Collapsed speech segment detection and suppression for WaveNet vocoder

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

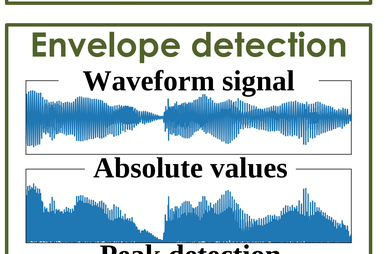
Nagoya University, Japan

Previous Work

- WaveNet : a deep autoregressive network capable of directly modeling speech waveform ＊1
- WaveNet vocoder : using acoustic features as auxiliary features to guide WaveNet generating speech samples ＊2

＊1. [A. van den Oord et al., 2016] ＊2. [Tamamori *et al.*, 2017]

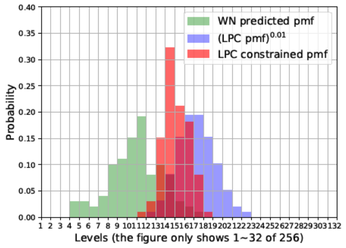
Problem



- For voice conversion (VC), speech generation using WaveNet conditioned on the converted features achieves better speech quality than WORLD vocoder
- Conditioned on converted features, WaveNet sometimes generates collapsed segments
  - Type-I: a collapsed segment has extremely large power at all frequencies like white-noise
  - Type-II: a collapsed segment has irregular short impulse
  - These collapsed segment are caused by the mismatch between training data (natural speech) and testing data (converted speech)
  - Training WaveNet vocoder using converted speech is not straightforward because of the limited parallel corpus for VC and the time alignments errors of converted-target speeches.

Proposed Method

LPC constrained pmf



- LPC pdf

$$f(y)=\frac{\exp(-(y-\mu)^2/2\sigma^2)}{\sigma\sqrt{2\pi}}$$

μ: LPC predicted value

σ^2: variance of LPC prediction error

Envelope detection

Collapsed speech suppression based on the LPC-constrained WaveNet vocoder

- Motivation: constrained the pmf of current sample using the relationship between current and past samples to prevent WaveNet from generating extremely non-speech like samples
- LPC coefficient ϕ: each sample is described by a linear combination of previous samples
- Constraint: (WaveNet predicted pmf)*(Gaussian(μ, σ^2))$^{\rho}$
  - μ: the previous samples*ϕ; σ^2: the variance of LPC prediction error; ρ: hyper parameter
- Problem: applying LPC-constraint causes over-smoothing side effects
- Using small ρ first to ease the over-smoothing issue, and then increasing ρ when WaveNet still has collapsed speech issue

Collapsed speech detection based on the difference of waveform envelopes

- Motivation: only applying LPC-constraint on collapsed speech segments
- Observation I: collapsed speech segments are easily detected by the waveform shapes
- Observation II: WORLD generated samples can be the reference waveform
- Criterion: envelope(WaveNet) - envelope(WORLD) > threshold ➔ collapsed speech detected

Envelope detection algorithm ＊3

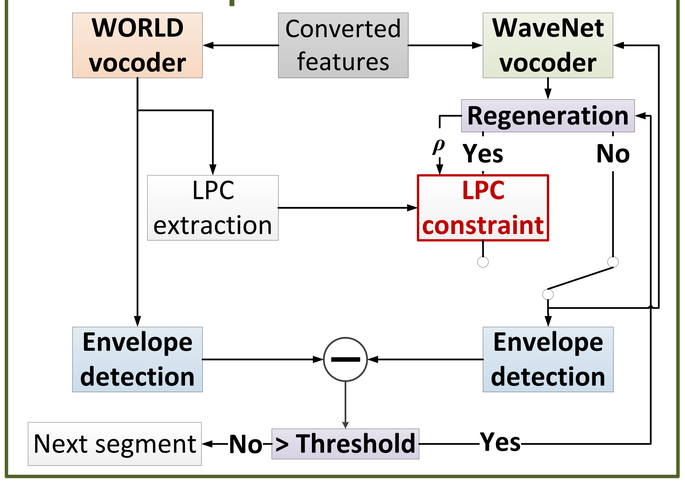
- Taking absolute values of waveform signal
- peak detection
- low-pass filtering

＊3. [Jarne, 2017]

Hyper parameters

- 30 dimensions of LPC coef.
- ρ =0.01➔0.1➔1
- 4000 samples/segment (~200ms)

Proposed flowchart



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 WaveNet vocoder
  - Multi-speaker WaveNet: using “bdl” and “slt” data from CMU-ARCTIC (1132 utts *2), and all training data from VCC2018 (81 utts *12).
  - Speaker-dependent WaveNet: using each target speaker’s training data to update the output layers of the multi-speaker WaveNet
- Collapsed speech detection evaluations
  - Statistical hypothesis test (verification)
- Collapsed speech detection criterions (between WaveNet and World samples)
  - maxPOW: the differences of maximum powers
  - maxMCD: the maximum MCDs
  - ENV: the differences of envelopes (proposed)

Verification Equal Error Rate			
Collapsed type	maxPOW	maxMCD	**ENV**
Type-I	10%	40%	**2%**
Type-II	40%	50%	**20%**

- Collapsed speech detection& suppression tests
  - Speech quality preference test
  - Speaker similarity test (4 measurements)

	w/ CL	w/o CL	p-value
Speech quality	**77%**	23%	1.09e-7
Speaker similarity	46%	48%	0.813
The same	11%	10%	0.810
Maybe the same	35%	38%	0.667
Maybe different	34%	33%	0.752
Different	20%	19%	0.963

Conclusions

- The proposed collapsed speech detection method achieves 20% equal error rate
- The proposed method gets 77 % preference in the quality test while keeping the same speaker similarity as the original WaveNet vocoder

VCC2018