

Speech-to-Speech Translation without Text

Andros Tjandra¹, Sakriani Sakti^{1,2}, Satoshi Nakamura^{1,2}

1) Nara Institute of Science & Technology, Nara, Japan

2) RIKEN AIP, Japan

Outline

- Introduction
- Technical Background
- Training and Inference
- Experimental Setup & Results
- Conclusion

1. Introduction

- Speech-to-speech translation technology overcomes language barrier from human communication
- Challenges:
 - Training requires speech-text pairs (cascade ASR-NMT-TTS).
 - Jia et al. 2019 proposed direct speech-to-speech, however can't converge without pre-training with text.
 - Not all languages has written form.

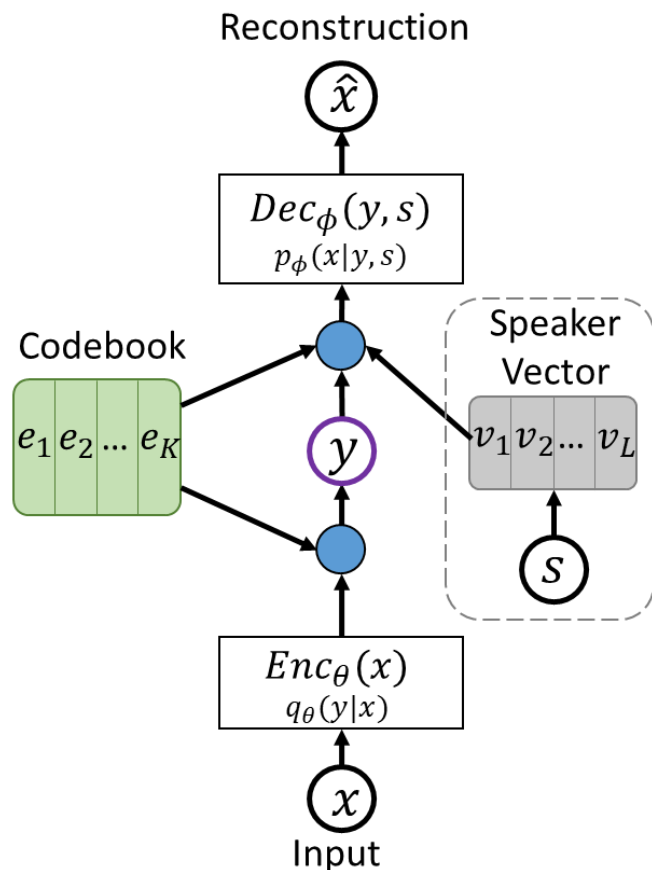
1. Our proposal ...

- Direct speech-to-speech translation for unknown languages (no prior knowledge about the language needed).
- No transcription needed for both source and target languages.

2. Technical Background

- We utilize 3 different models:
 - Unsupervised unit discovery with discrete autoencoder (VQ-VAE)
 - Sequence-to-sequence to translate audio to codebook
 - Codebook-to-spectrogram inverter to re-synthesize the translated audio

Unsupervised unit discovery with discrete autoencoder (VQ-VAE)



[de Oord et al., 2017]

Speech signal can be disentangled
into {**contexts**, speaking style}

$$Enc_\theta(x) = q_\theta(y|x)$$

$$Dec_\phi(y, s) = p_\phi(x|y, s)$$

$$\text{Codebook } E = [e_1, \dots, e_K]$$

$$\text{Speaker vec } V = [v_1, \dots, v_L]$$

Continuous speech (**harder target**)



Discrete symbol (**easier target**)

