Emotional Voice Conversion using Multitask Learning with Text-to-speech

Tae-Ho Kim, Sungjae Cho, Shinkook Choi, Sejik Park, and Soo-Young Lee

{ktho22, sungjaecho, sk.c, sejik.park, sylee}@kaist.ac.kr

Introduction

- Emotional Voice Conversion using Multitask Learning with Text-to-speech

- Contributions
  - Voice Conversion using Multitask Learning with Text-to-speech
  - Emotional Voice Conversion

Introduction

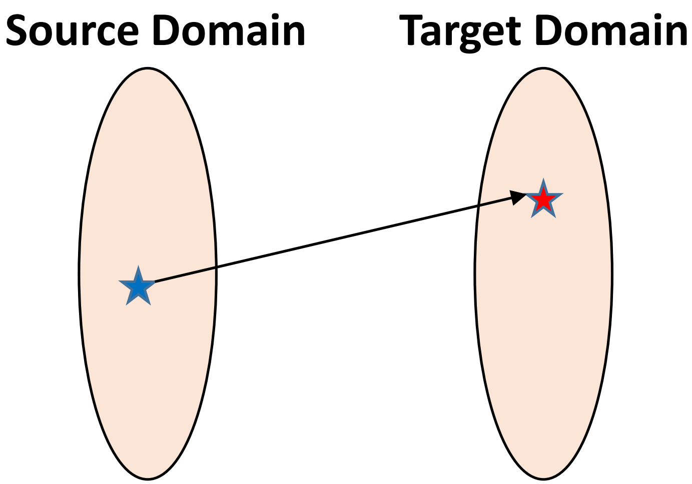
- Previous voice conversion uses frame-by-frame mapping with GMM, DNN, and RNN. [1, 2]
- Recently, voice conversion using seq2seq models has been proposed. [3]
- One drawback of current VC is lack of linguistic information.
- Linguistic information is additionally provided by auxillary classifier or complex attention modules. [4]
- In this paper, we propose 'Voice Conversion using Multitask Learning with Text-to-speech'

Introduction

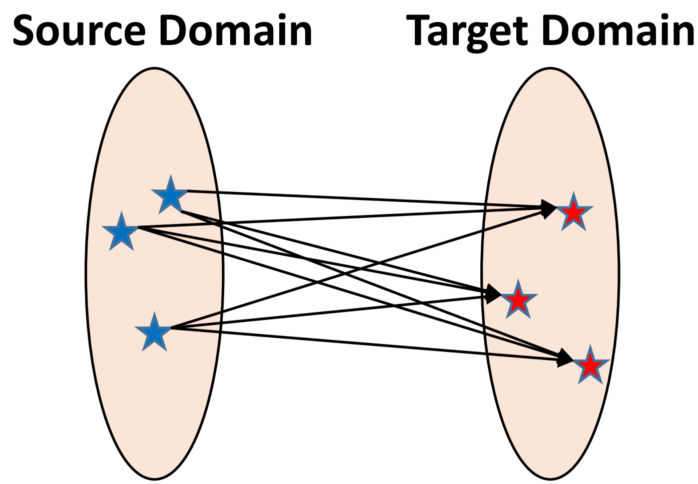
- Emotional Voice Conversion using Multitask Learning with Text-to-speech

- Contributions
  - Voice Conversion using Multitask Learning with Text-to-speech
  - Emotional Voice Conversion

