

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

IEEE ICASSP
June, 2021

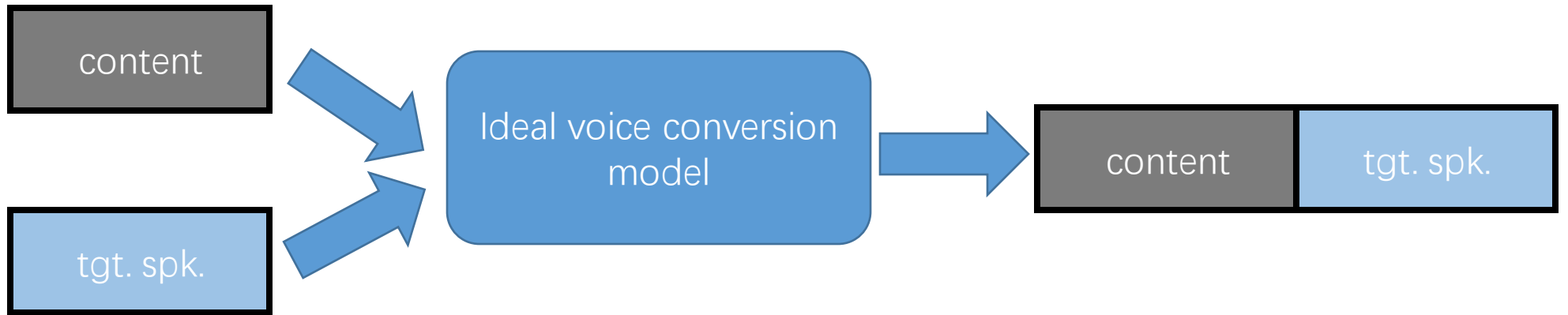
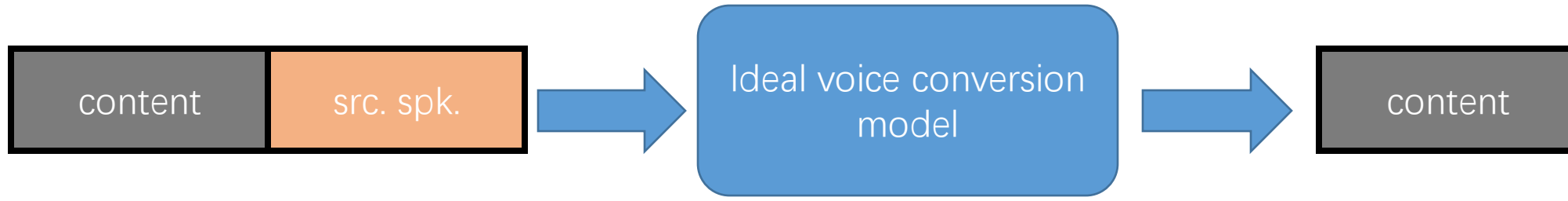
Introduction

Speech inherently carries different aspects of information.

