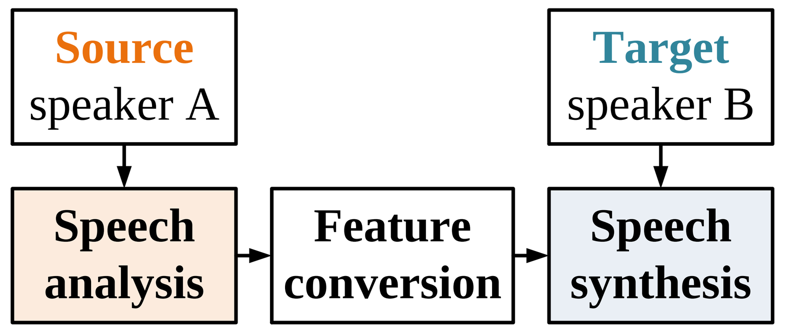
Statistical voice conversion with Quasi-Periodic WaveNet vocoder

Yi-Chiao Wu, Patrick Lumban Tobing, Tomoki Hayashi, Kazuhiro Kobayashi, Tomoki Toda

Nagoya University, Japan

Voice conversion



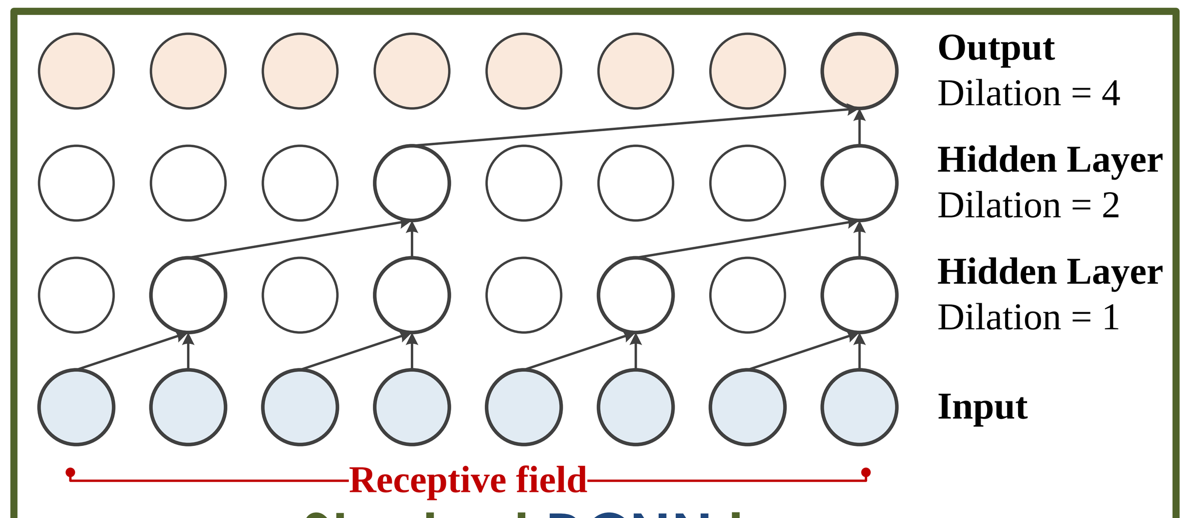
- VC: convert the speaker identity of speech while maintaining the same linguistic content
- Vocoder (analysis): encode speech into spectral and prosodic features
- Vocoder (synthesis): decode acoustic features to speech waveform
- Neural-Vocoder: replace the synthesizer of a conventional vocoder by an Neural-based speech generative model (ex: WaveNet, SampleRNN)
- Input of Neural-Vocoder: acoustic features
- Output of Neural-Vocoder: speech waveform

