Speaker-dependent WaveNet-based Delay-free ADPCM Speech Coding

Takenori Yoshimura[1,2], Kei Hashimoto[1], Keiichiro Oura[1], Yoshihiko Nankaku[1], and Keiichi Tokuda[1]
[1]Nagoya Institute of Technology, Department of Computer Science, Nagoya, Japan [2]Nagoya University, Institute of Innovation for Future Society, Nagoya, Japan

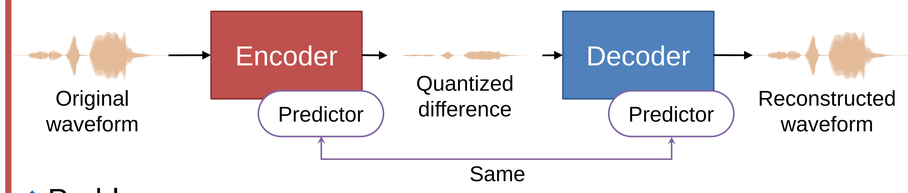
1. Introduction

- ADPCM speech coding
  - Widely used lossy (irreversible) coding technique
  - ITU-T G.726 [ITU-T; '90] provides a standardized ADPCM system
- ADPCM using neural network in 1990's
  - Effectiveness was shown [Haykin+; '93, Bartolini+; '93, Fundez-Zanuy+; '98]
  - Very simple architecture (one-hidden-layer with tens of units)
- ADPCM using WaveNet
  - WaveNet-based adaptive predictor
    - WaveNet [van den Oord+; '16] is a widely used waveform model
  - Mel-cepstrum-based noise shaping filter and postfilter
    - Mask noise using adaptive mel-cepstral analysis [Fukada+; '92]

High-quality digital communications without algorithmic delay

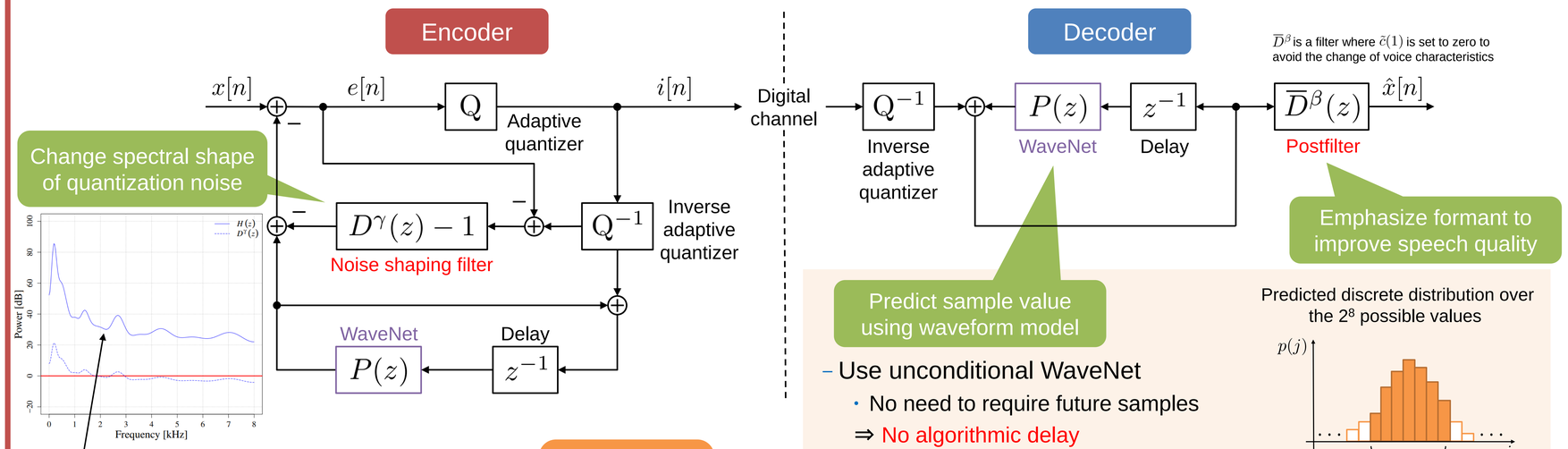
2. ADPCM

- Overview
  - Goal: transmit samples with high quality, low bit rate, and low latency
  - Assume: sample can be roughly predicted from its past samples
  - Idea: difference between actual sample and predicted one is quantized and transmitted instead of raw sample