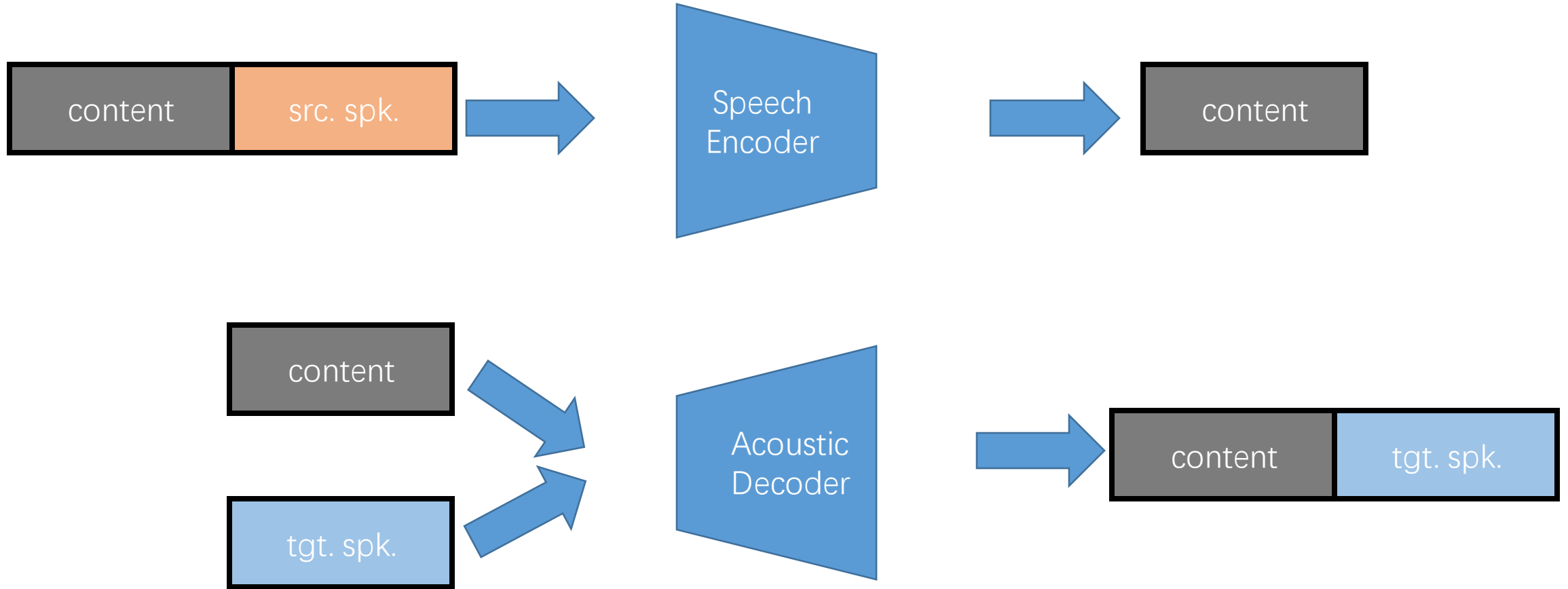


-  Content information: linguistic content of the speech.
-  Speaker information: other information that can be considered as speaker dependent.

Introduction

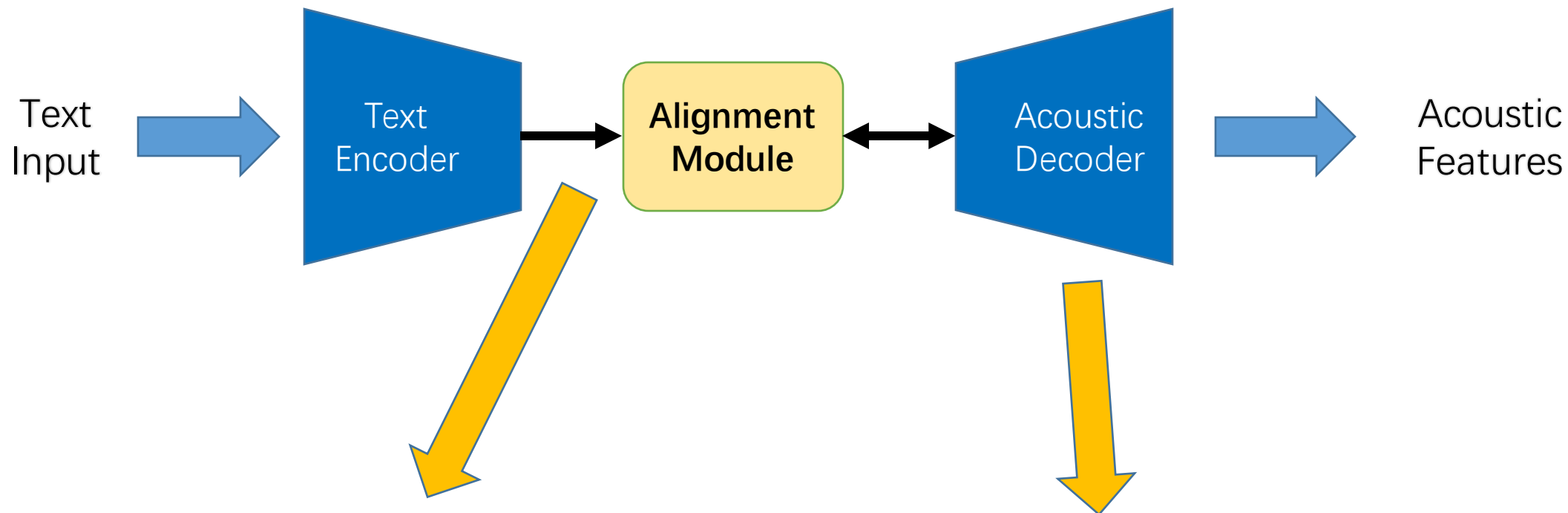


Introduction



This turns out to be challenging, as there is basically no supervision for the task.

Text-to-Speech Models



■ Content information: linguistic content of the speech.

