The NU non-parallel voice conversion system for the voice conversion challenge 2018

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

Nagoya University, Japan

Speaker Odyssey 2018

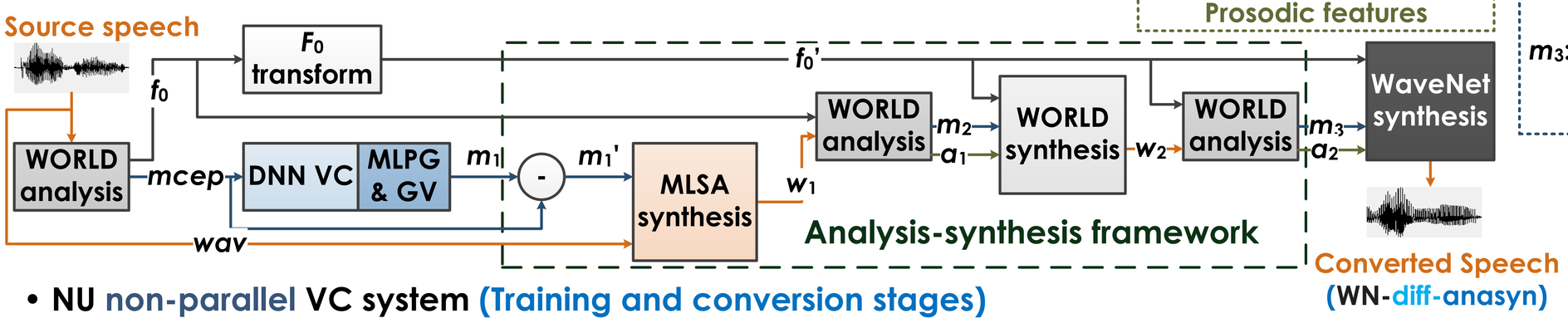
Introduction

- Conventional voice conversion (VC) usually needs a parallel corpus to train source-target mapping function
- Collecting parallel corpus is time consuming, expensive and inflexible
- Voice conversion challenge 2018 SPOKE task (Nonparallel VC)

VCC2018

System Framework

- NU parallel VC system
  - DNN VC + Analysis-synthesis framework + WaveNet vocoder

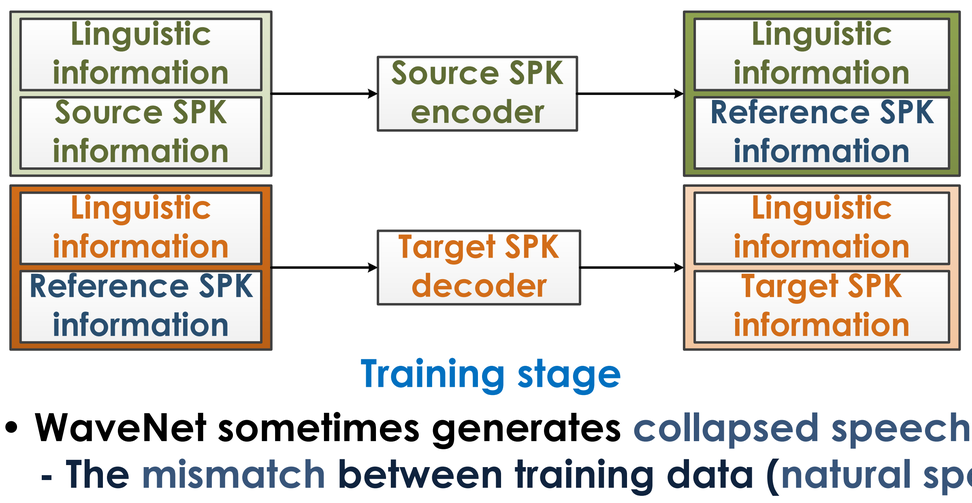


a_1: F_0 & *wav*-modified *ap*
a_2: analysis-synthesis *ap*
f_0: pitch
f_0': converted pitch
Prosodic features

