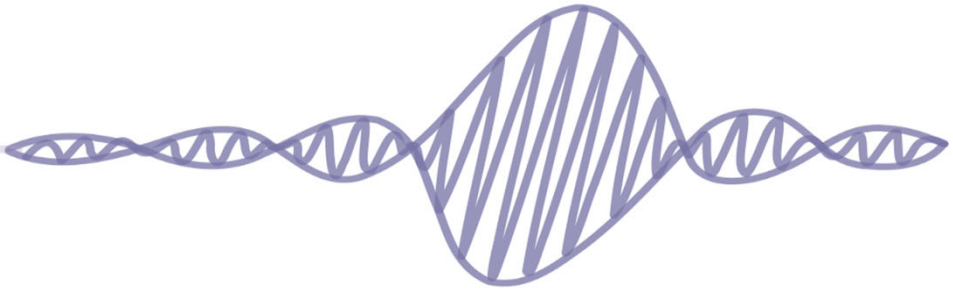
APSIPA
Asia-Pacific Signal and Information Processing Association



Speech Waveform Modeling for Advanced Voice Conversion

APSIPA Distinguished Lecture

Tomoki TODA

Nagoya University, JAPAN

Introduction to APSIPA

APSIPA Mission: To promote broad spectrum of research and education activities in signal and information processing in Asia Pacific

APSIPA Publications: Transactions on Signal and Information Processing in partnership with Cambridge Journals since 2012; APSIPA Newsletters

* Open-access e-only publications
  https://www.cambridge.org/sip

APSIPA Social Network: To link members together and to disseminate valuable information more effectively

* Friend labs
  http://www.apsipa.org/friendlab/Application/LabList.asp

Web page: http://www.apsipa.org/

APSIPA Conferences: ASPIPA Annual Summit and Conference (ASC)

12th APSIPA ASC 2020: Auckland, New Zealand, Dec. 7—10, 2020

July 1, 2020 Paper submissions

Sep. 1, 2020 Notification of acceptance

11th APSIPA ASC 2019: Lanzhou, China, Nov. 18—21, 2019

10th APSIPA ASC 2018: Honolulu, USA, Nov. 2018
9th APSIPA ASC 2017: Kuala Lumpur, Malaysia, Dec. 2017
8th APSIPA ASC 2016: Jeju, South Korea, Dec. 2016
7th APSIPA ASC 2015: Hong Kong, Dec. 2015
6th APSIPA ASC 2014: Siem Reap, Cambodia, Dec. 2014
5th APSIPA ASC 2013: Kaohsiung, Taiwan, Oct. 2013
4th APSIPA ASC 2012: Hollywood, USA, Dec. 2012
3rd APSIPA ASC 2011: Xi'an, China, Oct. 2011
2nd APSIPA ASC 2010: Biopolis, Singapore, Dec. 2010
1st APSIPA ASC 2009: Sapporo, Japan, Oct. 2009

APSIPA Distinguished Lecture 2019—2020

Speech Waveform Modeling for Advanced Voice Conversion

Outline

- Let’s review voice conversion (VC) progress!
  - Basics of VC
    - How to do VC?
    - For what?
  - Recent progress of VC
    - Which VC techniques are really helpful?
    - Let’s review recent Voice Conversion Challenge!
- Let’s review recent progress of waveform modeling!
  - Basics of waveform modeling
    - Let’s revisit vocoder!
  - Progress of waveform modeling in VC
    - How to avoid using vocoder?
    - How to improve vocoder?

Basis of VC

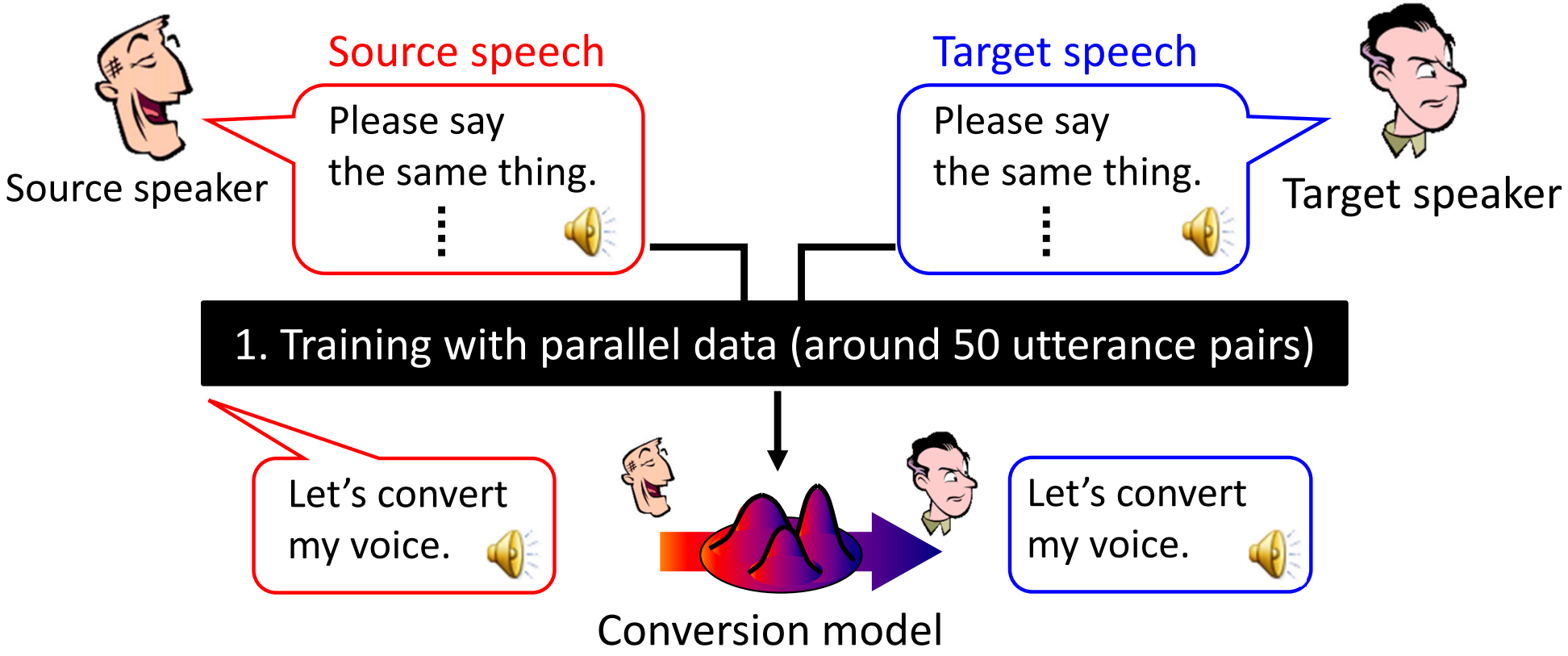
- Typical VC framework
- VC applications

Basic Framework of Statistical VC

[Abe; '90]

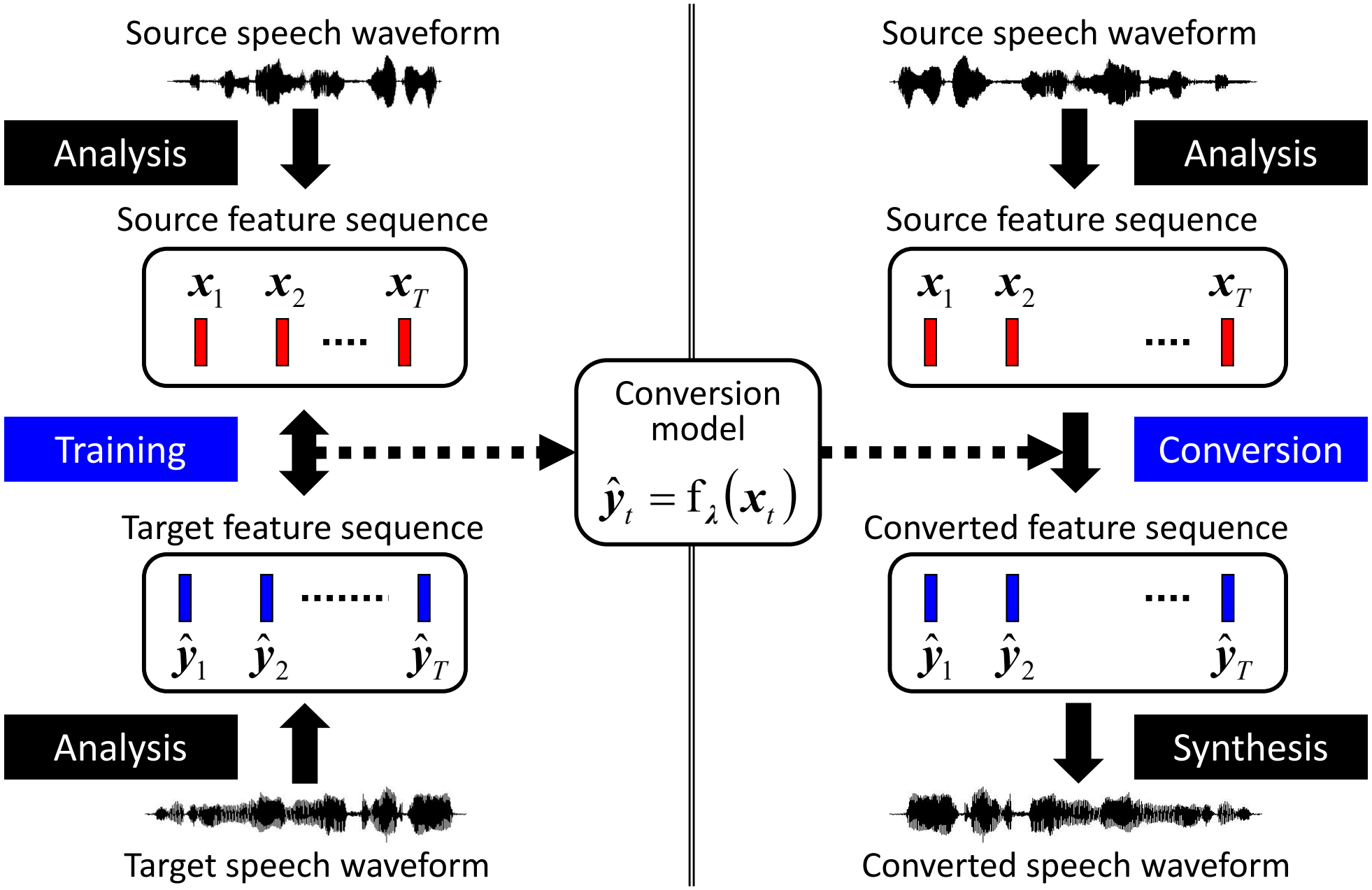
- Described as a regression problem
- Supervised training using utterance pairs of source & target speech

Example: speaker conversion



2. Conversion of any utterance while keeping linguistic contents unchanged

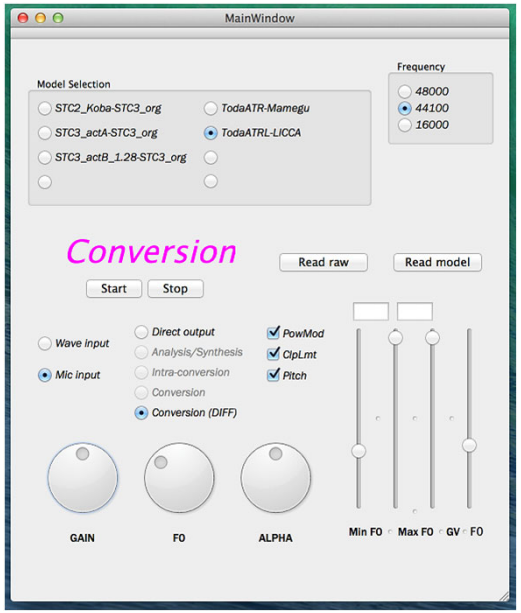
Training and Conversion Steps



Demo: Character Voice Changer

[Toda; ’12][Kobayashi; ’18a]

- Convert my voice into specific characters’ voices



Realtime statistical VC software

[Dr. Kobayashi, Nagoya Univ.]

An Example of VC Application

[Toda; '14]

- Development of augmented speech production

Voice changer or vocal effector to produce a desired voice

Create new expressions!

From current singing voice → to younger voice / to elder voice

Speaking aid to recover a lost voice

Break down barriers!

From vocal disorder's voice → to a naturally sounding voice

Silent speech interface to talk with cellphone while keeping silent!

Talk anytime and anywhere!

From very soft murmur → to intelligible voice

Risk of VC

- Need to look at a possibility that statistical VC is misused for spoofing…
  - Real-time VC makes it possible for someone to speak with your voices…
- Shall we stop VC research?

  ➡ No. There are many useful applications making our society better!
- What can we do?
  - Collaborate with anti-spoofing research [Wu; ’15]
    - ASVspoof (automatic speaker verification spoofing and countermeasures challenge) has been held since 2015. [Wu; ’17][Kinnunen; ’17]
  - Need to widely tell people how to use statistical VC correctly!