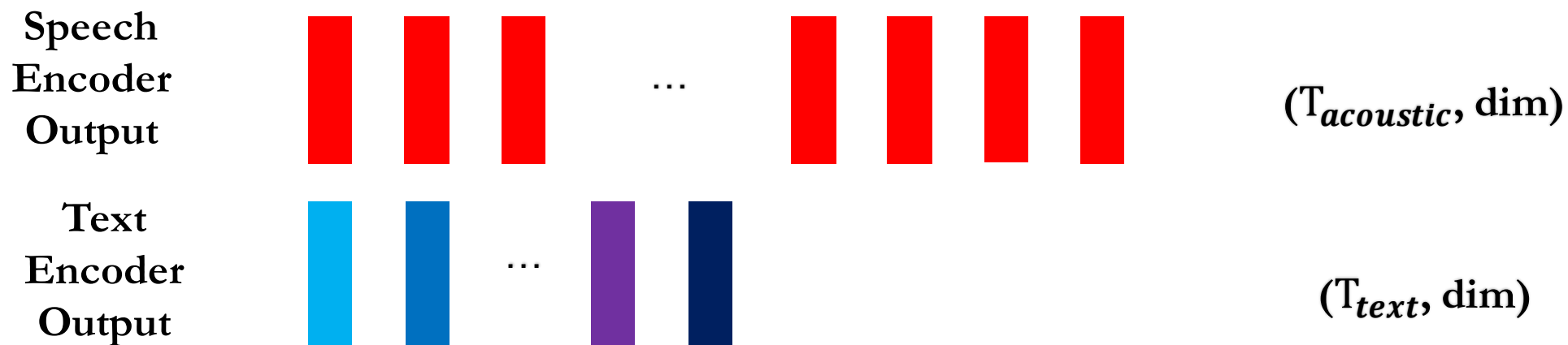
Acoustic decoder similar VC decoder.

Proposed Method

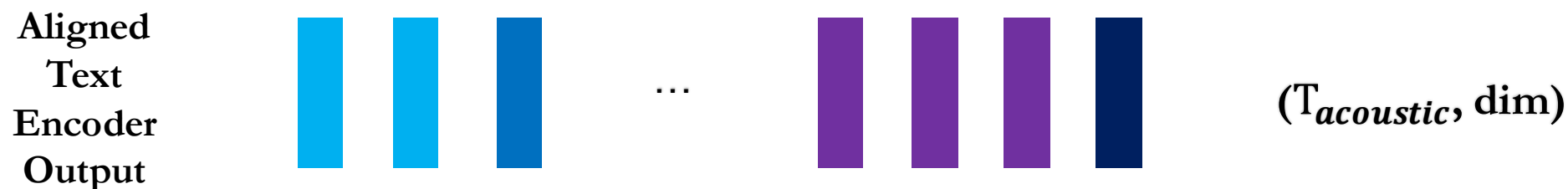
Use TTS text encoder output as target for speech encoder of our VC model.

Problem: text sequence and acoustic feature sequence typically have different lengths.

Solution: we use alignment matrix of the TTS model to align TTS encoder and VC encoder outputs, which have corresponding relationship to the TTS input and output.