- Problems
  - Poor predictor degrades the quality of reconstructed speech
  - The quantization introduces white quantization noise into speech

3. WaveNet-based delay-free APDCM speech coding



$$H(z) = \exp \sum_{m=0}^{M} \tilde{c}(m) \tilde{z}^{-m} = K \cdot D(z)$$

Used for noise shaping filter and postfilter

Spectral envelope | Mel-cepstrum | Gain factor

- Use adaptive mel-cepstral analysis for estimating mel-cepstrum
  - Sample-by-sample computation
  - ⇒ Can mask quantization noise in a time-variant manner
  - No need to require future samples for the computation
  - ⇒ No algorithmic delay

- Use unconditional WaveNet
  - No need to require future samples
  - ⇒ No algorithmic delay
  - Need to convert distribution to scalar
  - ⇒ Trimmed mean

Predicted discrete distribution over the 2^8 possible values

Convert index to sample value & calculate weighted mean

- Mel-cepstrum-based noise shaping and postfiltering [Tokuda+; '94]
  - Spectrum represented by mel-cepstrum can fit with the characteristics of human auditory system
  - Z-transform of reconstructed sample is:

$$\hat{X}(z) = X(z) + Q(z) \Rightarrow \hat{X}(z) = \{X(z) + D^{\gamma}(z)Q(z)\}\overline{D}^{\beta}(z)$$

Reconstructed sample | Original sample | Quantization noise | Noise shaping | Post-filtering

4. Evaluation

- Experimental setup

Database	CMU ARCTIC
# of training sentences	1092
# of test sentences	40
Sampling rate	8 kHz
# of bits	8 (μ-law compression)

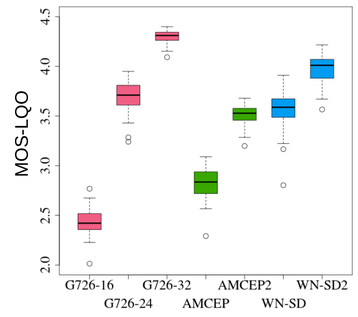
Dilations	[1, 2, 4, 8, …, 512] × 2
# of channels	32
Output layer	256-way softmax
Optimizer	Adam
Local/global conditions	N/A

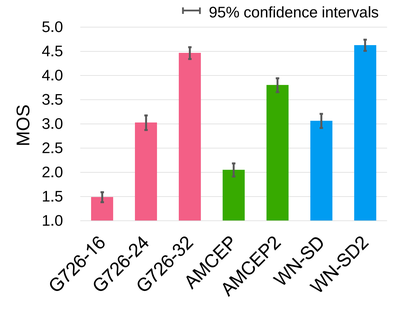
# of dimensions of mel-cepstrum M	12
Frequency warping parameter α	0.31
Postfiltering parameter β	0.3
Noise shaping parameter γ	0.3
Parameter of truncated mean B	200
Forgetting factor λ	0.98
Momentum τ	0.95

System	Bit rate [kbps]	Adaptive predictor	Speaker dependency	Noise shaping & postfiltering
G726-16	16	G.726		
G726-24	24	G.726		
G726-32	32	G.726		
AMCEP	16	AMCEP		
AMCEP2	16	AMCEP		✔
WN-SD	16	WaveNet	✔	
WN-SD2	16	WaveNet	✔	✔

* G.726 uses an ARMA predictor which has two poles and six zeros
* AMCEP [Tokuda+; '94] uses a backward predictor based on adaptive mel-cepstral analysis

- Experimental results





Proposed system achieved 32-kbps speech quality through 16-kbps digital channel

5. Conclusion

- WaveNet-based delay-free ADPCM speech coding
  - WaveNet is used as the adaptive predictor in ADPCM
    - Unconditional WaveNet for delay-free coding
    - Mel-cepstrum-based noise shaping and postfiltering
  - Outperformed conventional ADPCM systems
- Future work
  - Evaluate speaker-independent WaveNet-based predictors
  - Investigate robustness against bit errors
  - Optimize whole APDCM system in an end-to-end manner