Training VQ-VAE

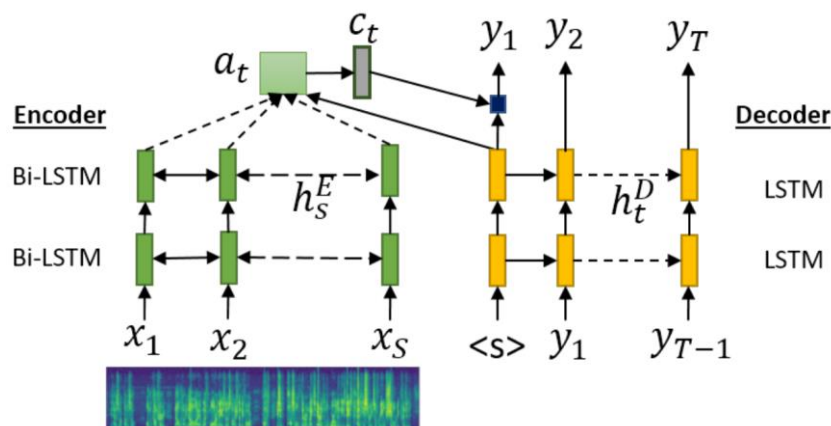
$$q_{\theta}(y = c|x) = \begin{cases} 1 & \text{if } c = \operatorname{argmin}_i \operatorname{Dist}(z, e_i) \\ 0 & \text{else} \end{cases}$$

$$e_c = \mathbb{E}_{q_{\theta}(y|c)}[\mathbf{E}]$$

$$= \sum_{i=1}^K q_{\theta}(y = i|x) e_i.$$

$$\mathcal{L}_{VQ} = \underbrace{-\log p_{\theta}(x|y, s)}_{\text{Reconstruction loss}} + \underbrace{\|sg(z) - e_c\|_2^2}_{\text{Embedding loss}} + \underbrace{\gamma \|z - sg(e_c)\|_2^2}_{\text{Commitement loss}}$$

Sequence-to-Sequence from Speech to Codebook



Input $X = [x_1, \dots, x_S]$ as the speech from source language

Output $Y = [y_1, \dots, y_T]$ as the codebook from target language

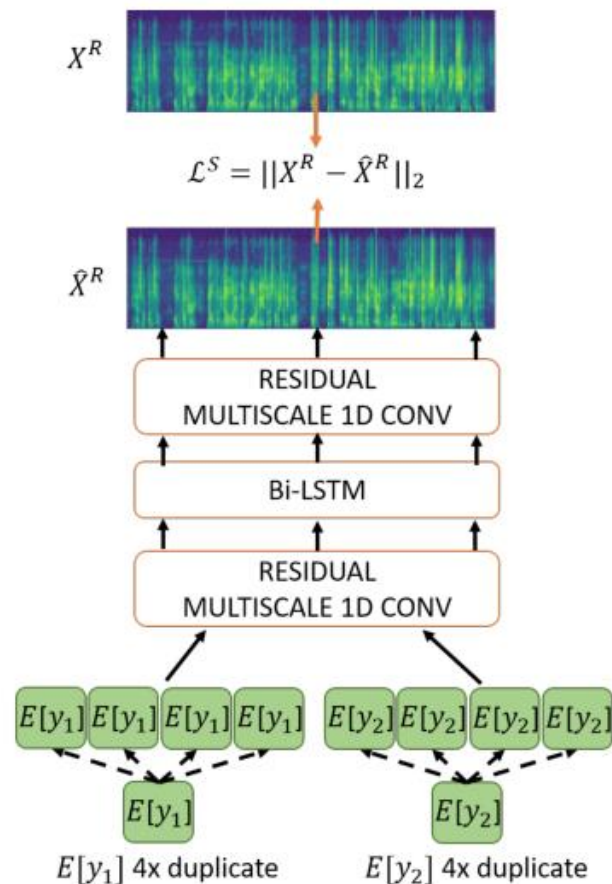
Encoder consisted of Bi-LSTMs and decoder consisted of LSTMs