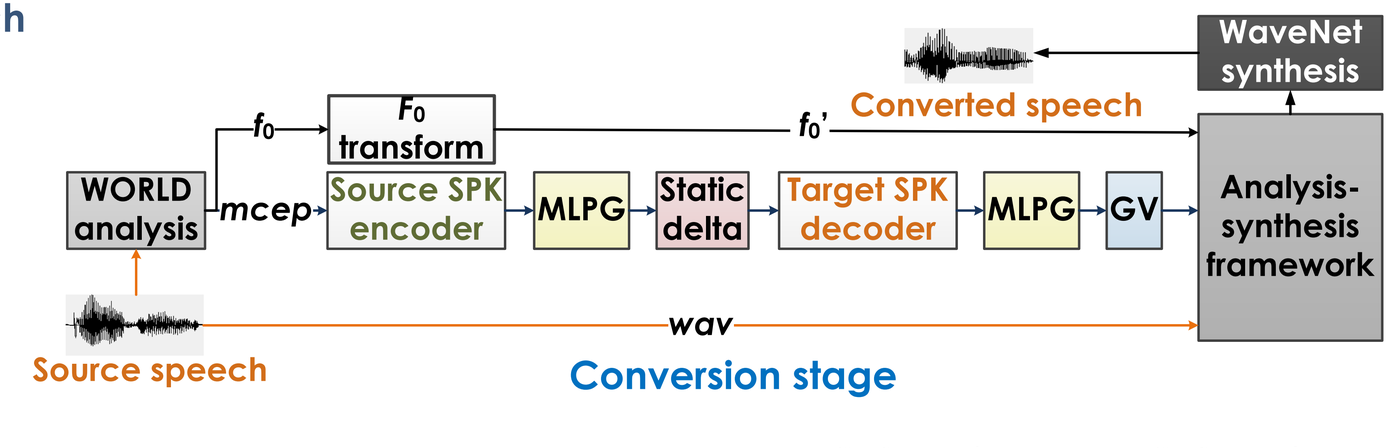
mcep: source mel-cepstral coef.
m_1: converted *mcep*
m_1': difference of source *mcep and* m_1
m_2: F_0 & *wav*-modified *mcep* (diff)
m_3: analysis-synthesis *mcep* (diff-anasyn)
Mel-cepstral features

wav: source speech waveform
w_1: *mcep* modified waveform
w_2: F_0 & *mcep*-modified waveform
Waveform signals

- NU non-parallel VC system (Training and conversion stages)
  - Cascade DNNs VC (source speaker ➔ reference speaker ➔ target speaker)
  - Use of TTS generated speech as VC reference speech

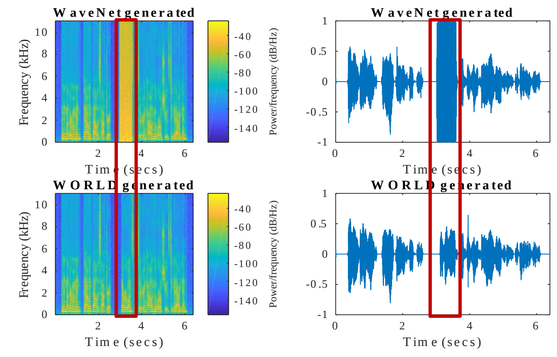


Training stage



Conversion stage

- WaveNet sometimes generates collapsed speech segments especially combined with VC
  - The mismatch between training data (natural speech) and testing data (converted speech)
  - Training of WaveNet vocoder using converted speech is not straightforward because of the limited parallel corpus
- System selection
  - Using WORLD-generated speech as detection reference speech
  - Detection criteria: The differences of maximum powers and Nyquist powers
  - Selection priority: WN-diff-anasyn > WN-diff-anasyn+LPC > WN-diff > WORLD-diff-anasyn



Collapsed speech segment

Experiments

- 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: "bdl" and "slt" speakers' data from CMU-ARCTIC (1132 utts *2), and all speakers' 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 *Fig. III*
  - Statistical hypothesis test (verification)
- Objective evaluations (internal) *Fig. I & II*
  - Only conducting on source speakers of SPOKE task
  - OtoO: One to one, parallel VC
  - MtoO: Many to one, non-parallel VC
  - wRTTS: VC with TTS ref. speech, non-parallel VC
  - wRspk: VC with natural ref. speech, non-parallel VC
- Subjective evaluations (from VCC2018) *Fig. IV*
  - Mean opinion score of speech quality
  - Speaker similarity test (the same, maybe the same, maybe different, different)