VC needs to be socially recognized as a kitchen knife.

Recent Progress of VC

- Evaluation of various techniques
- Important findings

Voice Conversion Challenges (VCCs)

- Conducted to better understand different VC techniques by comparing their performance using a freely-available dataset as a common dataset
- VCC2016 [Toda; '16] and VCC2018 [Lorenzo-Trueba; '18]
  - Tasks: speaker conversion
    - Parallel training (VCC2016 & VCC2018) and nonparallel training (VCC2018)
  - Perceptual evaluation: naturalness and speaker similarity by listening tests
  - Datasets: VCC 2016 and VCC2018 datasets designed using DAPS [Mysore, '15]

	VCC2018	# of speakers	# of sentences
Parallel training tsk	Source speakers	2 females & 2 males	81 for training & 35 for evaluation
	Target speakers	2 females & 2 males	81 for training
Nonparallel training task	Other source speakers	2 females & 2 males	Other 81 for training & 35 for evaluation

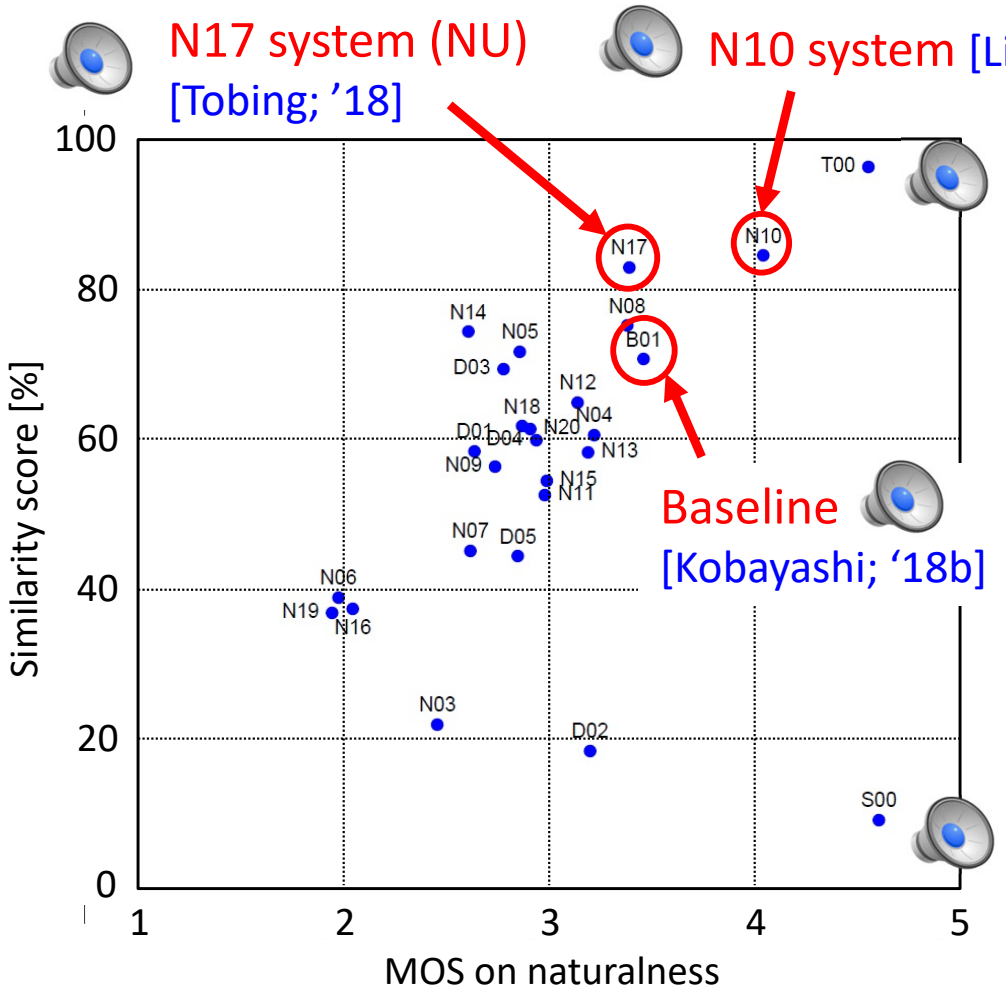
Overall Results of VCC2018 Listening Tests

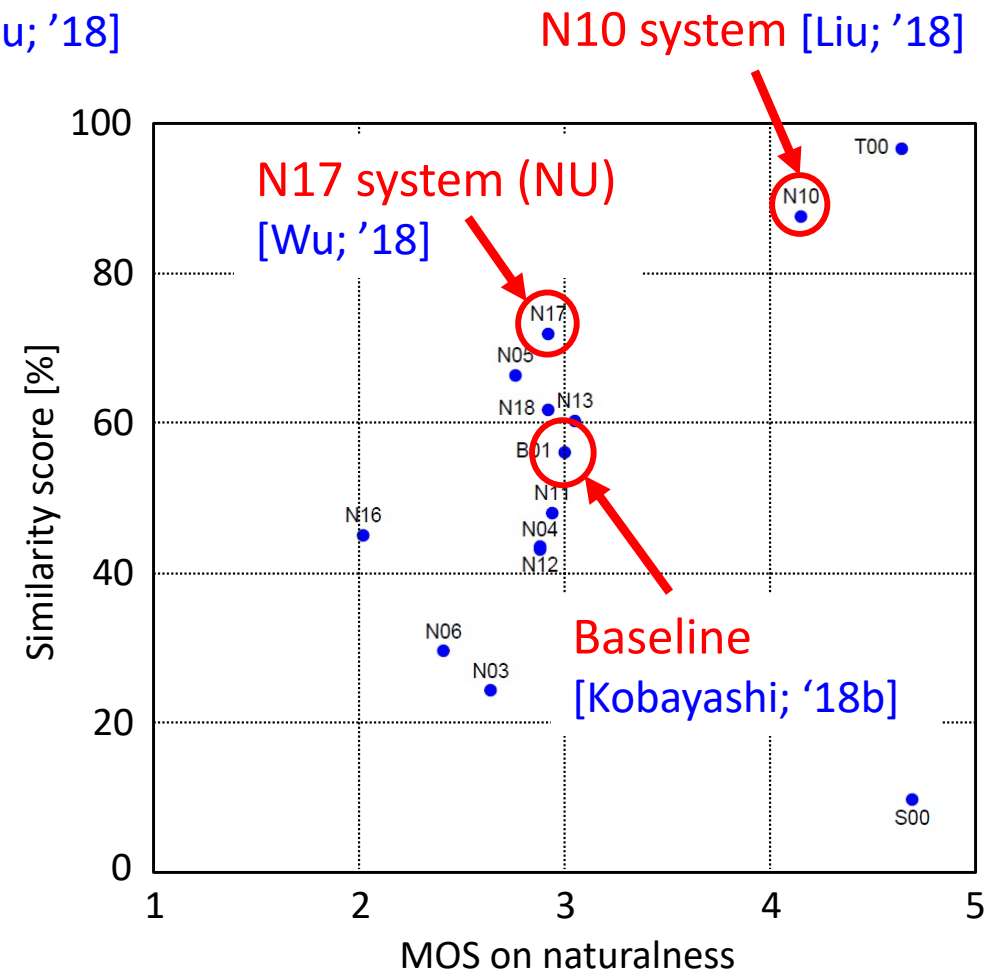
Parallel training task

- 23 submitted systems
- 1 baseline (developed w/ sprocket)

Nonparallel training task

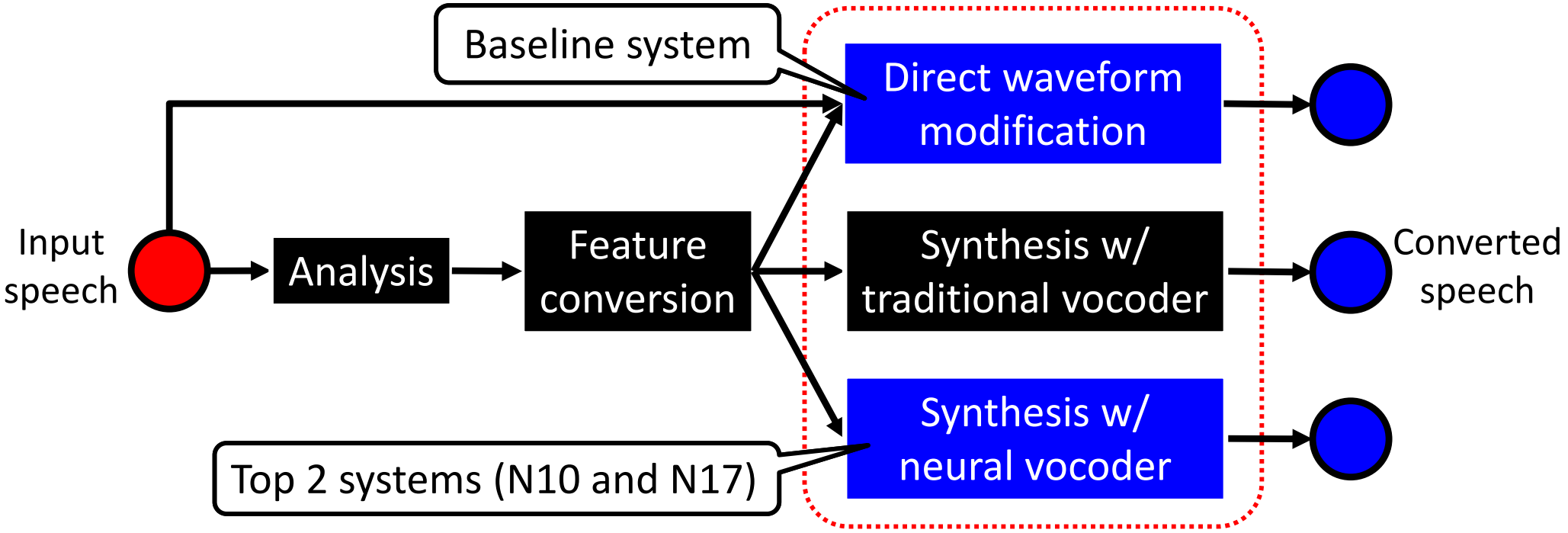
- 11 submitted systems
- 1 baseline (developed w/ sprocket)



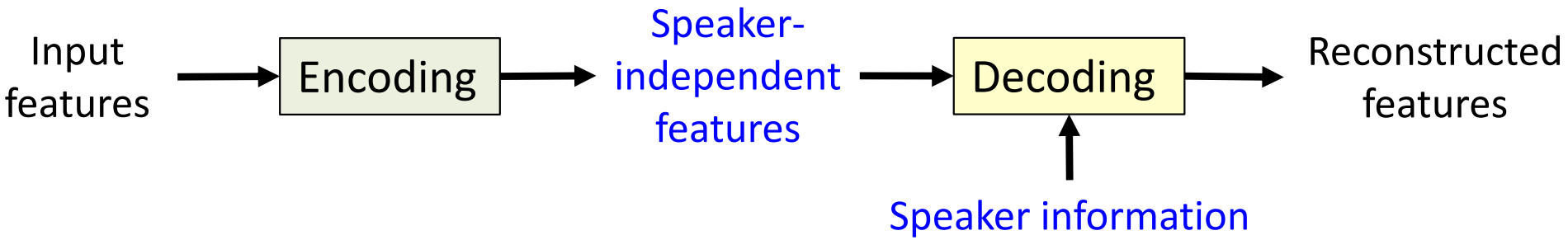


Findings through VCC2018

- Effectiveness of waveform generation process w/o traditional vocoder



- Effectiveness of alignment-free training based on reconstruction process

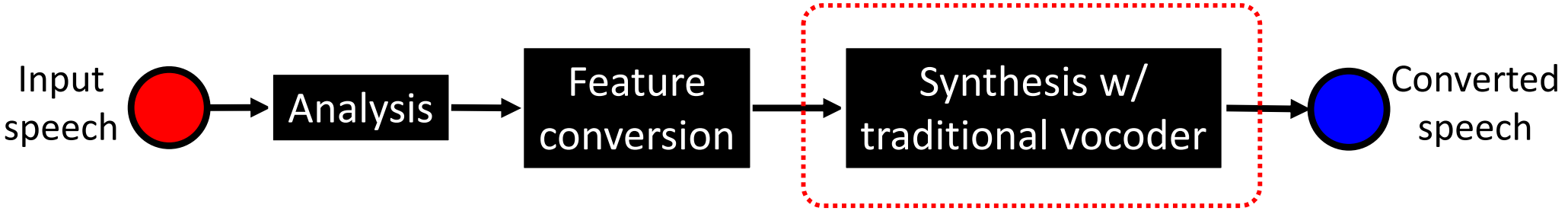


Outline

- Let’s review voice conversion (VC) progress!
  - Basics of VC
    - How to do VC?
    - For what?
  - Recent progress of VC
    - Which VC techniques are really helpful?
    - Let’s review recent Voice Conversion Challenge!
- Let’s review recent progress of waveform modeling!
  - Basics of waveform modeling
    - Let’s revisit vocoder!
  - Progress of waveform modeling in VC
    - How to avoid using vocoder?
    - How to improve vocoder?