Introduction



One-to-one VC



Many-to-many VC

Previous works mainly on One-to-one VC, especially for gender conversion

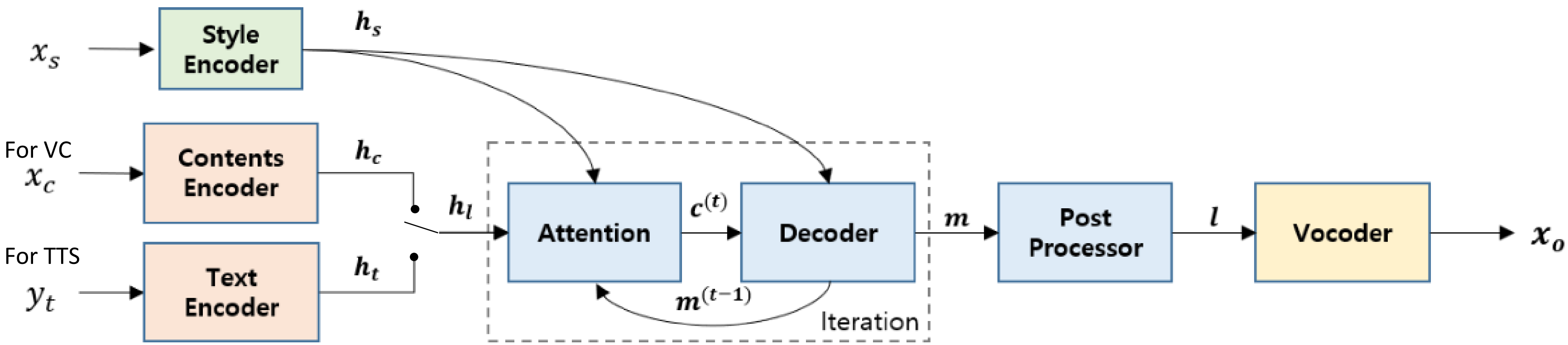
In this work, many-to-many VC on emotion conversion is proposed

Contributions

- Multitask learning with TTS could improve the performance of VC.
- Many-to-many emotional voice conversion was firstly conducted by a seq2seq model.
- A style reference speech could determine target domain of voice conversion

Emotional Voice Conversion using Multitask Learning with Text-to-speech

KI4AI
KAIST INSTITUTE



Main idea:

- without TTS path, VC can lose linguistic information.
- with TTS path, VC can capture linguistic information.

Training:

- For every batch, sample (x_s, x_c, x_o) or (x_s, y_t, x_o) with probability 0.5.
  (where $x_o$has same style with x_s, and same contents with x_c or y_t)

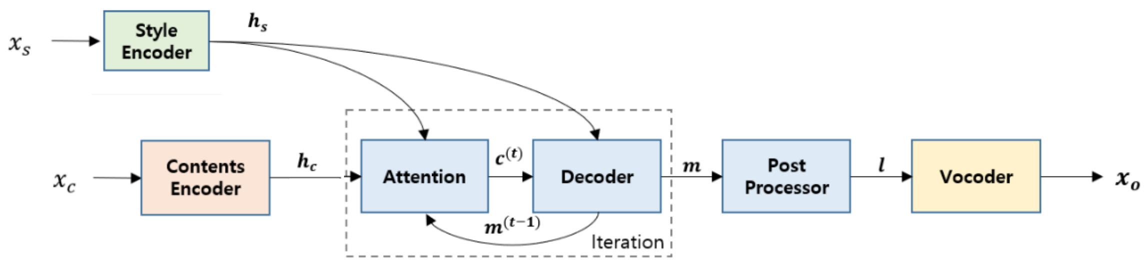
Experimental details

- Dataset: KETTS male
  - 7 emotions
  - 3,000 utterances per emotion
  - Across 3,000 utterances, same text set was used.
- Feature extraction
  - Downsampled to 16kHz
  - Silence removed using VAD[1)]
  - STFT with 50ms window and 12.5 shift.
  - 80 Log-mel spectrogram is used
  - Scaled to [0, 1]

