Non-parallel Many-to-many Voice Conversion by Knowledge Transfer from a Text-to-Speech Model

Xinyuan YU and Brian Mak

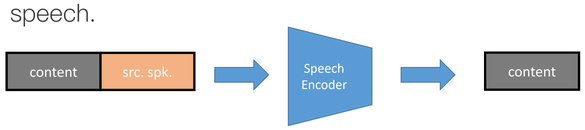
The Hong Kong University of Science and Technology
Department of Computer Science and Engineering

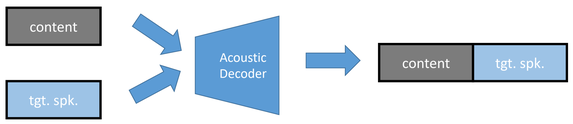
Introduction

- Typical non–parallel voice conversion models consist of an encoder and decoder.
- The encoder would disentangle speaker related information from content information in speech, and generate speaker–agnostic content representations.

$$H^{VC} = \{h_i^{VC}\}_{i=1}^{T_{speech}} = Enc^{VC}(\{s_i\}_{i=1}^{T_{speech}})$$

- The decoder then combines content representations with target speaker information to generate the converted speech.





- However, this turns out to be challenging as there is no supervision in the whole process.

Motivation

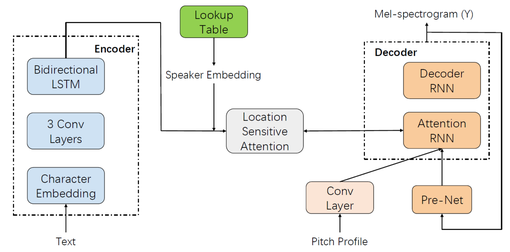
- Text–to–speech (TTS) models adopt similar encoder–decoder structures.
- TTS decoder is similar to VC decoder, which transforms hidden representations to acoustic features.
- TTS encoder output encodes linguistic information only, and is what the speech encoder of a VC model should learn

$$H^{TTS} = \{h_j^{TTS}\}_{j=1}^{T_{text}} = Enc^{TTS}(\{x_j\}_{j=1}^{T_{text}})$$

- The above suggests that there is room for transfer learning between the two tasks.

Proposed Method

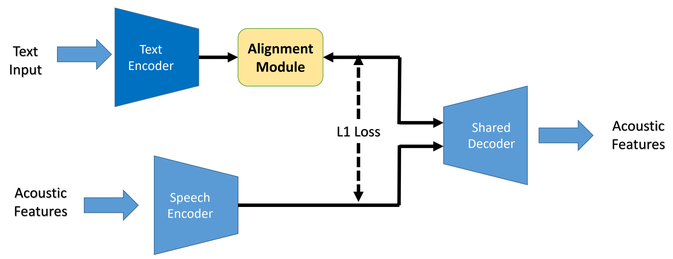
- Train a multi–speaker Tacotron2 TTS model
  Speaker identity is modeled by an embedding look up table.
- We condition the TTS decoder on pitch profile. This allows us to explicitly control the conversion of pitch profile for the VC model.



- Use TTS text encoder output as target for speech encoder of our VC model.
- Tacotron2 aligns text input and output speech frames
- We use this alignment matrix α to align the TTS encoder output and VC speech encoder output, which have corresponding relationship to the TTS input and output.

$$H^{aligned} = \alpha \cdot H^{TTS}$$

- In addition, we use TTS acoustic decoder as VC acoustic decoder
- The proposed model resembles an auto–encoder, as shown in the lower part of the figure below.

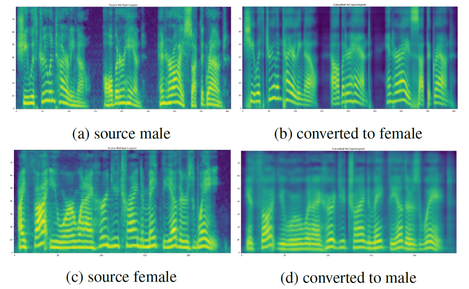


- At conversion stage, the TTS encoder and the alignment module are discarded.

Experimental Results

From conversion examples below, we can see that:

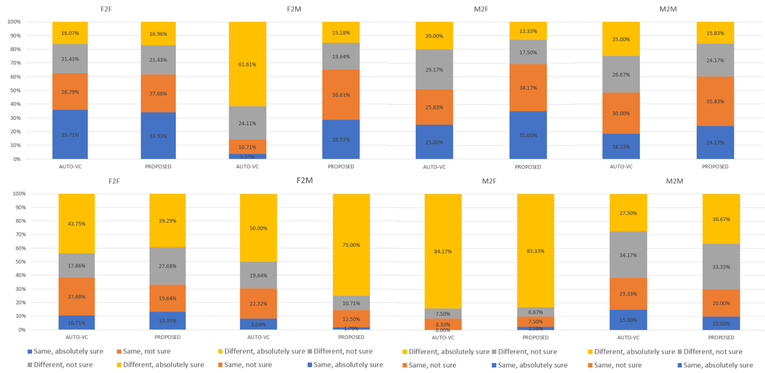
- The converted speeches keep most of the spectral structures in their original speech.
- Formants are shifted accordingly.



(a) source male (b) converted to female

(c) source female (d) converted to male

- Baseline: AutoVC.
- Speaker similarity test results compared to target speakers (up) and source speakers (down).
- The raters are asked if the converted speech is uttered by same/different speaker from the source /target speech.



Conclusions

- Using TTS text encoder output as VC speech encoder output targets can help disentangle speaker and content information.
- The acoustic decoder of a TTS model can be transferred to a VC model.
- The proposed method perform reasonably well on all conversion scenarios among many different speakers.