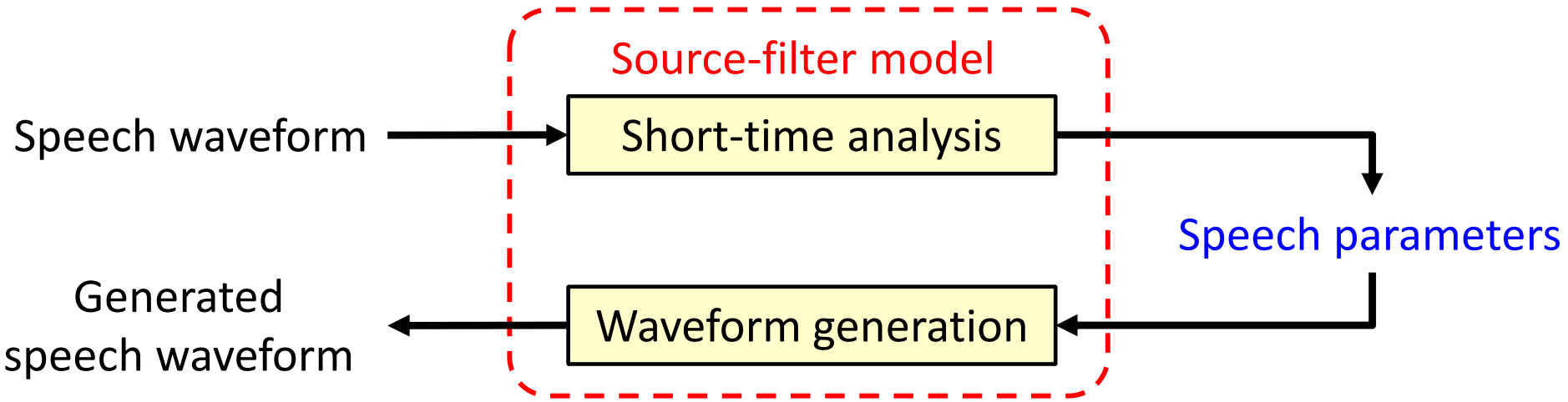
Basics of Waveform Modeling

- Typical approaches
- Probabilistic approach
- Issues to be addressed

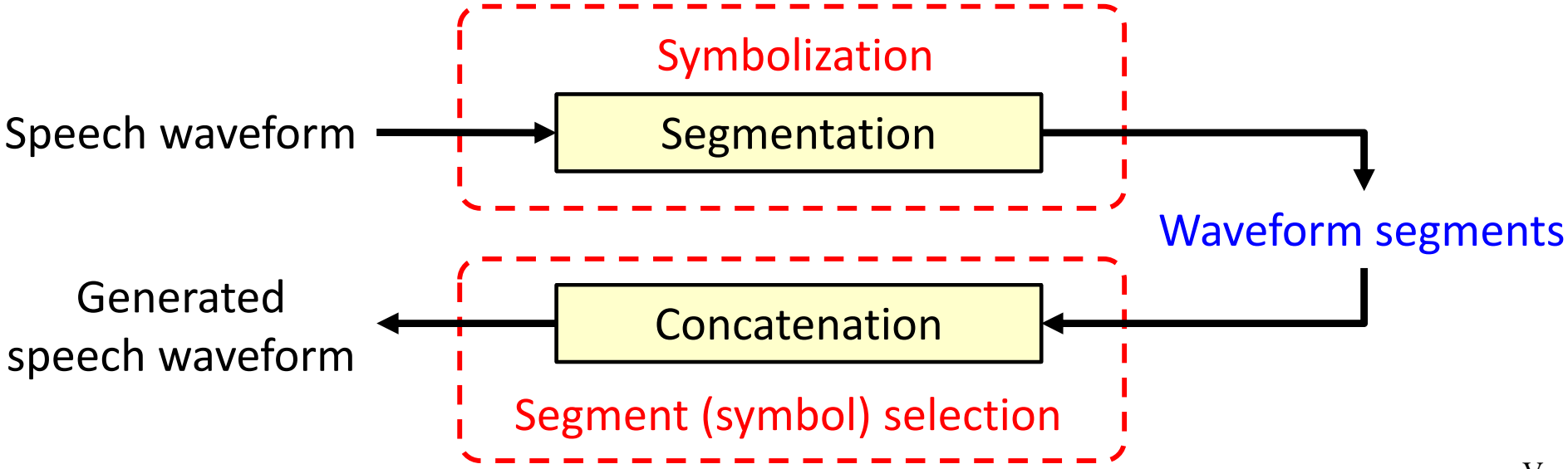


Typical Approaches to Waveform Generation

- Parametric approach (vocoder)



- Concatenative approach



Probabilistic Method for Vocoder

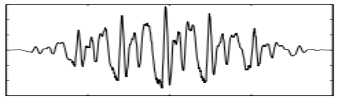
[Itakura; '68]

- Joint probability modeling of speech waveform

$$p(x[1], \ldots, x[N]) = \prod_{n=1}^{N} p(x[n] \mid x[1], \ldots, x[n-1])$$

- Autoregressive (AR) model w/ linear prediction $x[n] = \sum_{d=1}^{D} a_d x[n-d] + e[n]$

Gaussian noise $e[n]$ → $H(z)$: Resonance filter $X(z) = H(z)E(z)$ → Gaussian process $x[n]$

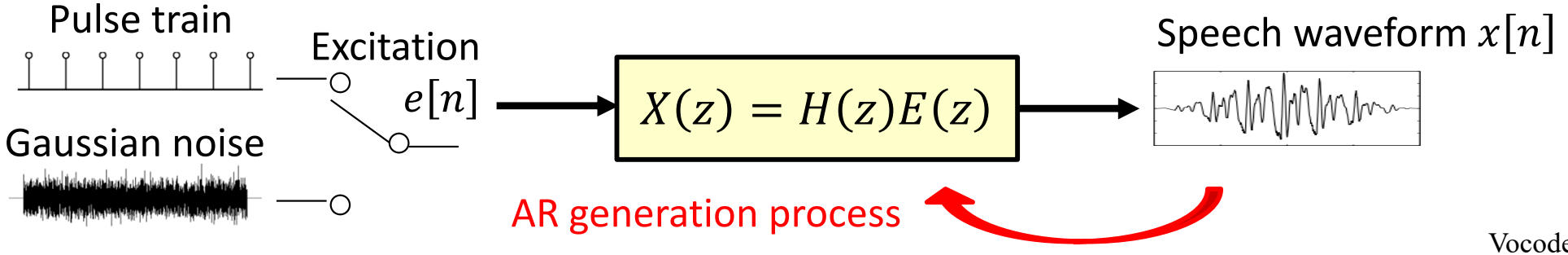


$$p(e[n] \mid \sigma) = \mathcal{N}(0, \sigma^2)$$

$$H(z) = \left(1 - \sum_{d=1}^{D} a_d z^{-d}\right)^{-1}$$

$$p(x[n] \mid x[1], \ldots, x[n-1], a_{1:D}, \sigma) = \mathcal{N}\left(x[n]; \sum_{d=1}^{D} a_d x[n-d], \sigma^2\right)$$

- **Analysis**: use maximum likelihood estimation
- **Synthesis**: use excitation model to generate an excitation signal



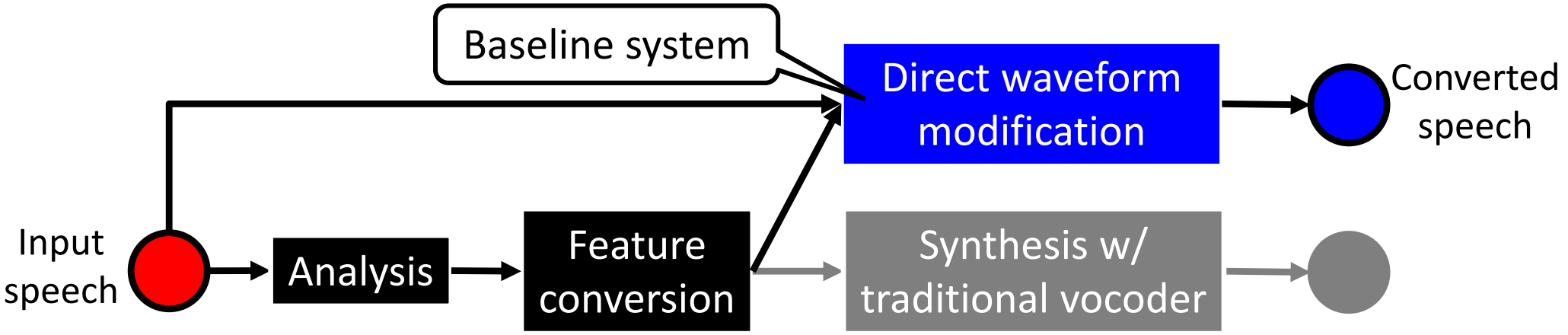
Essential Issues of Traditional Approaches

- Issues of speech waveform parameterization
  - Need to assume **stationary process** in frame analysis (*e.g.*, tackled in [Tokuda; ’15])
  - Need to assume **Gaussian process**
  - Hard to model **temporal structure** (phase components) (*e.g.*, tackled in [Maia; ’13] [Juvela; ’16])
  - Hard to accurately model **fluctuation** (stochastic components)
    - How to model source excitation parameters in the probabilistic approach
    - How to model spectral envelope parameters in the deterministic approach (*e.g.*, tackled in [Toda; ’07] [Takamichi; ’16])
- Issues of waveform segmentation and concatenation
  - **Less flexible** generation process
  - Hard to design a segment **selection function**