Codebook Inverter

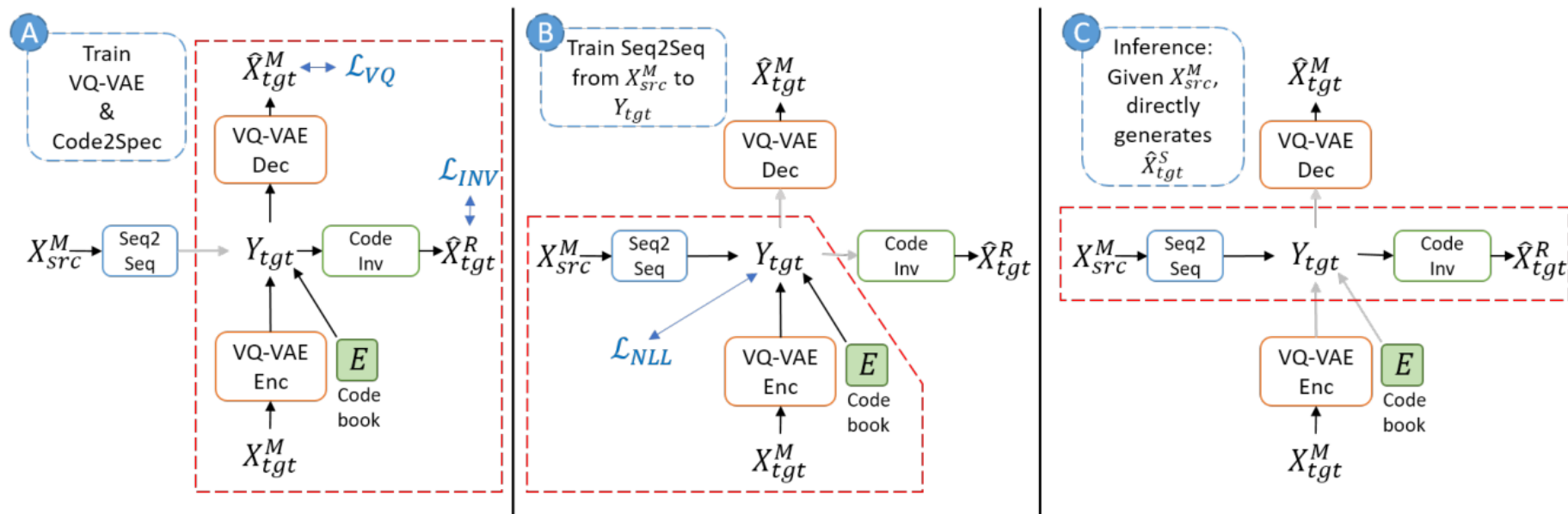
Input is codebook embedding $E_Y = [E[y_1], \dots, E[y_T]]$

Output is linear magnitude spectrogram

- We use Griffin-Lim to recover phase spectrogram and inverse Fourier transform to recover the waveform.



Training and inference



3. Experimental Setup

- Dataset: Basic Travel Expression (BTEC) corpus
- Language pairs:
 - French to English (similar grammatical structure)
 - Japanese to English (distant grammatical structure)
- Size:
 - Training 162.318 sentences pair
 - Test 510 sentences pair
- Speech features:
 - Input: MFCC (39 dimensions)
 - Output: Linear spectrogram (1025 dimensions)

Evaluation

- Because the large number of test samples, it is hard to do subjective evaluation.
- How we do evaluation:
 1. Train English ASR
 2. Translate source language speech to target language speech
 3. Use trained ASR (step 1) to recognize translated speech
 4. Calculate BLEU and Meteor between groundtruth and ASR transcription

Result

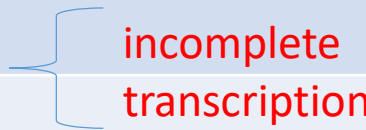
- Model:
 - **Baseline** (direct spectrogram-to-spectrogram)
 - **Proposed** SP2C (C=codebook size, T=time reduce)
 - **Topline** (speech src-> text tgt*-> speech tgt, *requires text transcription during training)

Model	BLEU4	METEOR
Baseline (FR-EN & JA-EN)	Not converged	
SP2C FR-EN C=64, T=12	25	23.2
Topline FR-EN (Cascade) *	47.4	41.2
SP2C JA-EN C=128, T=8	15.3	15.3
Topline JA-EN (Cascade) *	37.4	32.8

Additional result

- Translation samples : <https://sp2code-translation-v1.netlify.com/>

Model	Transcription
Groundtruth	how long are you going to stay
SP2C FR-EN	how long are you going to stay
SP2C JA-EN	how long will it take
Groundtruth	please tell him to call me as soon as he comes in
SP2C FR-EN	please tell him to call me back
SP2C JA-EN	please tell him that i called



incomplete transcription

Based on the example, 1) gives quite close result
However, 2) SP2C result left out the latter part

Conclusion

- We proposed a novel approach for training speech-to-speech translation w/o transcription
- Experiments was performed on French-English & Japanese-English

😊 Thank you for listening 😊