WaveNet

- Auto-regressive causal model
- Directly model the probability

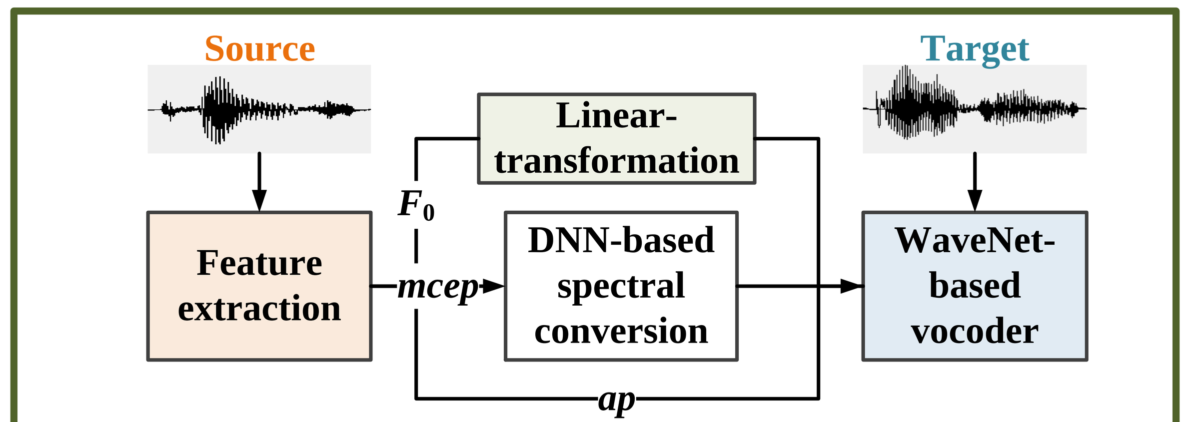
$$P(y_n \mid y_{n-r}, \dots, y_{n-1}, \mathbf{h})$$

- Conditioned on acoustic features $\mathbf{h}$
- Receptive field: previous samples y_{n-r}
- Dilated convolution (DCNN) layers efficiently extend the receptive field



Stacked DCNN layers

- Conditioned on the DNN-converted *mcep*, linearly transformed F_0, and source *ap* to generate target speech



Baseline VC system

- Inefficient speech modeling
  - huge network for long receptive field to cover all related samples
  - speech is a quasi-periodic signal ➔ the receptive field includes lots of redundant samples
- lack pitch-controllability
  - difficult to generate speech with accurate pitch while conditioned on the unseen F_0-*mcep* pair or F_0 not observed in the training data

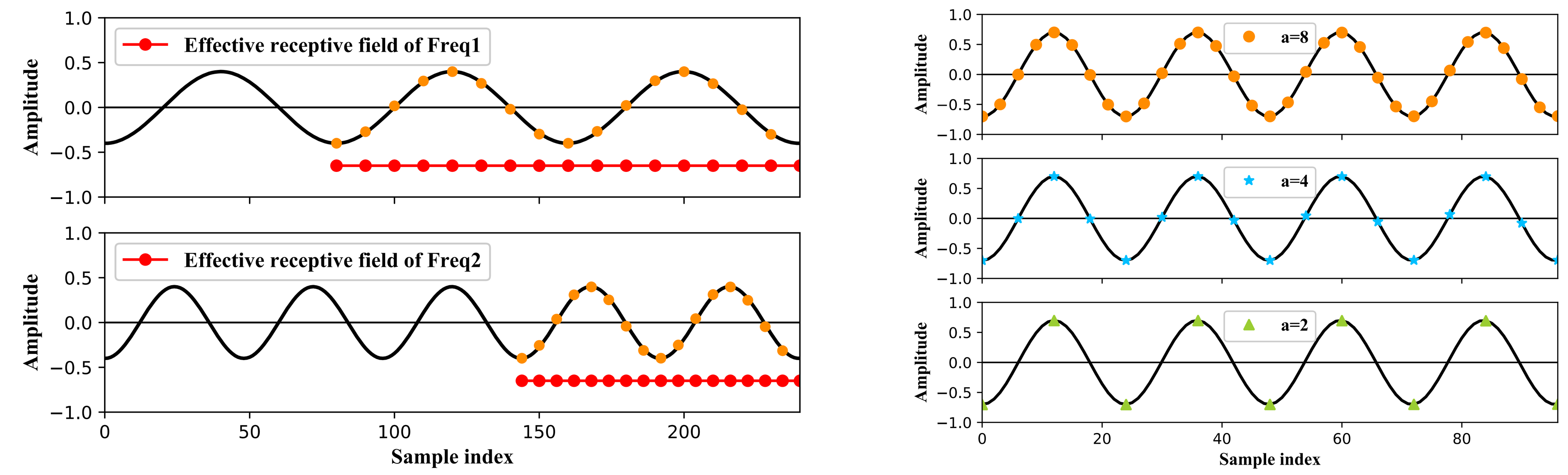
Proposed QPNet

Motivation

- Modeling the periodic part of speech with prior F_0 knowledge (long-term)
- Modeling the non-periodic part of speech with nearest samples (short-term)