I think we didn’t have any perfect solutions until Sep. 2016…

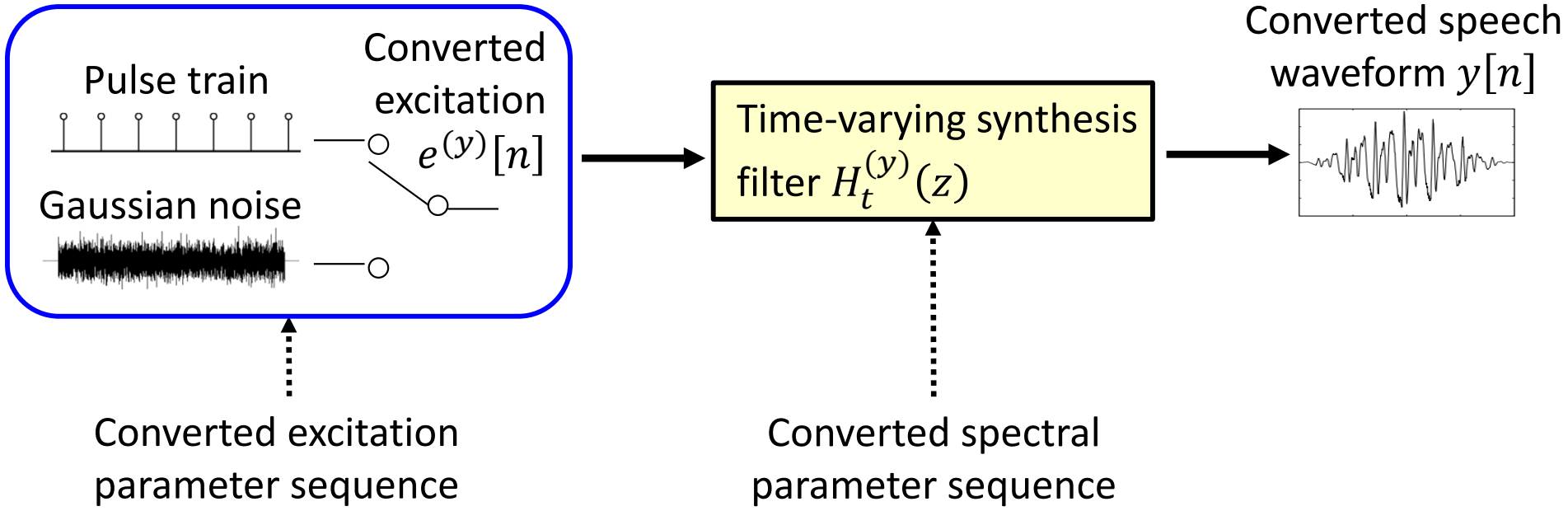
Progress of Waveform Modeling in VC

- **Direct waveform modification**
- **Implementation of neural vocoder**



Difficulties of Excitation Modeling

- Hard to generate a natural excitation signal by using excitation models…

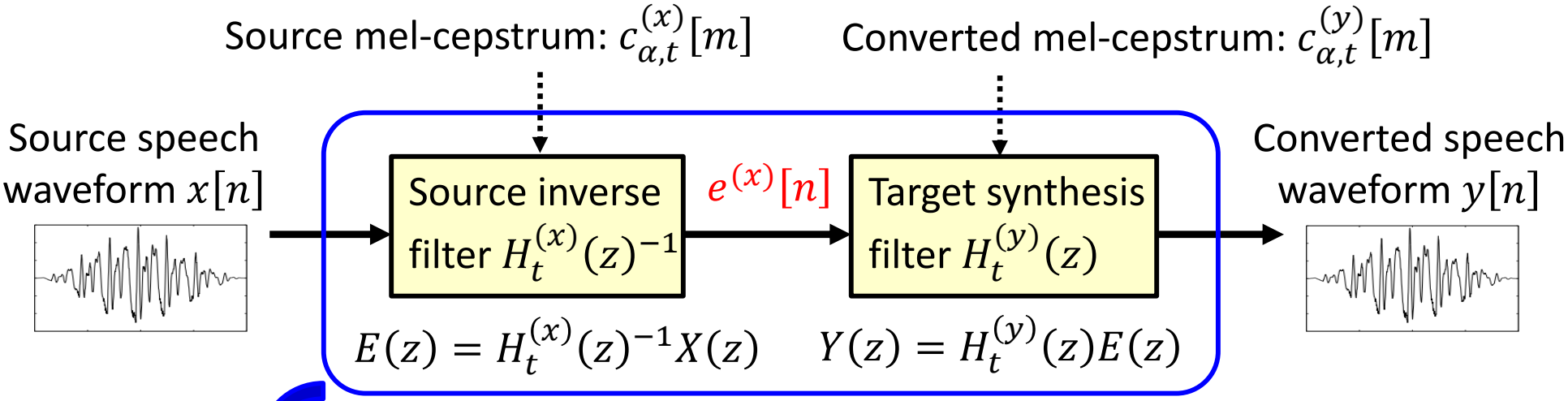


Not necessary to convert excitation parameters in some VC applications, *e.g.*, same-gender singing voice conversion, where F_0 values of source and target voices are similar to each other…

Shall we use natural excitation signals of source speech?

Filtering w/ Mel-Cepstrum Differential

- Convert only spectral parameter sequence (w/ MLSA filter [Imai; '83])



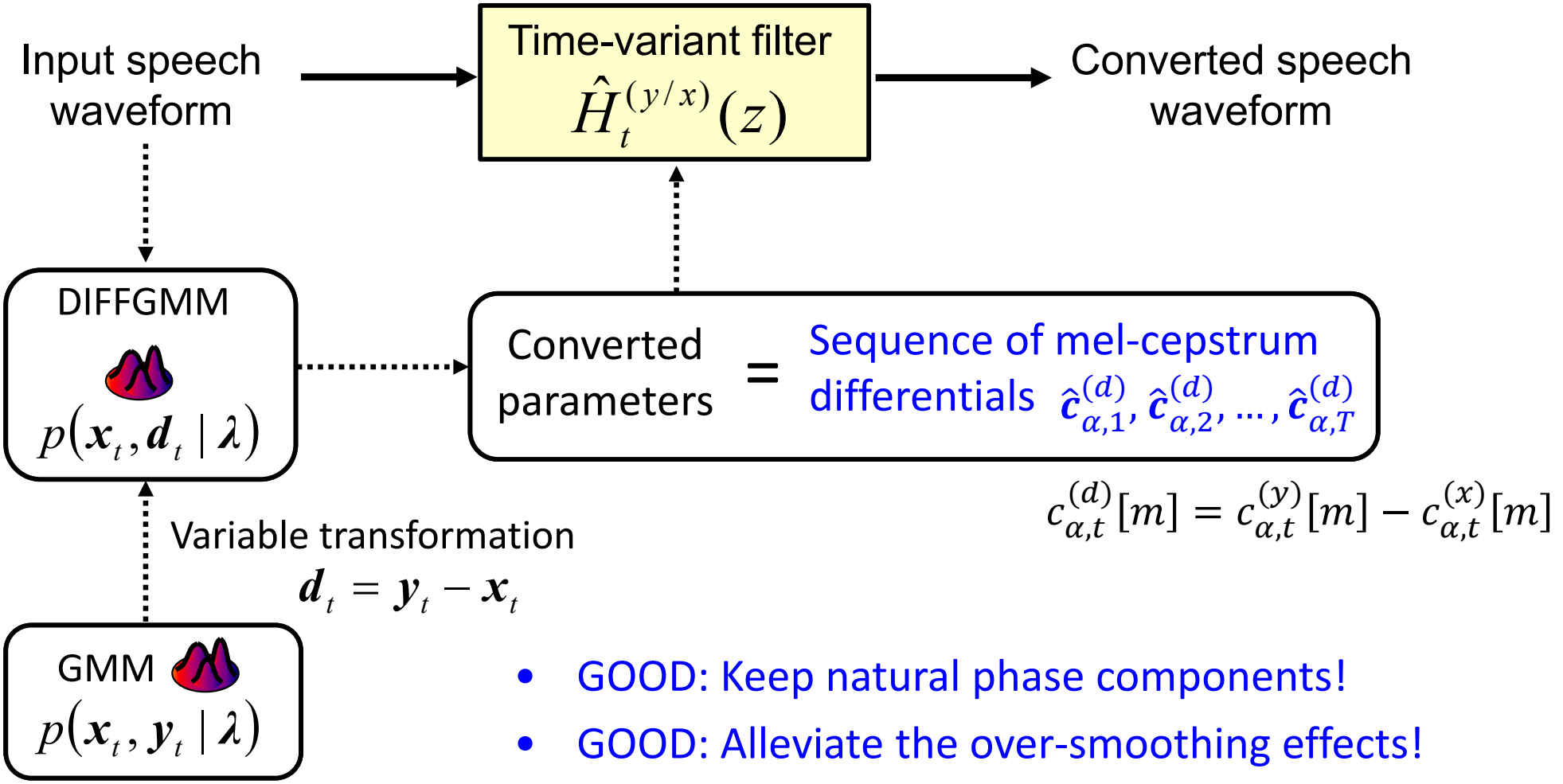
$$H_t^{(y/x)}(z) = \frac{H_t^{(y)}(z)}{H_t^{(x)}(z)} = \frac{\exp\sum_{m=0}^{M} c_{\alpha,t}^{(y)}[m]z_\alpha^{-m}}{\exp\sum_{m=0}^{M} c_{\alpha,t}^{(x)}[m]z_\alpha^{-m}} = \exp\sum_{m=0}^{M}\left(c_{\alpha,t}^{(y)}[m] - c_{\alpha,t}^{(x)}[m]\right)z_\alpha^{-m}$$

Mel-cepstrum differential

DIFFVC: VC w/ Direct Waveform Modification

[Kobayashi; '18a]

- Apply time-variant filtering to input speech waveform to convert its spectral envelope only



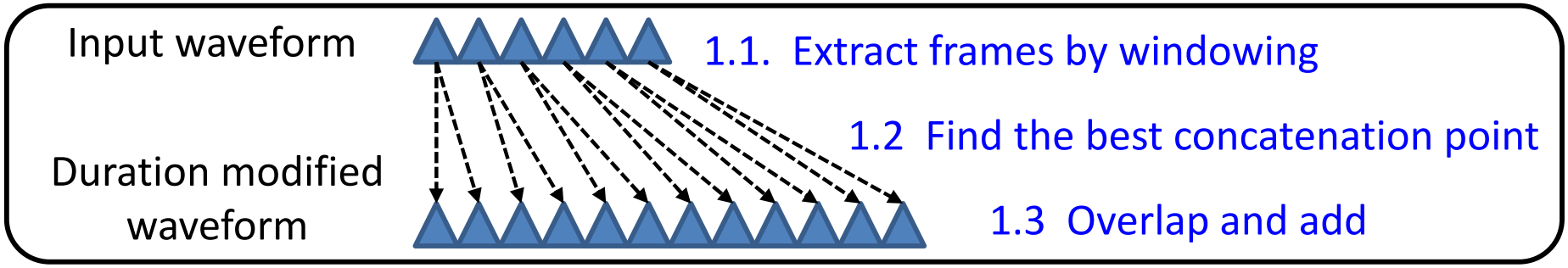
- GOOD: Keep natural phase components!
- GOOD: Alleviate the over-smoothing effects!
- BAD: Not convert excitation parameters (*e.g.*, F_0)

Waveform Modification for F_0 Conversion

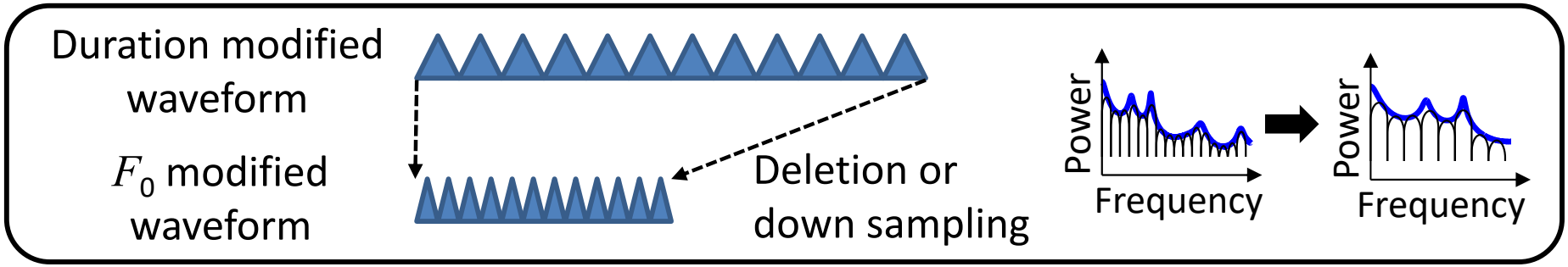
- Use of duration conversion w/ WSOLA and resampling for F_0 conversion

  $e.g.$, if setting F_0 transformation ratio to 2 ($i.e.$, 100 Hz to 200 Hz),

  1. Make duration of input waveform double w/ WSOLA while keeping F_0 values



  2. Resample the modified waveform to make its duration half

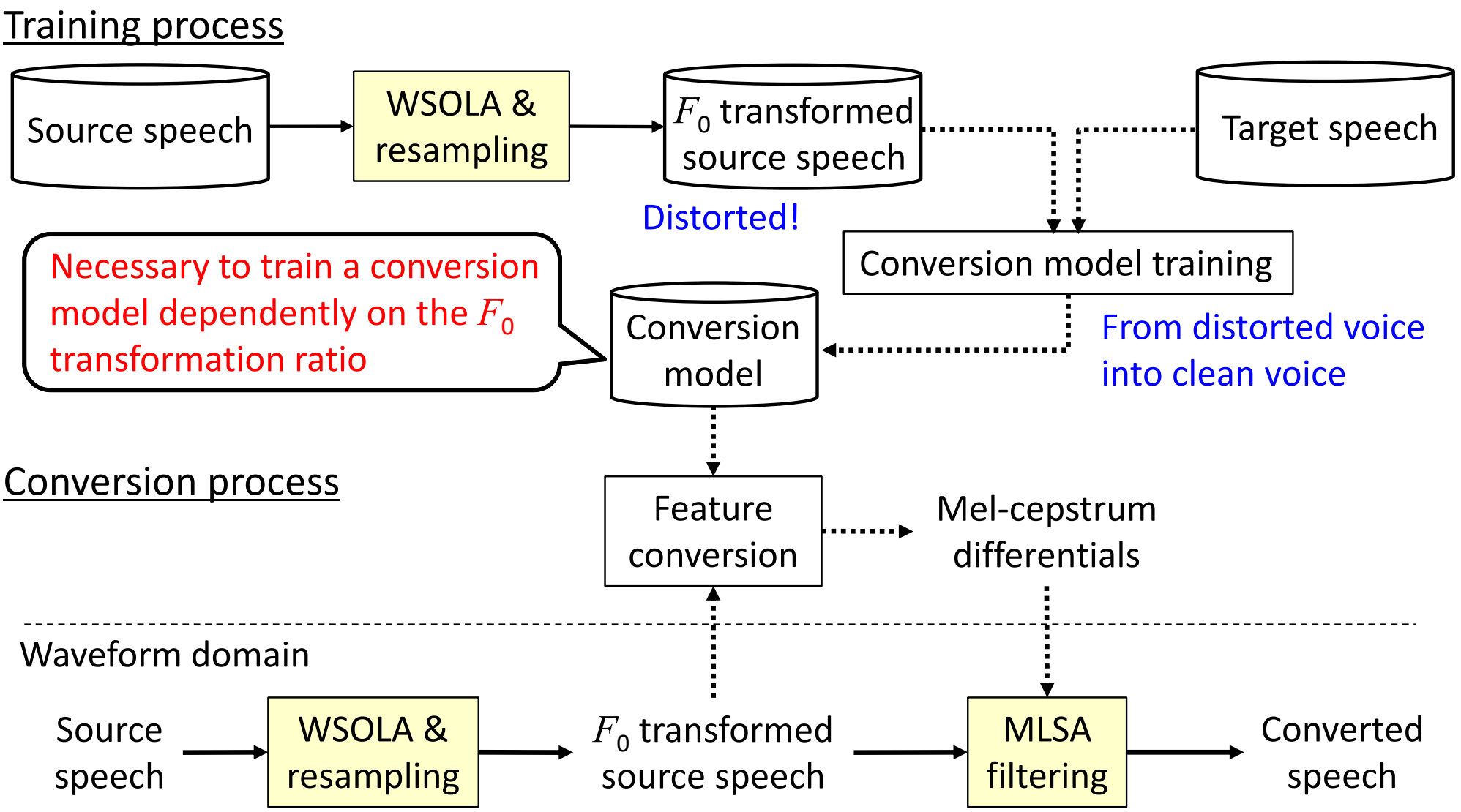


Note that spectrum envelope is also converted due to the frequency warping effect caused by resampling…