1) https://github.com/wiseman/py-webrtcvad

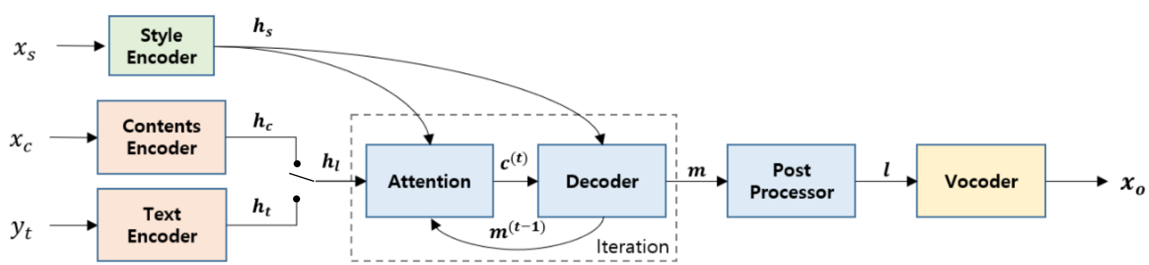
Model Comparison

VC



VCTTS-V

Infer with VC

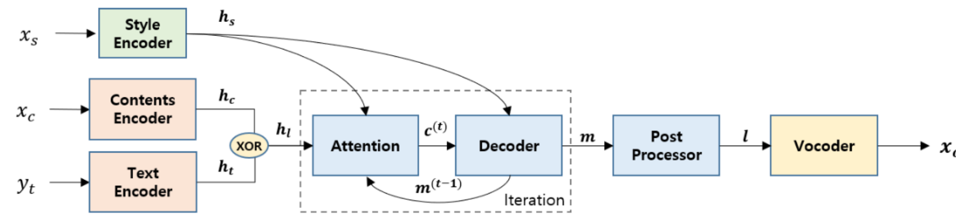


Experimental Results

Linguistic Consistency

x_s

(happy)



y_t	x_c	VC	VCTTS-V
방 안이 추우니 히터 좀 틀어주세요 bang an-i chuuni hiteo jom teul-eojuseyo			
어제보다 오늘, 더 너를 사랑해 eojeboda oneul, deo neoleul salanghae			
그 이야기는 할 필요가 없다. geu iyagineun hal pil-yoga eobda.			

Experimental Results

- Google ASR → Mecab POS Tagger → WER Measured

Table 1. Word error rate comparison

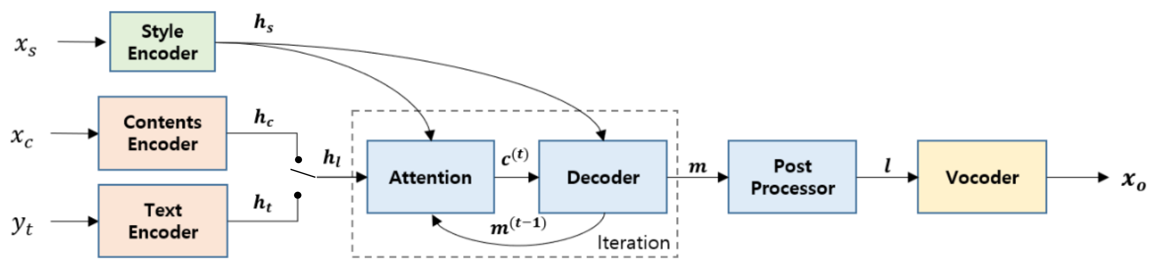
	VC	VCTTS-V	VCTTS-T	TTS	Train
WER	32.09	24.09	20.31	19.84	15.32

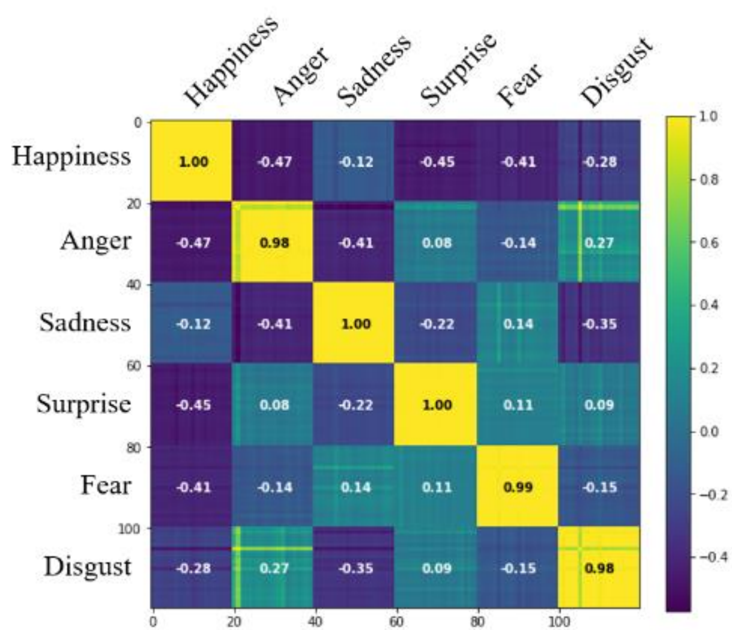
- Subjective Evaluation

Table 2. MOS and ABX preference score on clarity with 95% confidence intervals computed from the t-distribution