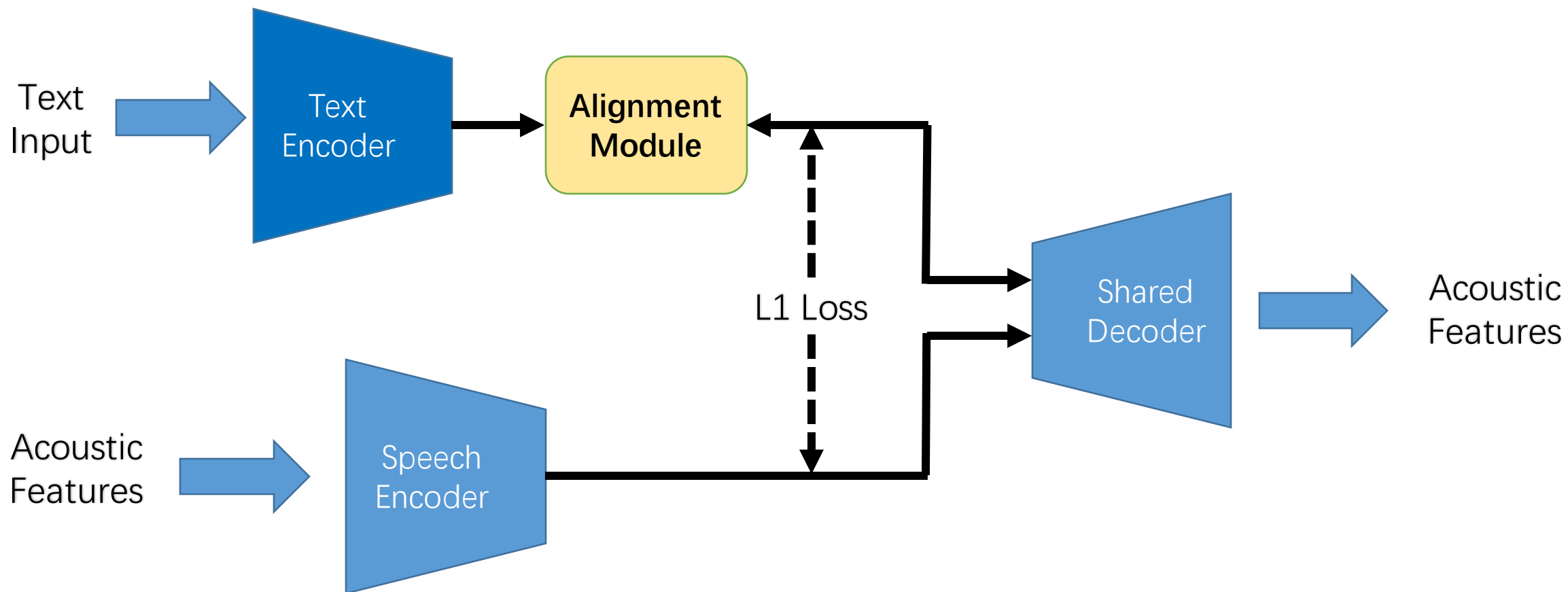


α is the alignment between text and acoustic features, which has shape $(T_{acoustic}, T_{text})$



Proposed Method

Use TTS model acoustic decoder as VC acoustic decoder.



Experiment and Results

All experiments are conducted using the train-clean-100 subset of LibriTTS.

To evaluate the performance of our model, we train AUTOVC model as our baseline.

We use melspectrogram as acoustic feature for both our TTS and VC models.

We use Waveglow as the universal vocoder to synthesize the final speech

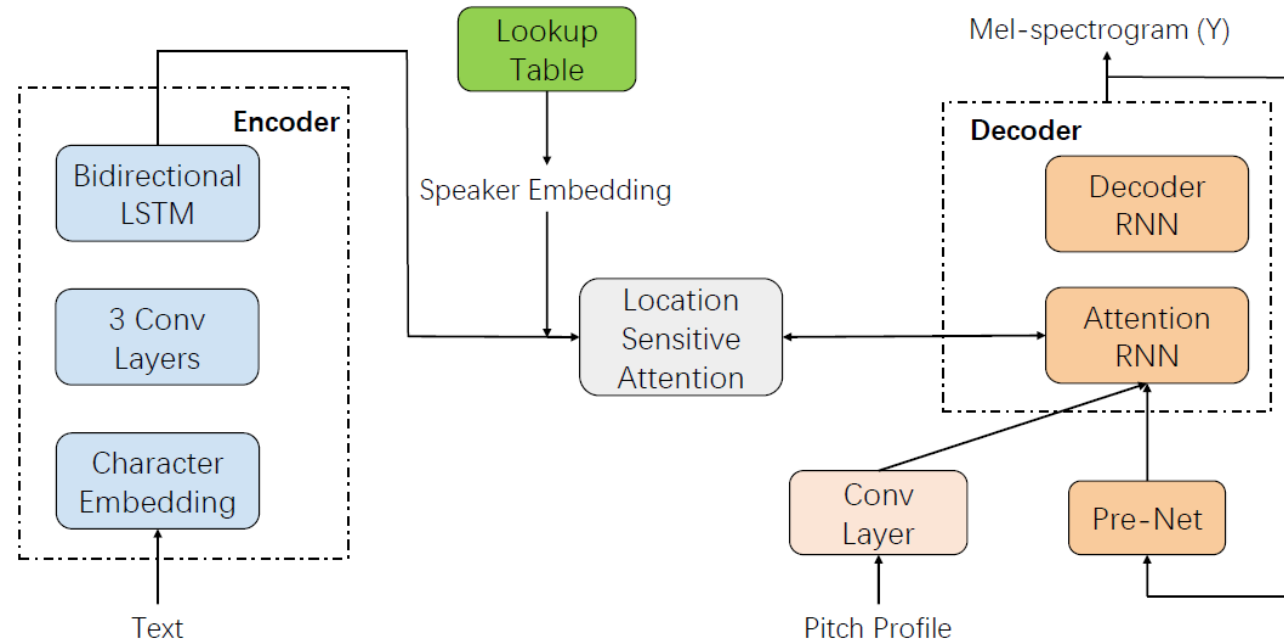
Experiment and Results

Train a multi-speaker Tacotron2 as our TTS model.