DIFFVC w/ F_0 Conversion

Implemented in freely available software: **sprocket**

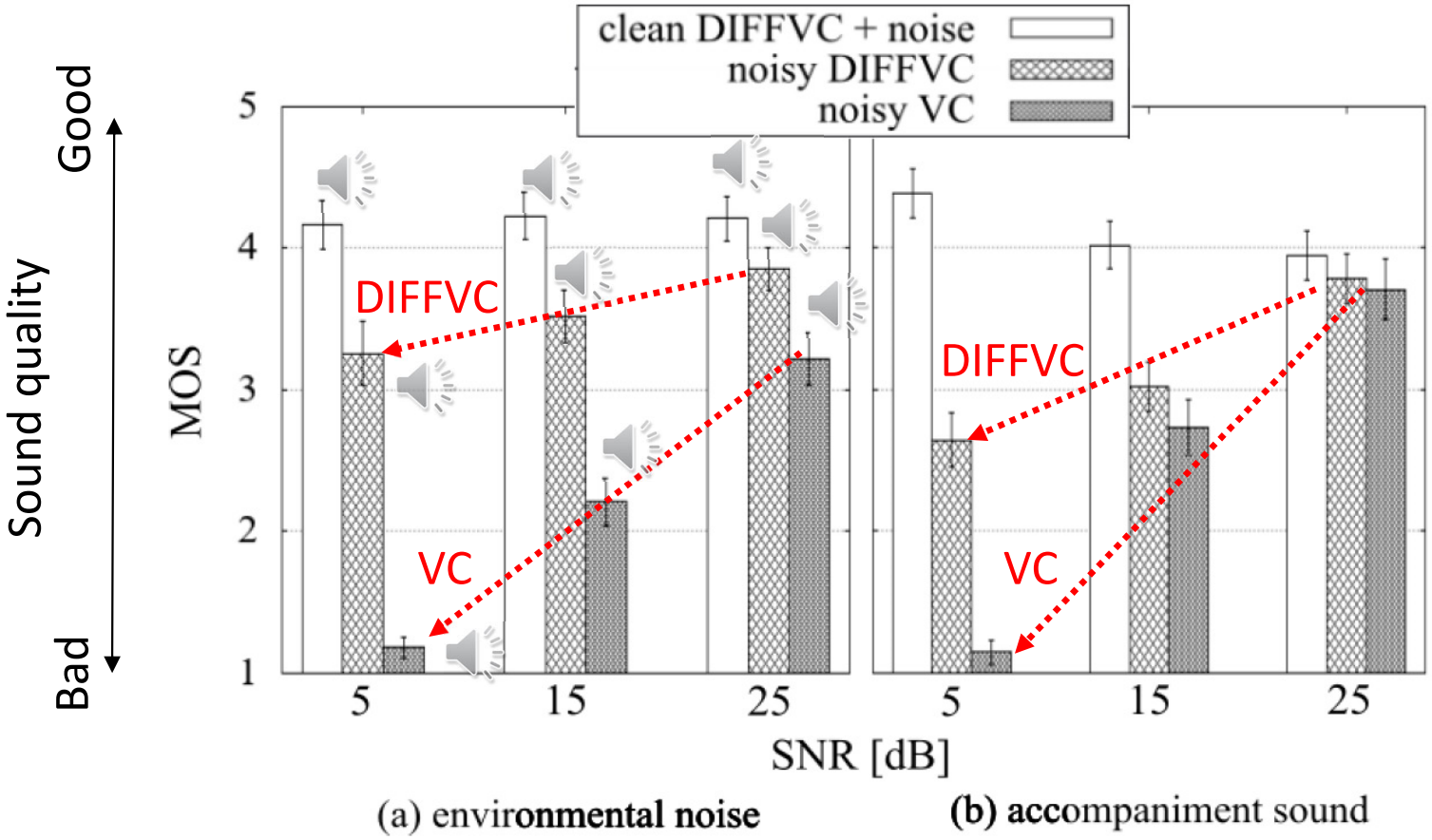
- Use F_0 modified waveform as input speech in spectral conversion



Noise Robustness of DIFFVC

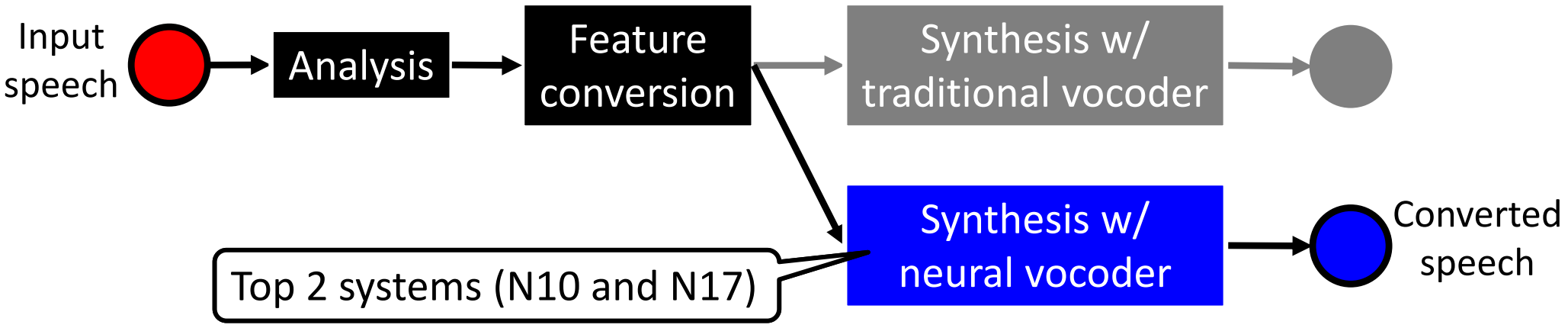
[Kurita; ’19]

- DIFFVC is more robust against background sounds than VC w/ vocoder!
  - Free from error of speech analysis
  - Keep phase components of noisy speech signal



Progress of Waveform Modeling in VC

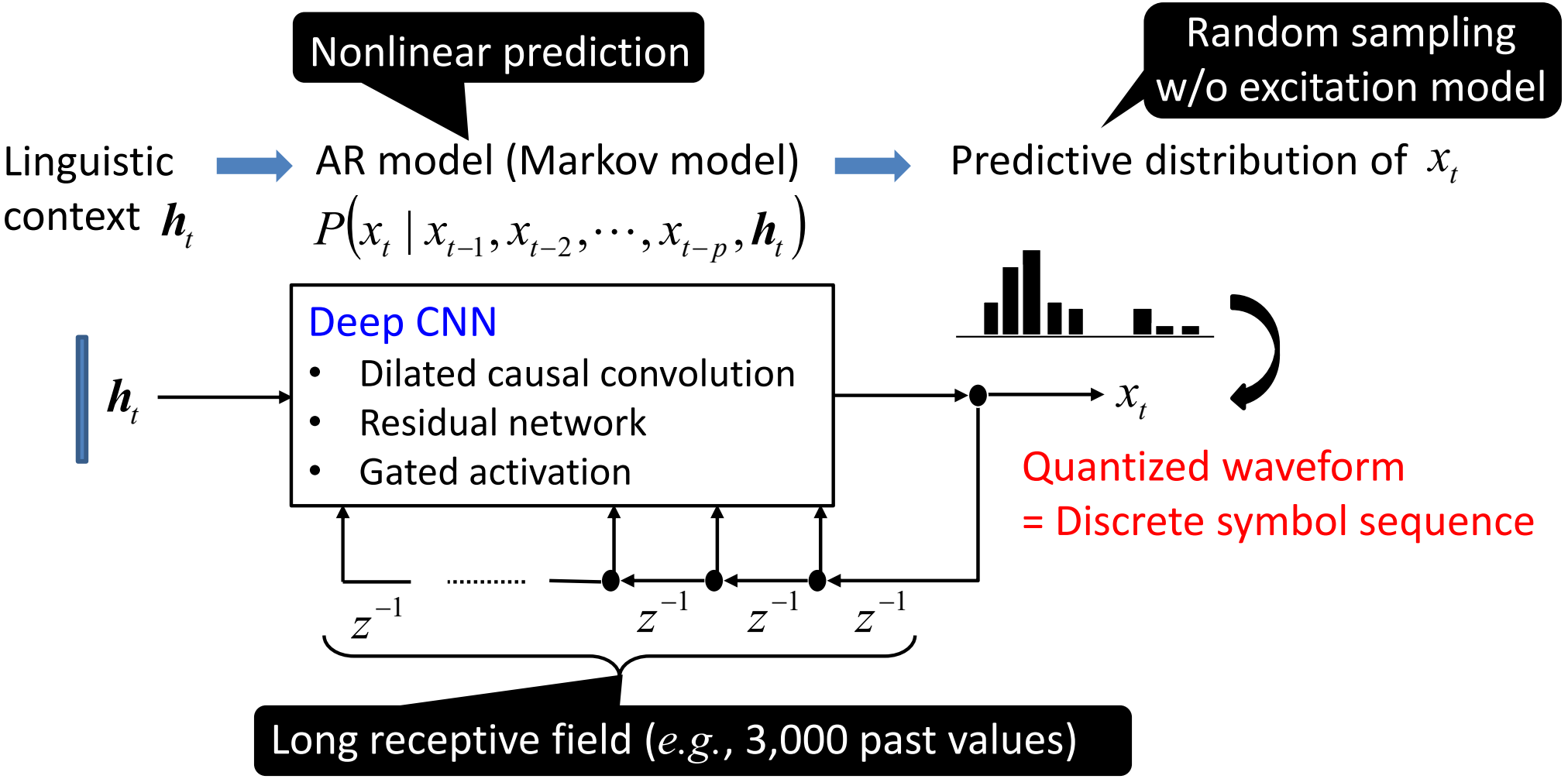
- Direct waveform modification
- **Implementation of neural vocoder**



Epoch-Making: WaveNet

[van den Oord; '16b]

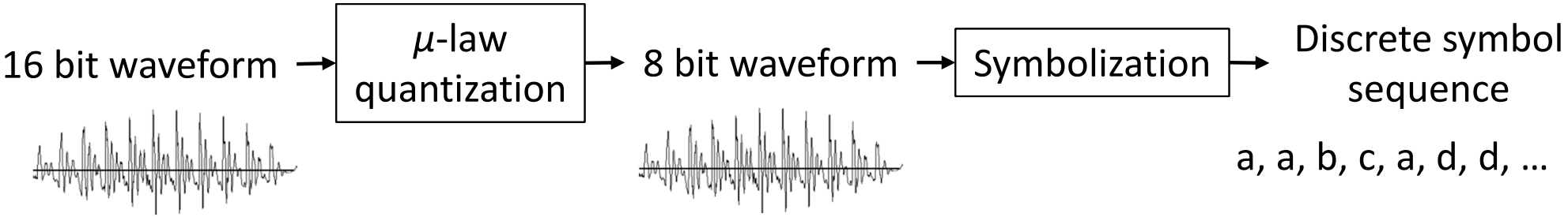
- Probabilistic generation model for waveforms
  - Naturally sounding speech generated by random sampling
  - Capable of well modeling stochastic components of speech signals



Discrete Symbol Sequence Modeling

[van den Oord; '16b]

- Represent speech waveform as discrete symbol sequence
  - 16 bits to 8 bits w/ μ-law quantization
  - Handle discrete symbols w/ 256 classes



- Probability mass modeling w/ higher-order Markov model ($i.e.$, AR model for discrete variables)
  - Formulated as classification problem (256 classes at each time sample)
  - Similar to the concatenative approach!

