Speech-to-Speech Translation without Text

Andros Tjandra[1] Sakriani Sakti[1,2] Satoshi Nakamura[1,2]

[1] Nara Institute of Science and Technology, Japan

[2] RIKEN AIP, Japan

{andros.tjandra.ai6, ssakti, s-nakamura}@.is.naist.jp

1 Introduction

The common system requires effort to construct several components, including automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) synthesis, all of which are trained and tuned independently. Given speech input, ASR processes and transforms speech into text in the source language, MT transforms the source language text to corresponding text in the target language, and finally TTS generates speech from the text in the target language. Significant progress has been made and various commercial speech translation systems are already available for several language pairs. However, more than 6000 languages, spoken by 350 million people, have not been covered yet. At the moment, it is difficult to scale-up the existing approach to unknown languages without written forms or transcription data available.

In this paper, we take a step beyond the current framework and propose a method for training speech to speech translation tasks without any transcription or linguistic supervision. Instead of only discovering subword units and synthesizing them within a certain language, our approach discovers subword units that are directly translated to another language. Our proposed method consists of two steps: (1) we train and generate discrete representation with unsupervised term discovery, which is also based on vector quantized variational autoencoder (VQ-VAE) [4]; (2) we train a sequence-to-sequence model to directly map the source language speech to the target language discrete representation. Our proposed method can directly generate target speech without any auxiliary or pre-training steps with source or target transcription. To the best of our knowledge, this is the first work that performed pure speech-to-speech translation between untranscribed unknown languages.

2 Training and Inference

In this section, we explain our proposed method in detail and step-by-step. To train our proposed model, we setup three different modules: VQ-VAE, a speech-to-codebook seq2seq to generate target codebook given the source language speech, and a codebook inverter to generate target speech given the target language codebook [3]. Fig. 1 shows which modules are trained in each step. Initially, we defined $\{\mathbf{X}^M_{src}, \mathbf{X}^M_{tgt}\}$ as paired parallel speech, $\mathbf{X}^M_{src}$ is the MFCC (mel frequency cepstral coefficients) features from the source language, and $\mathbf{X}^M_{tgt}$ is the MFCC features from the target language. $\mathbf{Y}_{tgt}$ is the codebook sequences generated by VQ-VAE encoder $\text{Enc}_\theta(x)$ given $\mathbf{X}^M_{tgt}$ as the input. $\hat{\mathbf{X}}^R_{tgt}$ is the predicted linear spectrogram of the target language. $\mathcal{L}_{VQ}$ is the VQ-VAE loss, $\mathcal{L}_{INV}$ is the raw spectogram loss, and $\mathcal{L}_{NLL}$ is the negative log-likelihood of codebook in target language given the source speech $\mathbf{X}^M_{src}$.

1. First, we trained the VQ-VAE model on target language MFCC $\mathbf{X}^M_{tgt}$. We also trained the codebook inverter to predict corresponding linear spectrogram $\mathbf{X}^R_{tgt}$.

2. Second, we trained the seq2seq model from the source language speech to the target language codebook. Given a paired parallel MFCC from source and target languages $\{\mathbf{X}^M_{src}, \mathbf{X}^M_{tgt}\}$, we extracted codebook sequence $\mathbf{Y}_{tgt} = \text{Enc}^{VQ}_\theta(\mathbf{X}^M_{tgt})$ from the VQ-VAE encoder. Later, we trained