Pitch-dependent dilated convolution (PDCNN)

- Number of samples in a receptive field is determined by the network size
- Effective receptive field can be changed by different dilation size
- Dilation size is dynamically changed according to the pitch
- Pitch-dependent dilated factor: $E_t = F_s / (F_{0,t} \times a)$

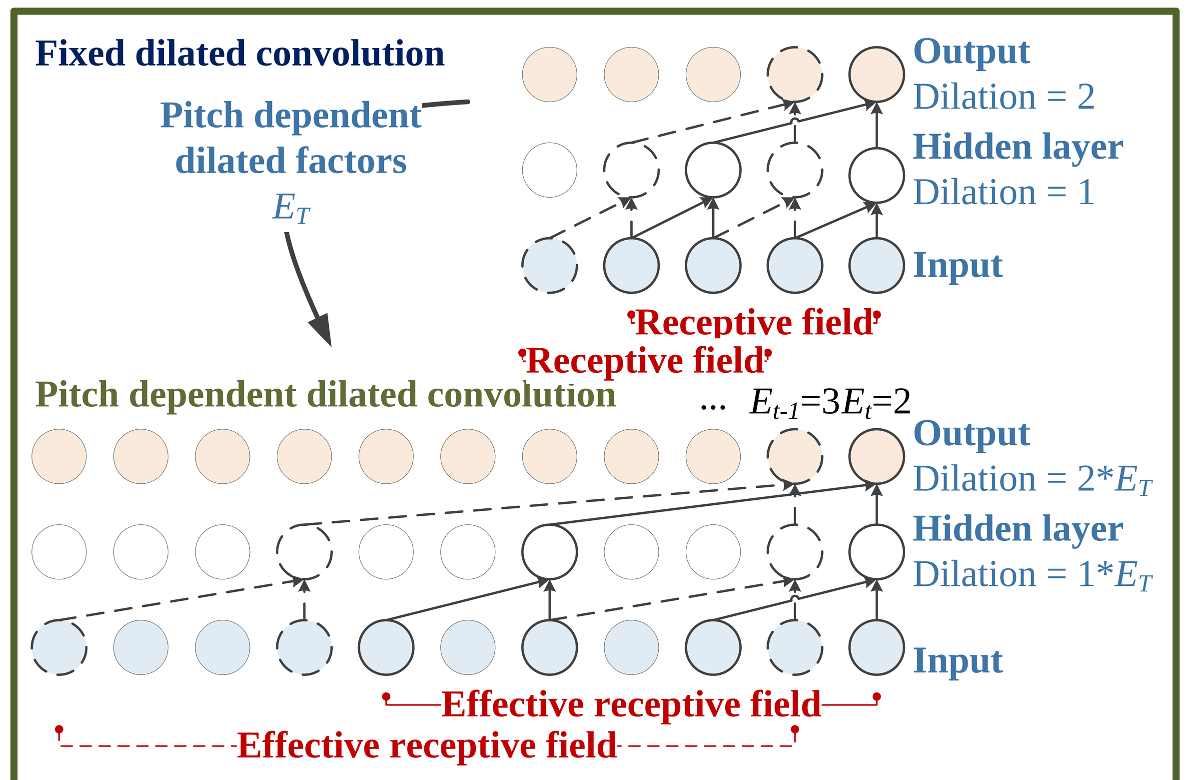


Cascaded autoregressive networks

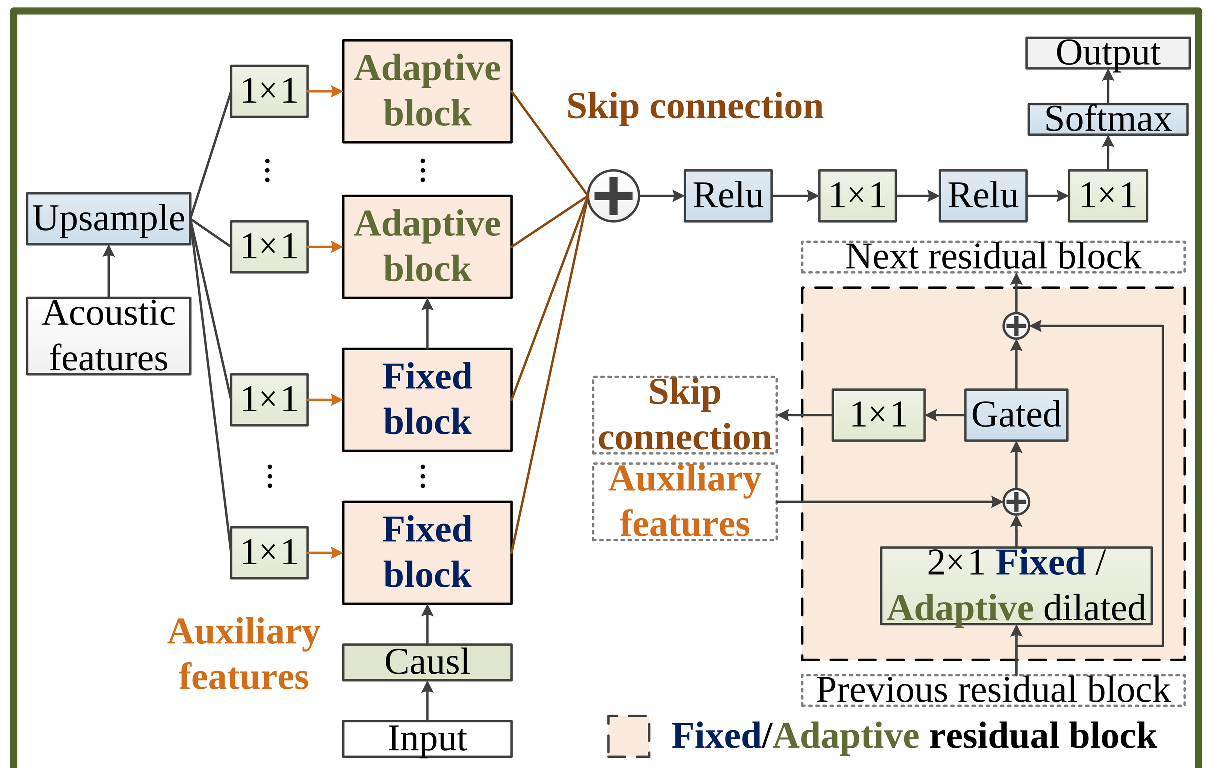
- Fixed modules (w/ DCNN) for short-term correlations
- Adaptive modules (w/ PDCNN) for long-term correlations

Speaker adaptation

- SDo: only update the output layers of the networks
- SDa: update the whole networks



Stacked PDCNN layers



QPNet vocoder

Experimental Evaluations

- Corpus for VC
  - SPOKE task of Voice Conversion Challenge 2018
  - 4 source speakers and 4 target speakers
  - 81 training utterances of each speaker
  - 35 testing utterances of each source speaker
- Corpus for Neural-Vocoder