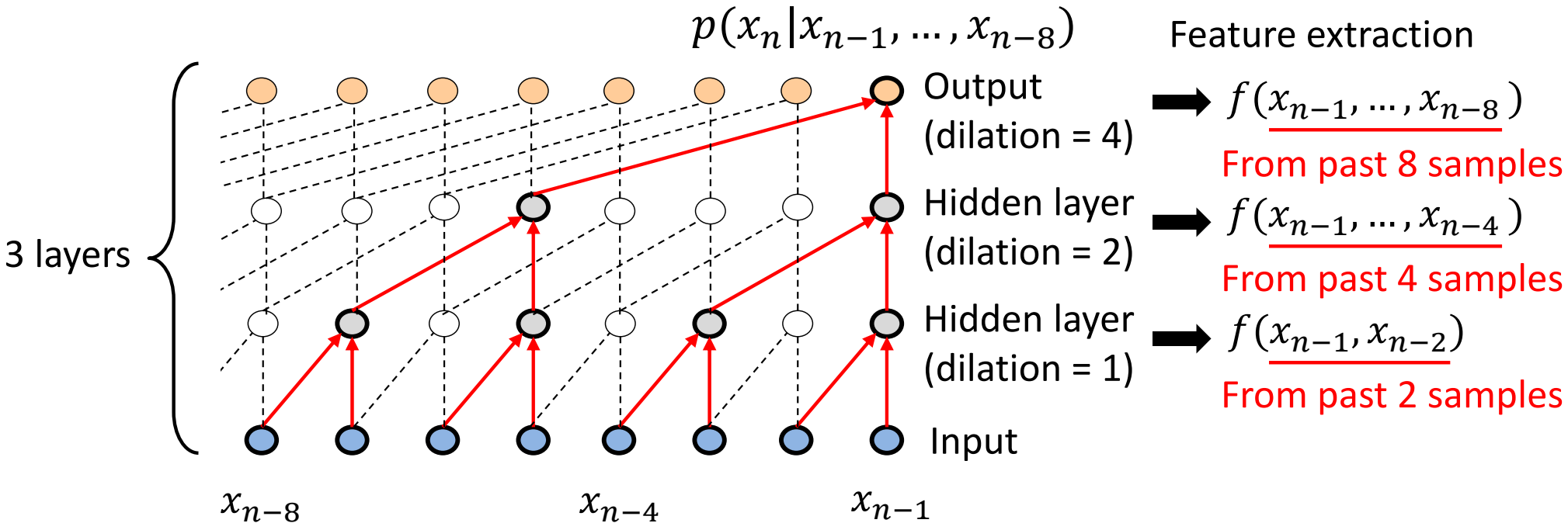
$$p(x_1, \ldots, x_N) = \prod_{n=1}^{N} p(x_n | x_1, \ldots, x_{n-1}) \cong \prod_{n=1}^{N} p(x_n | x_{n-L}, \ldots, x_{n-1})$$

Dependent on all past samples — Dependent only past L samples

Dilated Causal Convolution

[van den Oord; '16b]

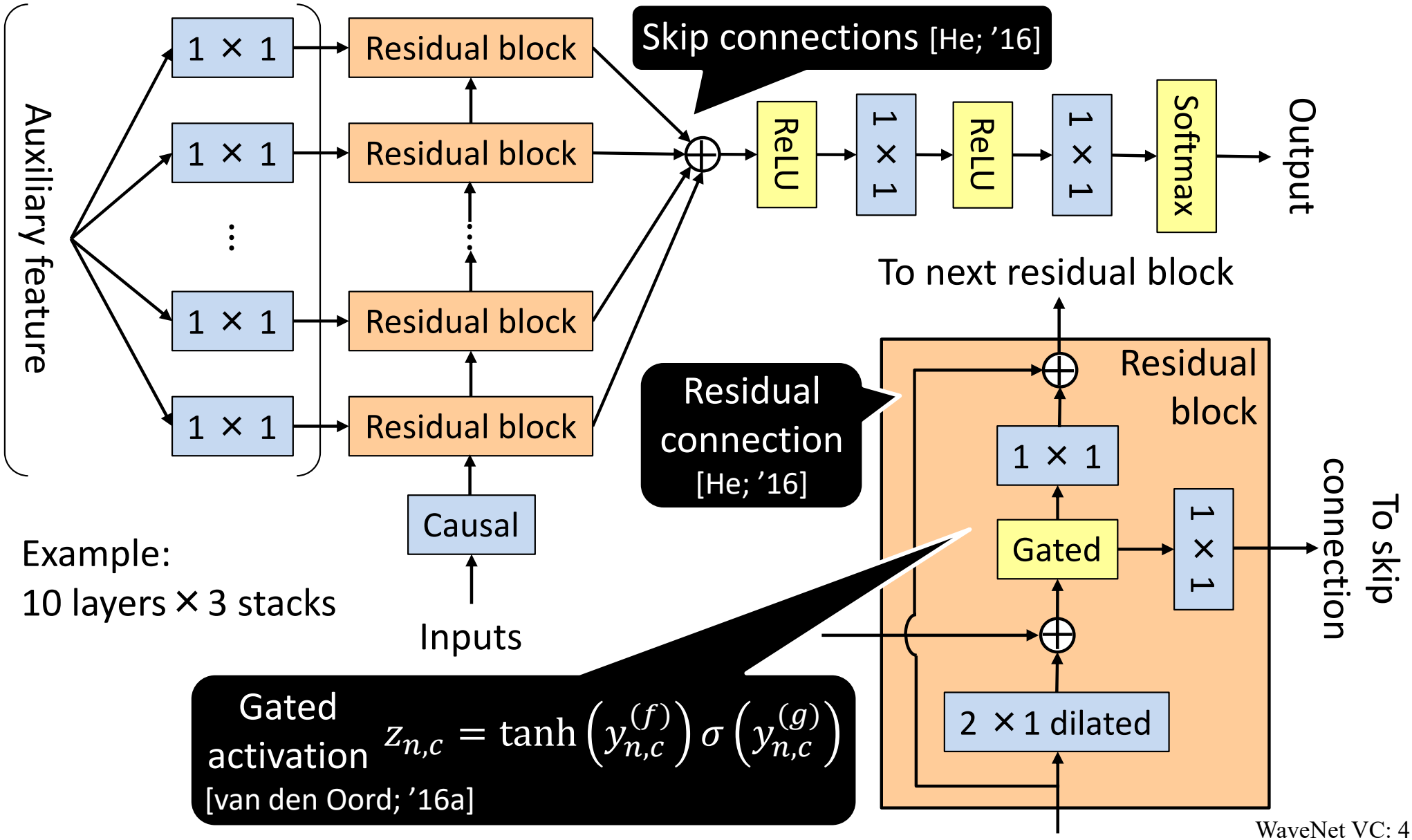
- Efficient convolution over many past samples (*i.e.*, looooong history)



8 × 1 convolution is achieved by using 2 × 1 convolution 3 times!

Network structure

- Predict output using all features extracted at individual layers



Training Process and Generation Process

- Training process
  - Maximize likelihood function of Markov model (= cross-entropy minimization)

$$\text{argmax}\, p(x_1, \ldots, x_N) = \text{argmin} \; -\sum_{n=1}^{N} \ln p(x_n | x_{n-L}, \ldots, x_{n-1})$$

- Generation process
  - Random sampling one by one as auto-regressive model

Predictive distribution (256 classes) at time step n

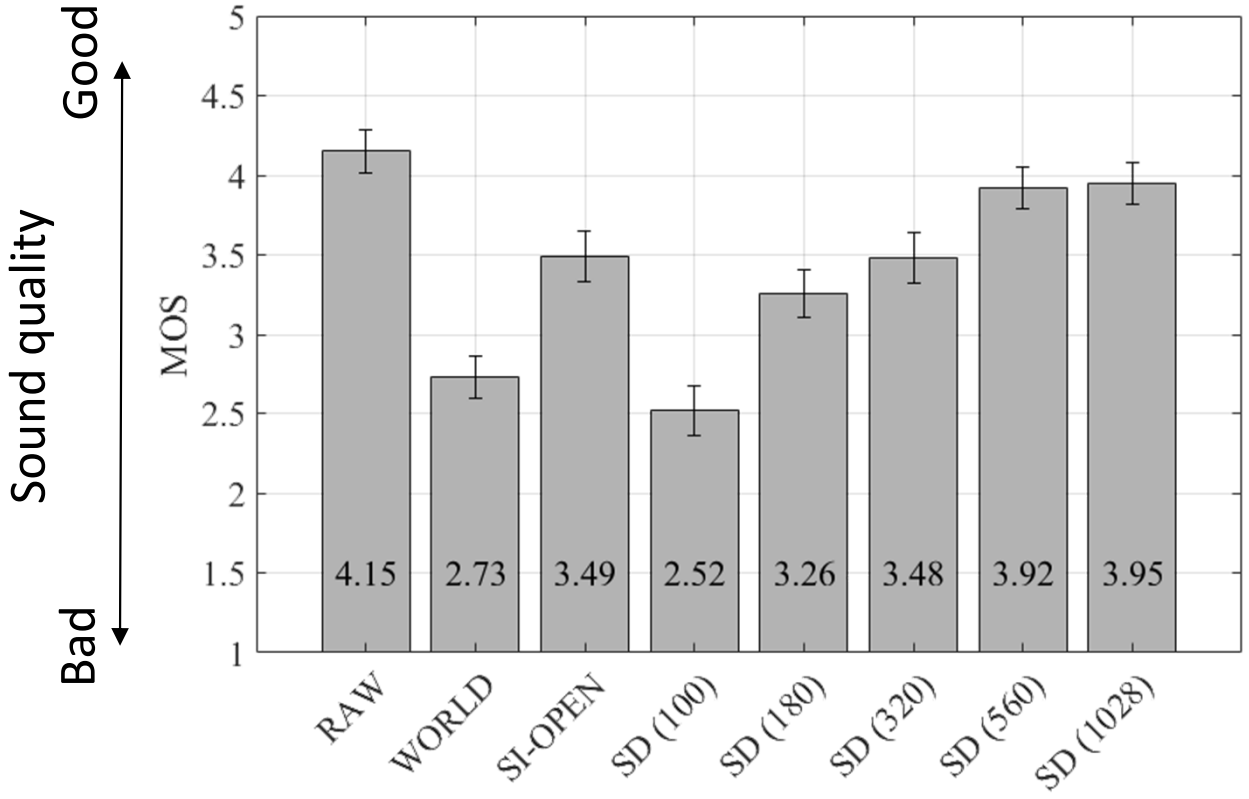
$$\hat{x}_n \sim p(x_n | \hat{x}_{n-L}, \ldots, \hat{x}_{n-1})$$

Already generated past L samples

Implementation of WaveNet as Vocoder

[Tamamori; '17]

- Use acoustic features, such as vocoder parameters or mel-spectrogram, as auxiliary features
  - Need to adjust their time-resolution to that of waveform, *e.g.*, use upsampling layer to convert 200 Hz feature sequence (*i.e.*, 5 ms shift) to 16 kHz
- Capable of generating naturally sounding speech waveform even if using only 500 utterances in speaker-dependent WaveNet training [Hayashi; '17]



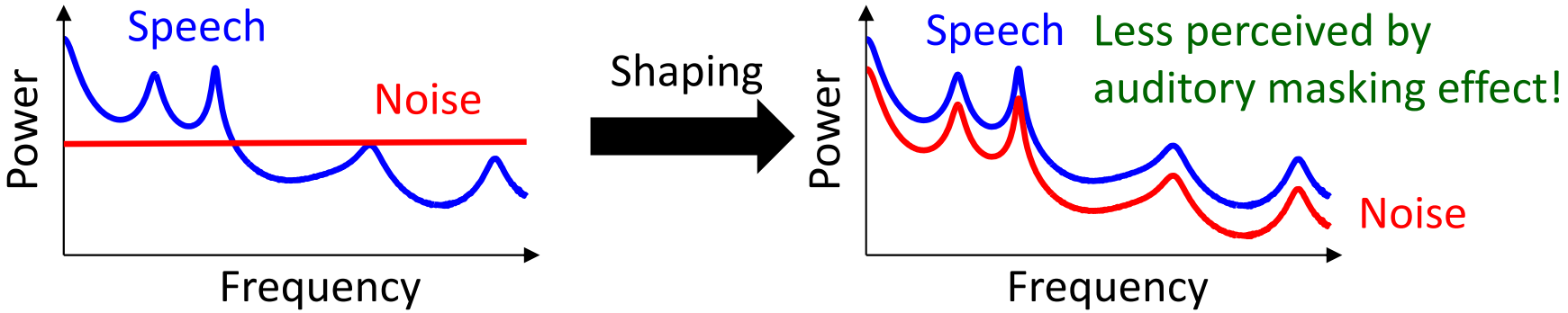
Comparison to Traditional Approaches

	Probabilistic approach (vocoder)	Concatenative approach	WaveNet vocoder
Stationary assumption	Necessary	Not necessary	Not necessary
Gaussian assumption	Necessary	Not necessary	Not necessary
Phase modeling	Hard	Copied w/ exemplar	Well handled
Fluctuation modeling	Hard	Copied w/ exemplar	Well handled
Generation process	Random sampling w/ excitation model	Exemplar selection	Random sampling w/o excitation model
Optimization	Well formulated	Not well formulated	Well formulated
Minimum unit	Sample-by-sample	Segment-by-segment	Sample-by-sample
Training data	Not necessary	Huge-sized data	Large-sized data
Controllability	Very high	Very limited	Quite high but still limited

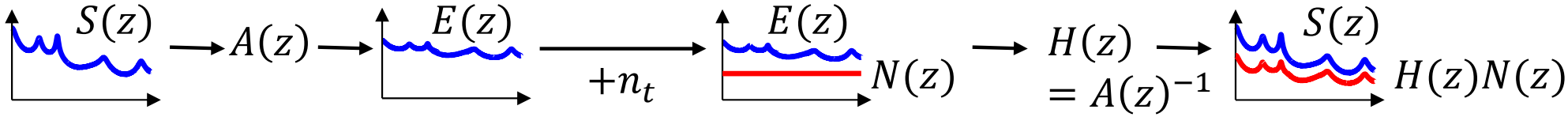
WaveNet vocoder may be regarded as a hybrid approach (*i.e.*, sample-by-sample selection)!