Condition the decoder of Tacotron2 on a pitch profile computed from the source speaker.

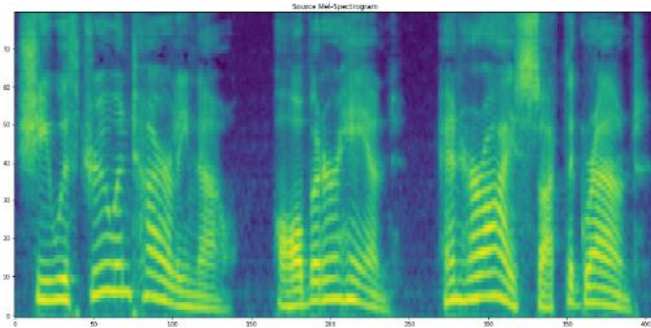
After training the TTS model, we train our VC model.

During conversion stage, scale the extracted pitch contour by the ratio between the mean pitches of the target and source speakers.

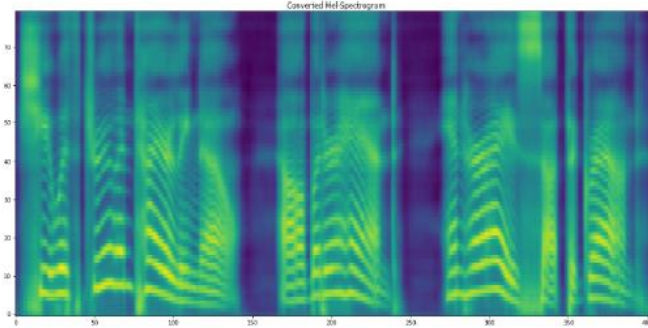


Experiment and Results

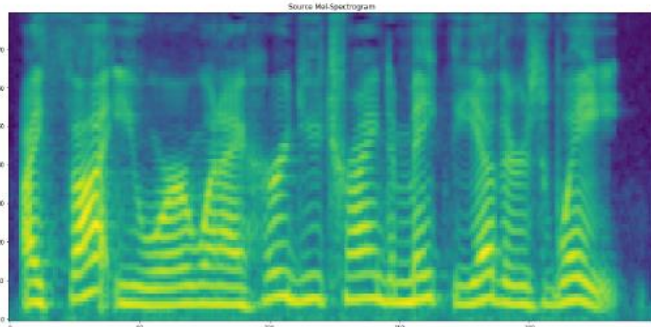
The converted speeches keep most of the spectral structures in their original speech.
Formants are shifted accordingly.



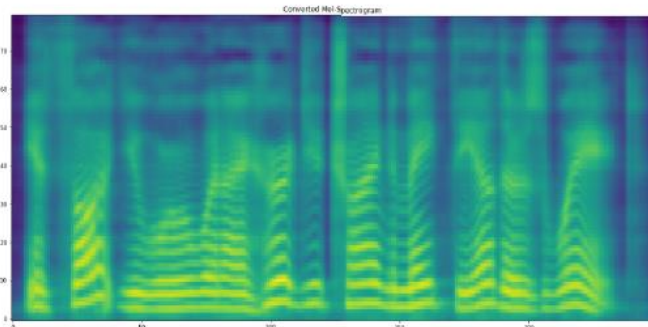
(a) source male



(b) converted to female



(c) source female



(d) converted to male



Experiments and Results

To evaluate our model, we conduct speaker similarity and speech naturalness tests.

We randomly select 4 speakers (2 male and 2 female) as the source speakers and another 4 different speakers (2 male and 2 female) as the target speakers.

For each speaker pair, we select 4 utterances of the source speakers from the test set and convert them to the voice of target speakers.

16 listeners (7 male, 9 female) are recruited to rate the utterances.

Experiments and Results

Speaker similarity test results.