	VC	VCTTS-V
MOS	4.08 ± 0.17	4.54 ± 0.08
ABX	0.11 ± 0.06	0.55 ± 0.12

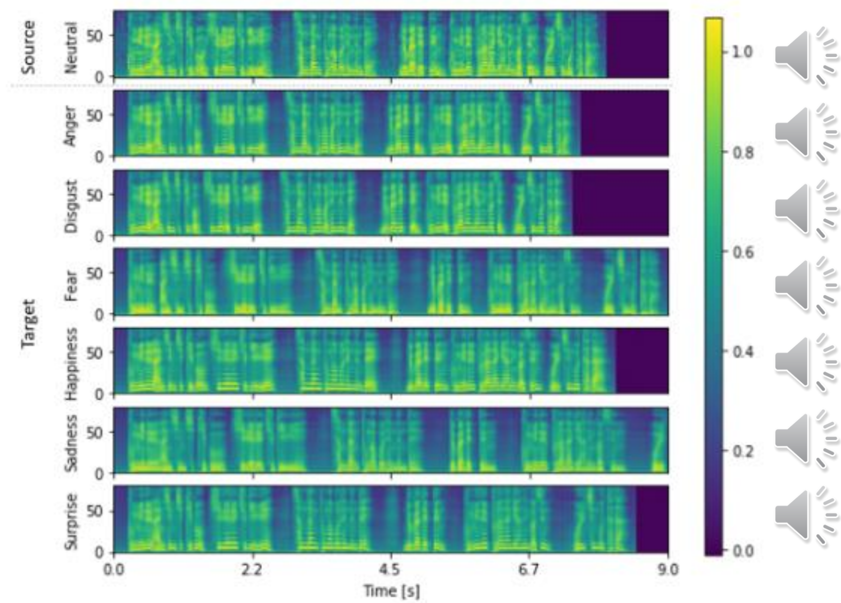
Experimental Results





h_s Similarity matrix

"나는 수업시간에 책을 읽는 척하면서 고개를 숙이고 잤다."
naneun sueobsigan-e chaeg-eul ilgneun cheoghamyeonseo gogaeleul sug-igo jassda



Voice conversion result

Conclusion

- We presented the emotional VC using multi-task learning with TTS.
- Although there have been abundant researches on VC, the performance of VC lacks in terms of preserving linguistic information, emotional information, and many-to-many VC.
- Unlike previous methods, the linguistic contents of VC were preserved by multitask learning with TTS.
- The results showed that using mul-titask learning significantly reduces the WER.

Future Work

KI4AI
KAIST INSTITUTE

- This research can be extended into many other directions.
- First, TTS can also be improved by the VC as some characterscan be pronounced differently.
- Second, the content encoder can make synergy with speech recognition as the content encoder was trained to extract linguistic information.
- Third, more explicit loss can be added to minimize the difference between the linguistic embedding of VC and TTS.

References

[1] Srinivas Desai, Alan W Black, B Yegnanarayana, andKishore Prahallad, "Spectral mapping using artificialneural networks for voice conversion,"IEEE Transac-tions on Audio, Speech, and Language Processing, vol.18, no. 5, pp. 954–964, 2010.

[2] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng,"Voice conversion using deep bidirectional long short-term memory based recurrent neural networks,"in ICASSP 2015. IEEE, 2015, pp. 4869–4873.

[3] Jing-Xuan Zhang, Zhen-Hua Ling, Li-Juan Liu, YuanJiang, and Li-Rong Dai, "Sequence-to-sequence acoustic modeling for voice conversion,"IEEE/ACM Trans-actions on Audio, Speech and Language Processing(TASLP), vol. 27, no. 3, pp. 631–644, 2019.

[4] Jing-Xuan Zhang, Zhen-Hua Ling, Yuan Jiang, Li-Juan Liu, Chen Liang, and Li-Rong Dai, "Improvingsequence-to-sequence voice conversion by adding text-supervision," inICASSP 2019. IEEE, 2019, pp. 6785–6789.