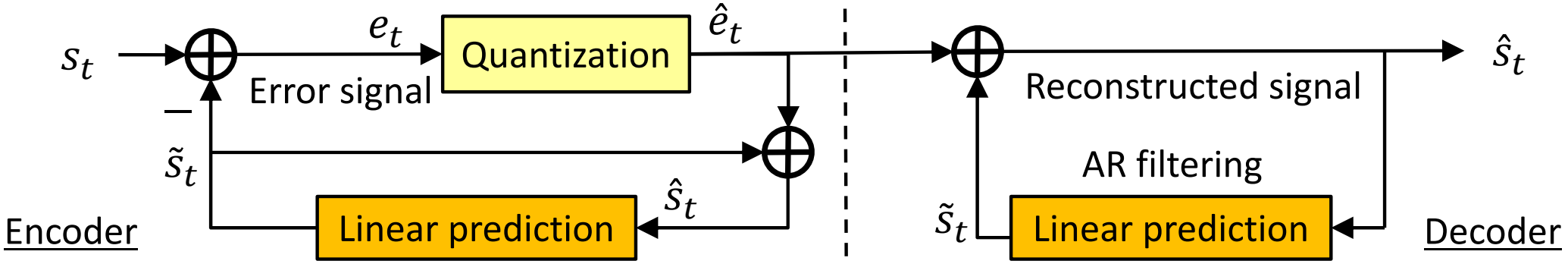
Effective Technique: Noise Shaping

- Perceptually suppress noises caused in waveform generation process
  - Control their frequency patterns to make them hardly perceived



Example: predictive pulse code modulation (PPCM) [Atal; '78]





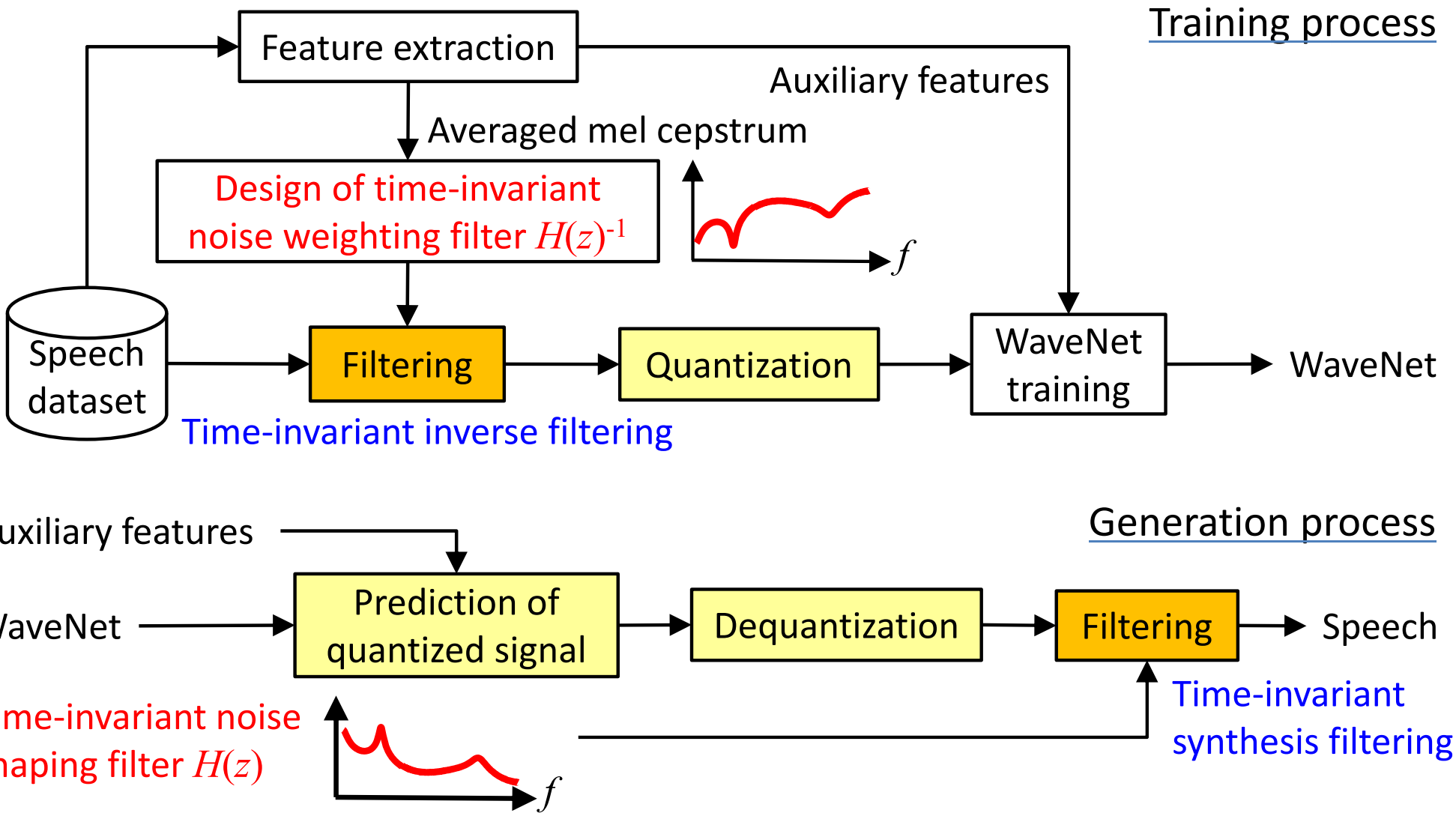
Quantize the error signal e_t (with flatter spectral envelope) generated by LP analysis

Reconstruct the signal by inverse-filtering the quantized error signal $\hat{e}_t$ $(= e_t + n_t)$

Implementation of Noise Shaping

Implemented in freely-available software: **PytorchWaveNetVocoder**

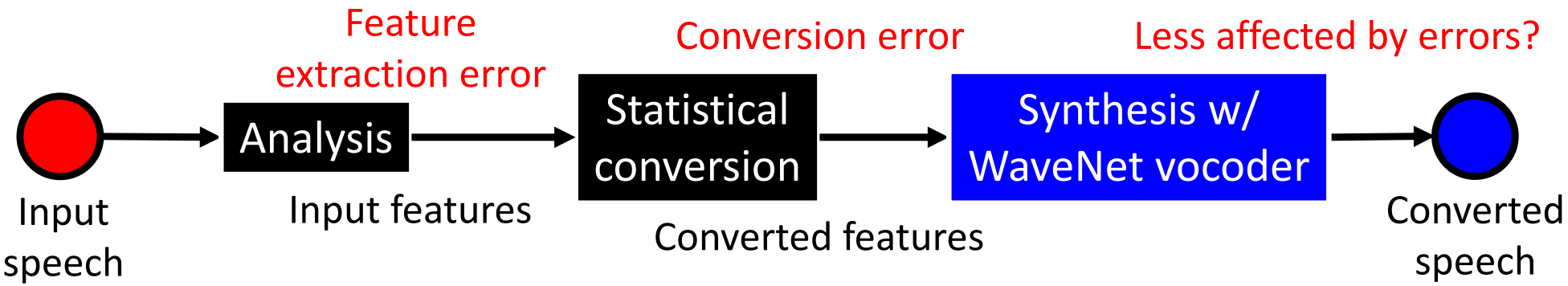
- Applied to both prediction and quantization noises [Tachibana; '18] rather than only quantization noise [Yoshimura; '18]



VC with WaveNet Vocoder

Can be developed with **sprocket** & **PytorchWaveNetVocoder**

- Implementation of WaveNet as a data-driven vocoder for VC
  - Significant improvement of speaker similarity yielded by just using WaveNet vocoder in VC [Kobayashi; '17]



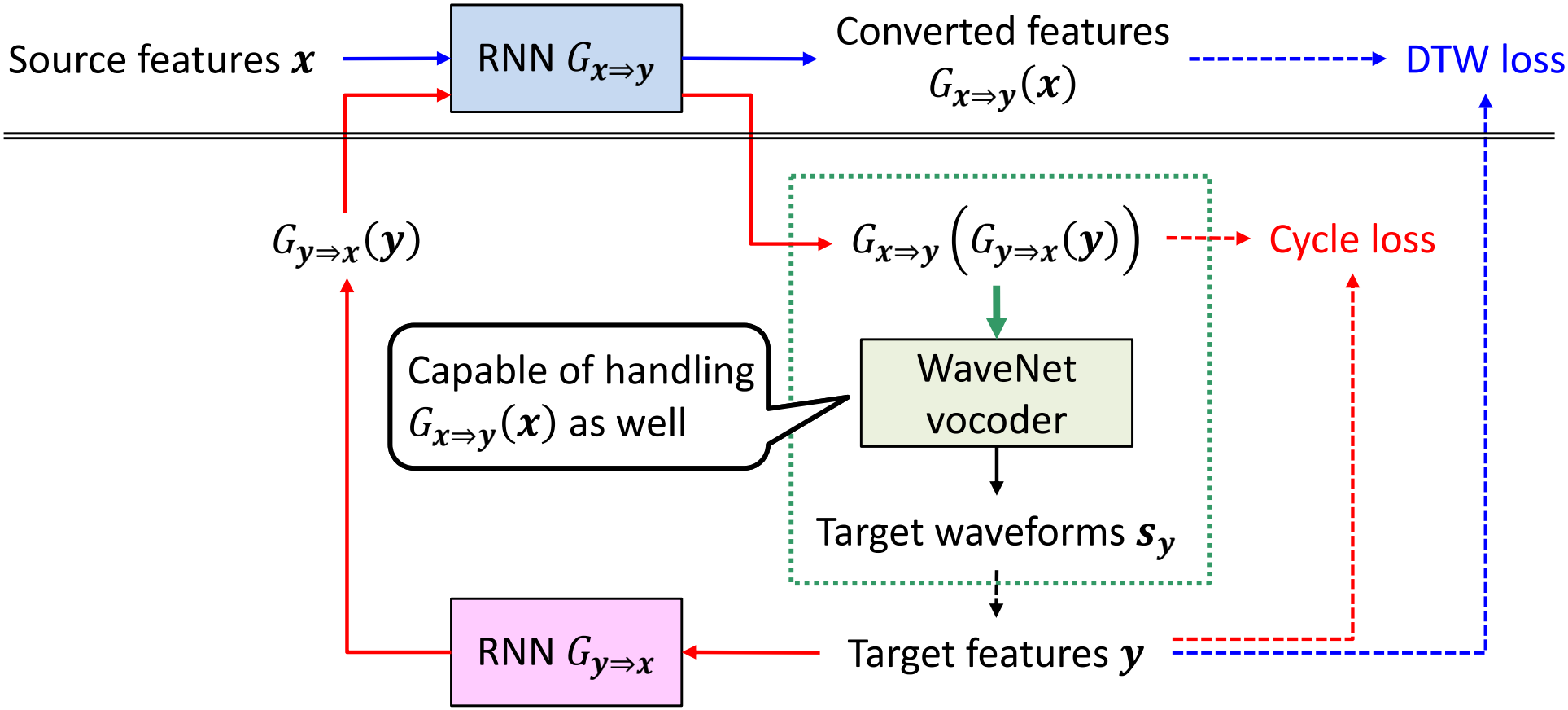
  - Could also reduce adverse effects of some errors on converted speech by training WaveNet vocoder using the converted features

However, it is hard to train WaveNet vocoder directly using the converted features owing to different temporal structures (*i.e.*, time-alignment issue)…

WaveNet Fine-Tuning w/ CycleRNN

[Tobing; '19]