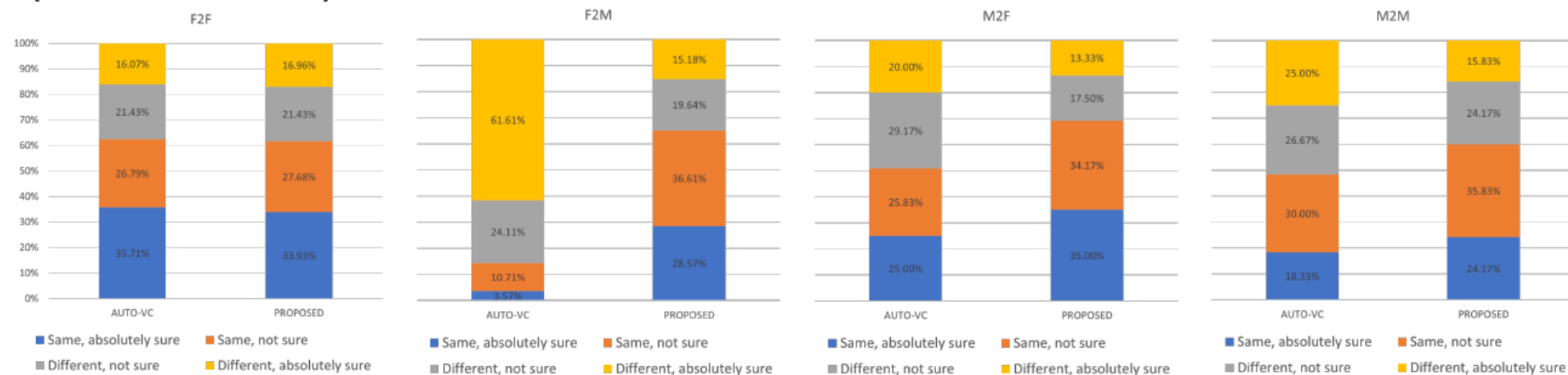


Fig. 3: Speaker similarity test result compared to target speakers.

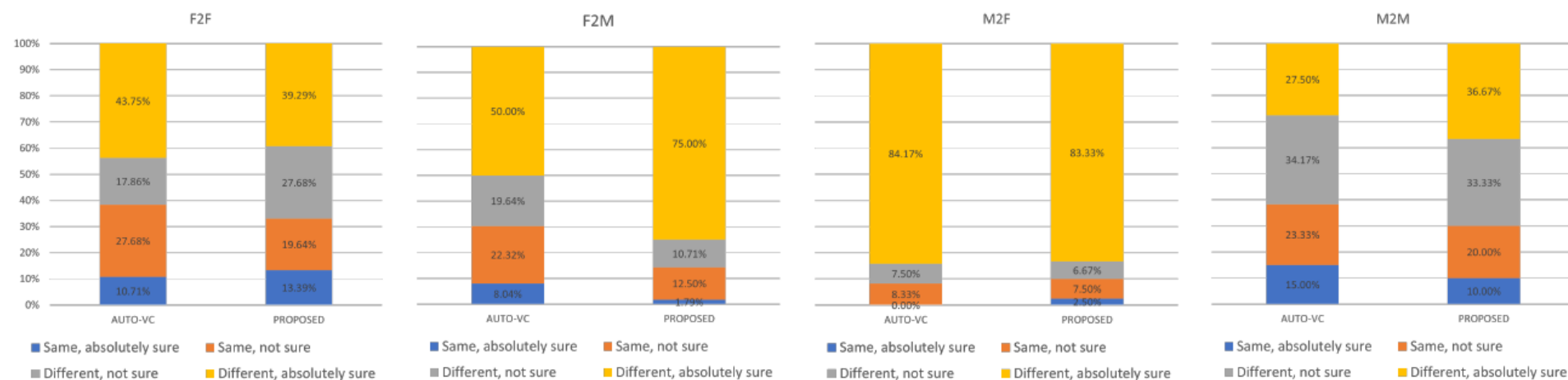


Fig. 4: Speaker similarity test result compared to source speakers.

Experiments and Results

Speech naturalness test results.

Table 1: MOS result for naturalness.

Model	MOS
AutoVC	3.813 (± 0.340)
Proposed Model	3.719 (± 0.315)
Ground Truth	4.377 (± 0.225)

Summary

- We show that using TTS text encoder output as VC speech encoder output targets can help disentangle speaker and content information.
- We demonstrate that the acoustic decoder of a TTS model can be transferred to a VC model.
- The proposed method perform reasonably well on same-gender and cross-gender voice conversion task among many different speakers.

Thanks for Listening