MCD [dB]				
	Source	OtoO	wRTTS	wRspk
F - F	8.27	**5.37**	5.54	5.73
F - M	8.46	**5.51**	5.66	5.67
M - F	8.46	**5.54**	5.68	5.67
M - M	7.89	**5.44**	5.65	5.63
Avg.	8.33	**5.48**	5.64	5.67

Fig. I: Parallel v.s. non-parallel (Quality of proposed method only slightly degrades than parallel VC)

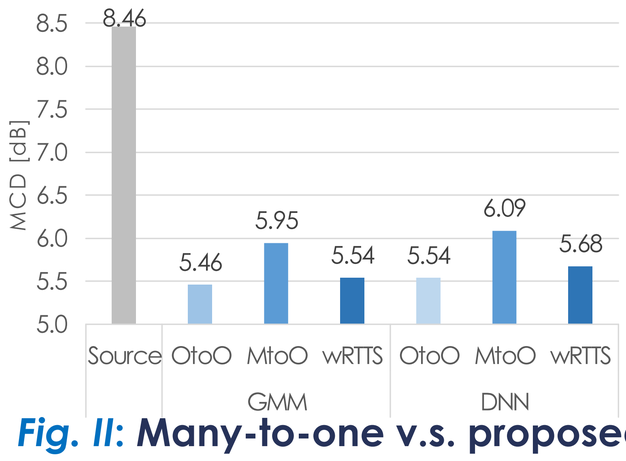


Fig. II: Many-to-one v.s. proposed (proposed method gets better quality than Many-to-one VC)

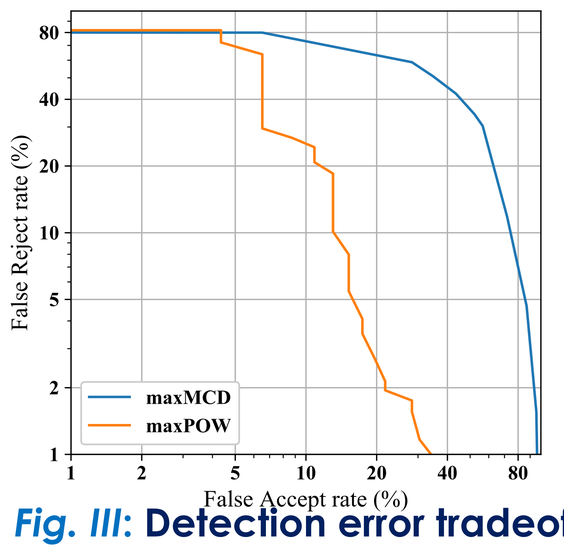


Fig. III: Detection error tradeoff (Detection performance)

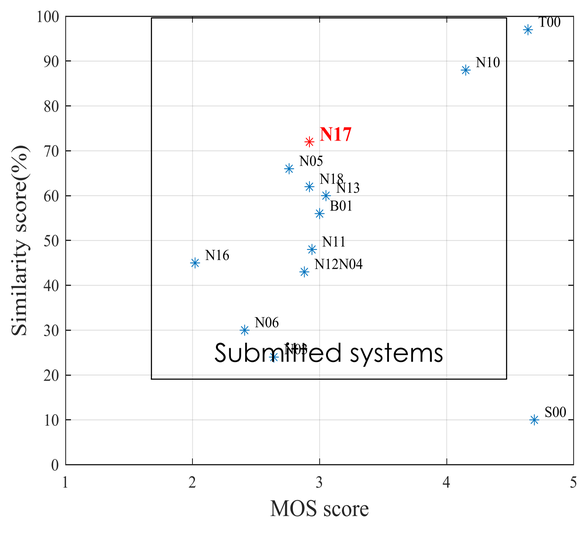


Fig. IV: VCC2018 results

Conclusions

- NU non-parallel VC system for VCC2018 achieves the above average performance in speech quality and the 2nd place in speaker similarity
- Objective results also show the effectiveness of the proposed non-parallel VC with reference speech
- The detected collapsed utterances are about 5% of all converted utterances

More details about WaveNet+LPC (INTERSPEECH 2018)