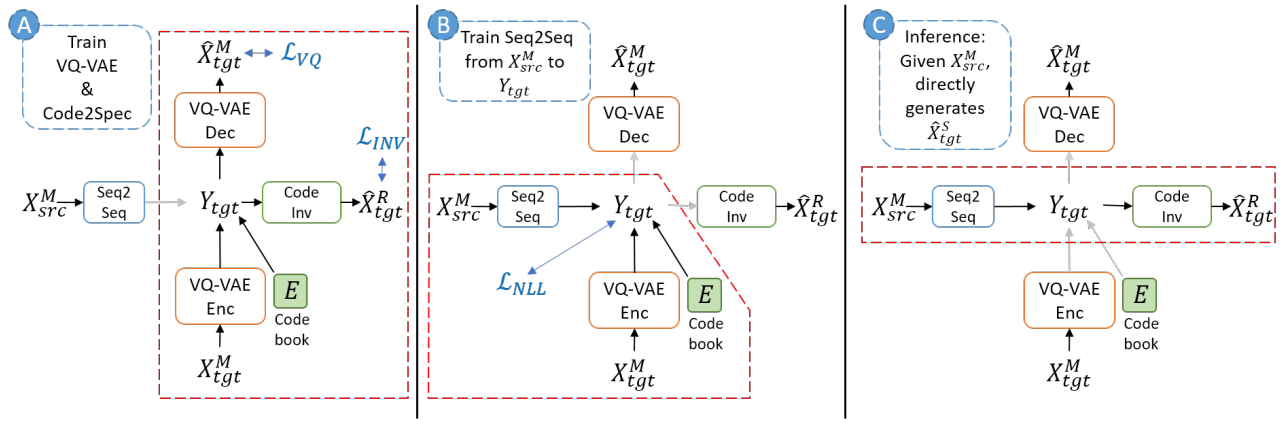

Figure 1: a) Train VQ-VAE to represent continuous MFCC vectors with codebook sequence and train codebook inverter to generate a linear magnitude spectrogram based on generated codebook sequence; b) Train a seq2seq model from source language MFCC to target language codebook. c) In inference stage, seq2seq model takes source language MFCC and predicts codebook sequences, and then codebook inverter generates target language speech representation.

the seq2seq translation model to predict $\hat{\mathbf{Y}}_{tgt} = \text{Seq2Seq}(\mathbf{X}^M_{src})$ and minimize loss $\mathcal{L}_{NLL}$ between $\mathbf{X}^M_{src}$ and X^M_{src}.

3. In the inference step, given source language speech $\mathbf{X}^M_{src}$, we decoded a target language codebook index sequence $\hat{\mathbf{Y}}_{tgt} = \text{Seq2Seq}_{\psi}(\mathbf{X}^M_{src})$ and synthesized it into target language speech $\hat{\mathbf{X}}^R_{tgt} = \text{Inverter}(\hat{\mathbf{Y}}_{tgt})$.

2.1 Evaluation

For an objective evaluation of the target speech utterances, currently there is no standard method can be used to measure translation quality directly on the speech utterances. Therefore, we utilized a pre-trained ASR on the English BTEC dataset and the generated transcription for our evaluation. For the ASR architecture, the encoder module has three stacked Bi-LSTMs with 512 hidden units, and the decoder has one LSTM with 512 hidden units. For the attention module, we utilized MLP attention with multiscale location history [2]. For the output unit, we used a word-level token from the English transcription. Because there is a performance gap between the ASR and the ground truth cause by imperfect transcription, we assume the metric (calculated based on the ASR transcription) is the lower-bound for the related translation model. We utilized two metrics to evaluate the translation performance from the transcribed text: BLEU scores and METEOR with a Multeval toolkit. Our pre-trained ASR model resulted in a 2.84% WER, a 94.9 BLEU, and a 69.1 METEOR on English speech utterances from the BTEC test set, and we set those scores as the groundtruth topline scores.

3 Results and Discussion

In this section, we present our experimental result and followed by the discussion.

3.1 Baseline

For the baseline translation task, we trained a sequence-to-sequence model from speech-to-speech directly. However, this approach did not converge at all and produced no audible speech. [1] also observed a similar result with a similar scenario.

3.2 Topline with Cascade ASR-TTS

In this paper, we set the topline performance by using the cascade of ASR and TTS system. First,

Table 1: Our experiment results based on BTEC French-English speech-to-speech translation.

Model (FR-EN)		**BLEU**	**METEOR**
Baseline Tacotron with MFCC input		-	-
Proposed Speech2Code			
Codebook	Time Reduction		
64	4	16.1	16.9
64	8	24.4	22.9
64	12	25.0	23.2
Topline (Cascade ASR ->TTS)		47.4	41.2

Table 2: Our experiment results based on BTEC Japanese-English speech-to-speech translation.

Model (JA-EN)		**BLEU**	**METEOR**
Baseline Tacotron with MFCC input		-	-
Proposed Speech2Code			
Codebook	Time Reduction		
128	4	11.9	13.5
128	8	15.3	15.3
128	12	14.9	14.5
Topline (Cascade ASR ->TTS)		37.4	32.8

we train the ASR system by using the source language MFCC as the input and target language character transcription. Second, we train a TTS based on Tacotron [5] to generate a speech from the target language characters to the target language speech representation.

3.3 Speech to Codebook

Table 1 shows the result for French-to-English Speech translation. Our best performance was produced by codebook of 64 and a time-reduction factor of 12 with a score of 25.0 BLEU and 23.2 METEOR.

Table 2 shows the result for Japanese-to-English Speech translation. Our best performance was produced by a codebook of 128 and a time-reduction factor 8 with a score of 15.3 BLEU and 15.3 METEOR.

4 Conclusion

In this paper, we proposed a novel approach for training a speech-to-speech translation between two languages without any transcription. Our model can perform a direct speech-to-speech translation on French-English and Japanese-English.

5 Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

References

[1] Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*, 2019.

[2] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Multi-scale alignment and contextual history for attention mechanism in sequence-to-sequence model. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 648–655. IEEE, 2018.

[3] Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura. VQVAE unsupervised unit discovery and multi-scale code2spec inverter for zerospeech challenge 2019. *CoRR*, Vol. abs/1905.11449, , 2019.

[4] Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pp. 6306–6315, 2017.

[5] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.