ac.kr
Homepage: http://dsp.yonsei.ac.kr/tts_woo

REFERENCE

[Atal' 82] B. S. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. Comm.*, vol. 30, no. 4, pp. 600-614, 1982.

[Choy' 98] E. L. Choy, "Waveform interpolation speech coder at 4kbps," Ph.D. dissertation, McGill University, 1998.

[Kawahara' 97] K. Kawahara, et al., "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," *in Proc. of ICASSP*, 1997.

[McCree' 95] A. McCree, and T. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech and Audio Process.*, vol. 3, no.4, 1995.

[Shinoda' 00] K. Shinoda, et al., "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn(E)*, vol.21, no. 2, pp. 79-86, 2000.

[Song' 15-1] E. Song, et al., "Improved time-frequency trajectory excitation modeling for statistical parametric speech synthesis system," *in Proc. ICASSP*, 2015.

[Song' 15-2] E. Song, and H.G. Kang," Deep neural network-based statistical parametric speech synthesis system using improved time-frequency trajectory excitation model," *in Proc. Interspeech*, 2015.

[Zen' 09] H. Zen, et al., "Statistical parametric speech synthesis," *Speech Comm.,* vol. 51, no. 11, pp. 1039-1064, 2009.

[Zen' 13] H. Zen, et al., "Statistical parametric speech synthesis using deep neural networks," *in Proc. ICASSP*, 2013.

Thank you