**Objective: Non-parallel VC**

(Typical) parallel VC: Requires parallel utterances for training

(Our) non-parallel VC: Does not require parallel utterances

Baseline: CycleGAN-VC [Kaneko+2017]

1. CycleGAN [Zhu+2017]

   - Pros: Easy to learn
   - Cons: Hard to collect

   - Adversarial loss
   - Cycle-consistency loss
   - Gated CNN
   - IM Loss
   - Propagates information selectively
   - Represents sequential & hierarchical structures in speech
   - Forward & inverse mappings are simultaneously learned using adversarial loss and cycle-consistency loss

   - Finds optimal pseudo pair from non-parallel data

2. Gated CNN [Dauphin+2017]

3. IM Loss [Taigman+2017]

   - Identity-mapping loss
   - Preserves structure

   - Propagates linear unit (GLU)

   - Propagates structure

   - Preserves composition between input and output

Proposed: CycleGAN-VC2

1. Improved objective: Two-step adversarial losses

   - First adversarial loss
   - Second adversarial loss

   - Adversarial loss
   - Cycle-consistency loss

   - GLU

   - Inverse mapping

   - Forward mapping

   - Preserves structure

   - Translates dynamically

   - Preserves structure

2. Improved generator: 2-1-2D CNN

   - 2-1-2D CNN (Proposed)

   - 2D CNN

   - 1D CNN (Previous)

   - Adversarial loss

   - Cycle-consistency loss

   - Reshape

   - Upsample

   - Downsample

   - 1D CNN (Previous)

   - 2D CNN

   - 1x1Conv

   - ResBlocks

   - UpSample

   - DownSample

   - 1x1Conv

   - Reshape

   - 2-1-2D CNN (Proposed)

3. Improved discriminator: PatchGAN [Li+2016]

   - PatchGAN (Proposed)

   - FullGAN (Previous)

   - Requires many parameters

   - Makes training hard

   - Improves results for every speaker pair

Experiments

1. Experimental conditions

   i) Data

      - Dataset: Voice Conversion Challenge 2018 (Spoke (non-parallel) task)
      - Speakers: Professional US English speakers
      - Sentences: 81 sentences (about 5 min., relatively few for VC)
      - Sampling Rate: 22.05 kHz
      - Features: 54 MCEPs, log F0, APs (WORLD, 5 ms)

   ii) Conversion process (Follow VCC 2018 baseline)

      - Inter-gender: Vocoder-based VC
        - MCEP: CycleGAN-VC2
        - log F0: Linear transformation
        - AP: No conversion

      - Intra-gender: Vocoder-free VC [Kobayashi+2016]
        - DiffMCEP: CycleGAN-VC2 ➔ Waveform conversion (MLSA filter)

   iii) Training

      - Does not use any extra data, modules, or time alignment procedure

2. Objective evaluation

   i) Mel-cepstral distortion (MCD): Global structural difference

      - In both metrics, smaller is better

   ii) Modulation spectra distance (MSD): Local structural difference

   iii) Training

   - See our paper for detailed ablation studies

3. Subjective evaluation (10 participants)

   i) MOS test on naturalness

   ii) XAB test on speaker similarity

   - Note: See our paper for detailed ablation studies

Listen to speech samples at:

Copyright © 2019 NTT corporation All Rights Reserved.