  - Multi-speaker (SI) models: training data of "bdl" and "slt" from CMU-ARCTIC (1132 utts *2) and all training data of VCC2018 (81 utts *12)
  - Speaker-adapted (SD) models: 81 utts for each target speaker adaptation
- Objective evaluation
  - MCD for spectral prediction accuracy
  - RMSE of log F_0 for pitch prediction accuracy

		WN full	WN half	QPNet
MCD	SI	3.25	3.83	3.57
	SDo	3.11	3.73	3.51
	SDa	**3.02**	3.68	3.46
RMSE of log F_0	SI	0.15	0.21	0.15
	SDo	0.15	0.20	**0.13**
	SDa	0.15	0.19	0.14

- Subjective evaluation
  - MOS for speech quality (1:bad ~ 5:excellent)

	World	WN full	WN half	QPNet
SI	2.83 ± 0.10	2.72 ± 0.10	1.70 ± 0.07	2.83 ± 0.11
SDa	-	**3.26** ± 0.11	1.93 ± 0.07	**3.24** ± 0.11

  - Speaker similarity (same as real target or not)

	SDa-WN full	SDa-WN half	SDa-QPNet
Same	60.3 ± 6.5	44.4 ± 8.1	**61.9** ± 10.6
Different	39.7 ± 6.5	55.6 ± 8.1	38.1 ± 10.6

Conclusions

- Combined with DNN-based VC, QPNet vocoder achieves comparable speaker similarity and speech quality to WaveNet vocoder with only half the network size

VCC2018 QPNet Demo