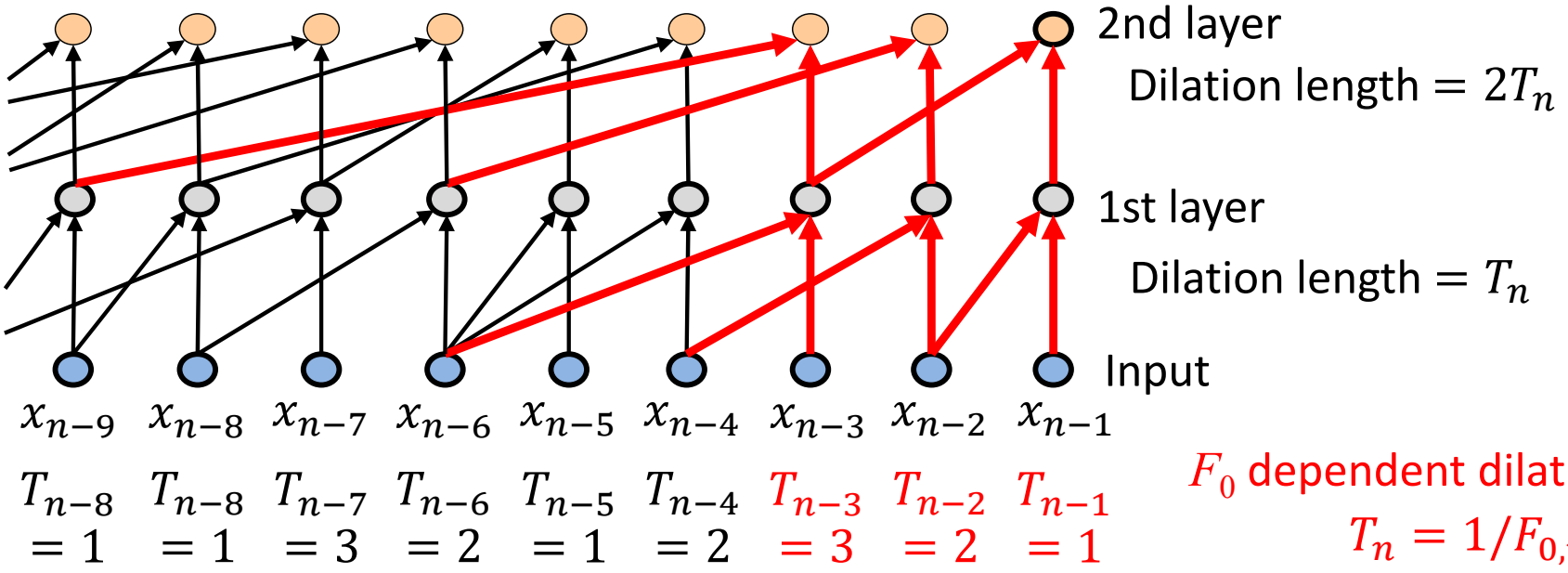
- Generate training data for training WaveNet vocoder
  - Use cyclic conversion (as intra-speaker conversion [Kobayashi; '17])
  - Reduce acoustic mismatches between training and conversion
  - Free from temporal structure mismatches between features and waveforms



Quasi-Periodic WaveNet (QPNet)

[Wu; '19]

- Dynamically change dilation length based on F_0 value
  - Significantly improve F_0 controllability and reduce the network size



F_0 dependent dilation length:

$T_n = 1/F_{0,n}$

- QPNet structure
  - Lower layers: dilated causal convolution for short-term prediction
  - Upper layers: F_0-dependent dilated causal convolution for long-term prediction

Summary

- Reviewed VC progress!
  - Basics of VC
    - Basic framework of statistical VC
    - Many useful applications
    - Statistical VC = kitchen knife
  - Improvements of VC
    - Evaluation through voice conversion challenges
    - Improvements of waveform generation and nonparallel training
- Reviewed recent progress of waveform modeling!
  - Basics of waveform modeling
    - Essential issues of waveform generation with traditional vocoder
  - Progress of waveform modeling in VC
    - DIFFVC based on direct waveform modification to avoid using vocoder
    - Implementation of WaveNet vocoder for VC and further improvements

Available Resources

- Tutorial materials at INTERSPEECH 2019
  - **https://bit.ly/328LwSS**
  - Lecture slides
  - Hands-on
    - Google Colab note
    - Development of VC w/ WaveNet vocoder
      - Baseline system: sprocket
      - WaveNet vocoder: PytorchWaveNetVocoder

- Summer school materials at SPCC 2018 (& 2019)
  - Lecture slides on “Advanced Voice Conversion”
    - **https://bit.ly/2PpWEYx**
    - More details of recent progress of VC techniques
  - Hands-on slides
    - **https://bit.ly/2pmwuLC**
    - More details of sprocket to develop VCC2018 baseline system

References

[Abe; ’90] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara. Voice conversion through vector quantization. *J. Acoust. Soc. Jpn (E)*, Vol. 11, No. 2, pp. 71–76, 1990.

[Atal; ’78] B.S. Atal, M.R. Scroeder .Predictive coding of speech signals and subjective error criteria. *Proc. IEEE ICASSP*, pp. 247–254, 1978.

[He; ’16] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. *Proc. CVPR*, pp. 770–778, 2016.

[Hayashi; ’17] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, T. Toda. An investigation of multi-speaker training for WaveNet vocoder. *Proc. IEEE ASRU*, pp. 698–704, 2017.

[Imai; ’83] S. Imai, K. Sumita, C. Furuichi. Mel log spectrum approximation (MLSA) filter for speech synthesis. Electron. Commun. Japan (Part 1: Communications), Vol. 66, No. 2, pp. 10–18, 1983.

[Itakura; ’68] F. Itakura, S. Saito. Analysis synthesis telephony based upon the maximum likelihood method. *Proc. ICA*, C-5-5, pp. C17–20, 1968.

[Juvela; ’16] L. Juvela, B. Bollepalli, M. Airaksinen, P. Alku. High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network. *Proc. IEEE ICASSP*, pp. 5120–5124, 2016.

[Kawahara; ’99] H. Kawahara, I. Masuda-Katsuse, A. de Cheveigne. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based $F0$ extraction: possible role of a repetitive structure in sounds. *Speech Commun.,* Vol. 27, No. 3–4, pp. 187–207, 1999.

[Kinnunen; ’17] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, K.A. Lee. The ASVspoof 2017 Challenge: assessing the limits of replay spoofing attack detection. *Proc. INTERSPEECH*, pp. 2--6, 2017.

[Kobayashi; ’17] K. Kobayashi, T. Hayashi, A. Tamamori, T. Toda. Statistical voice conversion with WaveNet-based waveform generation. *Proc. INTERSPEECH*, pp. 1138–1142, 2017.

[Kobayashi; ’18a] K. Kobayashi, T. Toda, S. Nakamura. Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential. *Speech Commun.*, Vol. 99, pp. 211–220, 2018.

[Kobayashi; ’18b] K. Kobayashi, T. Toda. sprocket: open-source voice conversion software. *Proc. Odyssey*, pp. 203–210, 2018.

[Kurita; ’19] Y. Kurita, K. Kobayashi, K. Takeda, T. Toda. Robustness of statistical voice conversion based on direct waveform modification against background sounds. *Proc. INTERSPEECH*, pp. 684–688, 2018.

[Liu; ’18] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, L.-R. Dai. WaveNet Vocoder with Limited Training Data for Voice Conversion. *Proc. INTERSPEECH*, pp. 1983–1987, 2018.

[Lorenzo-Trueba; ’18] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, Z. Ling. The voice conversion challenge 2018: promoting development of parallel and nonparallel methods. *Proc. Odyssey*, pp. 195–202, 2018.

[Maia; ’13] R. Maia, M. Akamine, M. Gales. Complex cepstrum for statistical parametric speech synthesis. *Speech Commun.*, Vol. 55, No. 5, pp. 606–618, 2013.

[Morise; ’16] M. Morise, F. Yokomori, K. Ozawa. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. & Syst.*, Vol. E99-D, No. 7, pp. 1877–1884, 2016.

[Mysore, ’15] G. J. Mysore. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? – a dataset, insights, and challenges. *IEEE Signal Process. Letters*, Vol. 22, No. 8, pp. 1006–1010, 2015.

[Pantazis; ’11] Y. Pantazis, O. Rosec, Y. Stylianou. Adaptive AM–FM signal decomposition with application to speech analysis. *IEEE Trans. Audio, Speech, & Lang. Process.*, Vol. 19, No. 2, pp. 290–300, 2011.

[Tachibana; ’18] K. Tachibana, T. Toda, Y. Shiga, H. Kawai. An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation. *Proc. IEEE ICASSP*, pp. 5664–5668, 2018.

[Takamichi; ’16] S. Takamichi, T. Toda, A.W. Black, G. Neubig, S. Sakti, S. Nakamura. Post-filters to modify the modulation spectrum for statistical parametric speech synthesis. *IEEE/ACM Trans. Audio, Speech & Lang. Process.*, Vol. 24, No. 4, pp. 755–767, 2016.

[Tamamori; ’17] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, T. Toda. Speaker-dependent WaveNet vocoder. *Proc. INTERSPEECH*, pp. 1118–1122, 2017.

[Tobing; ’18] P.L. Tobing, Y. Wu, T. Hayashi, K. Kobayashi, T. Toda. NU voice conversion system for the voice conversion challenge 2018. *Proc. Odyssey*, pp. 219–226, 2018.

[Tobing; ’19] P.L. Tobing, Y. Wu, T. Hayashi, K. Kobayashi, T. Toda. Voice conversion with cyclic recurrent neural network and fine-tuned WaveNet vocoder. *Proc. IEEE ICASSP*, pp. 6815–6819, 2019.

[Toda; ’07] T. Toda, A.W. Black, K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio, Speech & Lang. Process.*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[Toda, ’12] T. Toda, T. Muramatsu, H. Banno. Implementation of computationally efficient real-time voice conversion. *Proc. INTERSPEECH*, 4 pages, 2012.

[Toda, ’14] T. Toda. Augmented speech production based on real-time statistical voice conversion. *Proc. GlobalSIP*, pp. 755–759, 2014.

[Toda; ’16] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, J. Yamagishi. The Voice Conversion Challenge 2016. *Proc. INTERSPEECH*, pp. 1632–1636, 2016.

[Tokuda; ’94] K. Tokuda, T. Kobayashi, T. Masuko, S. Imai. Mel-generalized cepstral analysis —a unified approach to speech spectral estimation. *Proc. ICSLP*, vol.3, pp.1043–1046, 1994.

[Tokuda; ’15] K. Tokuda, H. Zen. Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis. *Proc. IEEE ICASSP*, pp. 4215–4219, 2015

[van den Oord; ’16a] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, K. Kavukcuoglu. Conditional image generation with PixelCNN decoders. *arXiv preprint*, arXiv:1606.05328, 13 pages, 2016.

[van den Oord; ’16b] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. WaveNet: a generative model for raw audio. *arXiv preprint*, arXiv:1609.03499, 15 pages, 2016.

[Wu; ’15] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, H. Li. Spoofing and countermeasures for speaker verification: A survey. *Speech Commun*. Vol. 66, pp. 130–153, 2015.

[Wu; ’17] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilci, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, H. Delgado. ASVspoof: the automatic speaker verification spoofing and countermeasures challenge. *IEEE J. Sel. Topics in Signal Process*., Vol. 11, No. 4, pp. 588–604, 2017.
[Wu; ’18] Y.-C. Wu, P.L. Tobing, T. Hayashi, K. Kobayashi, T. Toda. The NU non-parallel voice conversion system for the voice conversion challenge 2018. *Proc. Odyssey,* pp. 211–218, 2018.
[Wu; ’19] Y.-C. Wu, T. Hayashi, P.L. Tobing, K. Kobayashi, T. Toda. Quasi-periodic WaveNet vocoder: a pitch dependent dilated convolution model for parametric speech generation. *Proc. INTERSPEECH,* pp. 196–200, 2019.
[Yoshimura; ’18] T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda. Mel-cepstrum-based quantization noise shaping applied to neural-network-based speech waveform synthesis. IEEE/ACM Trans. Audio, Speech & Lang. Process., Vol. 26, No. 7, pp. 1173–1180, 2018.

<Special issues>

- E. Moulines, Y. Sagisaka, Voice conversion: state of the art and perspectives. *Speech Commun.*, Vol. 16, No. 2, 1995.
- Y. Stylianou, T. Toda, C.-H. Wu, A. Kain, O. Rosec. The special section on voice transformation. *IEEE Trans. Audio, Speech & Lang.*, Vol. 18, No. 5, 2010.

<Survey>

- H. Mohammadi, A. Kain. An overview of voice conversion systems. *Speech Commun*. Vol. 88, pp. 65–82, 2017.

<Software>

- K. Kobayashi. sprocket. *https://github.com/k2kobayashi/sprocket*
- T. Hayashi. PytorchWaveNetVocoder. *https://github.com/kan-bayashi/PytorchWaveNetVocoder*