From Pixels to Voice: A Simple and Efficient End-to-End Spoken Image Description Approach via Vision Codec Language Models

Author: Chung Tran[1] - Sakriani Sakti[1]

1 Nara Institute of Science and Technology (NAIST), Japan

Outline

1. Introduction
2. Related Works
3. Methodology
4. Results
5. Conclusion

Introduction & Related Work

Introduction

- Recently, generative models in NLP have extended/applied to speech processing tasks (TTS)
- By using speech tokenizer (e.g. Encodec[1])

  Speech signal (raw) ➔ Speech Unit ➔ Speech signal (reconstruction)
- Methods:
  - Standard: Phoneme ➔ Mel-spectrogram ➔ Speech (Tacotron2, Fastspeech2)
  - New: Phoneme ➔ Speech Unit ➔ Speech (VALL-E, VALL-E X)

[1] Alexandre Défossez et al, High Fidelity Neural Audio Compression

Introduction

- VioLA[2], LauraGPT[3] has extended the new concept to perform many speech processing tasks
  - → Audio generation, speech generation, speech translation
- While these models excel in processing sequential inputs such as text or speech
  - → Processing non-sequential data, like images, remains underexplored
- Human communication involves not only text, and speech but also images
- The model that takes images as input and outputs text/speech can be applied to many applications:
  - → Helping visually impaired people understand their surroundings
- However, many languages do not have standardized writing systems [4]
  - → Limits the applicability of text-based technologies

[2] Tianrui Wang et al, VioLA: Unified Codec Language Models for Speech Recognition, Synthesis, and Translation
[3] Zhihao Du et al, LauraGPT: Listen, Attend, Understand, and Regenerate Audio with GPT
[4] Gilles Adda et al, Breaking the Unwritten Language Barrier

Related Works

- ❖ Recent studies (SAT[5], E-I2S[6], Im2Sp[7]) have proposed models that can describe images in speech without using text representation
- ❖ These studies need to train multiple components
  - → Image-to-Unit (I2U) + Unit-to-Speech (U2S) [SAT[5], Im2Sp[7]]
  - → Image-to-Unit (I2U) + VQ-VAE + Vocoder [E-I2S[6]]
- ❖ Limitation:
  - → Train multiple components ➔ increase complexity
  - → Retrain all models if speech-unit changes
- ❖ Show-and-Speak (SAS[8]) use a pretrained Faster-RCNN + a modified Tacotron2 (E2E)
  - → Depend on the external model to extract feature (36 objects)
  - → Performance is low due to limitation of Faster-RCNN (missed detection, misidentifications)

[5] Wei-Ning Hsu et al, Text- Free Image-to-Speech Synthesis Using Learned Segmental Units
[6] Johanes Effendi et al, End-to-end image-to-speech generation for untranscribed unknown languages
[7] Minsu Kim et al, Towards practical and efficient image-to-speech captioning with vision-language pre-training and multi-modal tokens
[8] Xinsheng Wang et al, Show and Speak: Directly Synthesize Spoken Description of Images

Proposal Approach

- Ours is the first use an off-the-shelf audio codec model to extract discrete representations and reconstruct it into speech
  - → Simplifies I2S training by focusing exclusively on the vision-language model.
- Ours use vision transformer to learn feature end-to-end, reducing the need for external or hand-crafted feature extraction.
- Experiments on the Flickr8k dataset:
  - → Our model is easier to train and infer
  - → Delivering promising results compared to existing I2S methods.

Methodology

Problem Formulation

- Step 1: Speech Encoding and Reconstruction

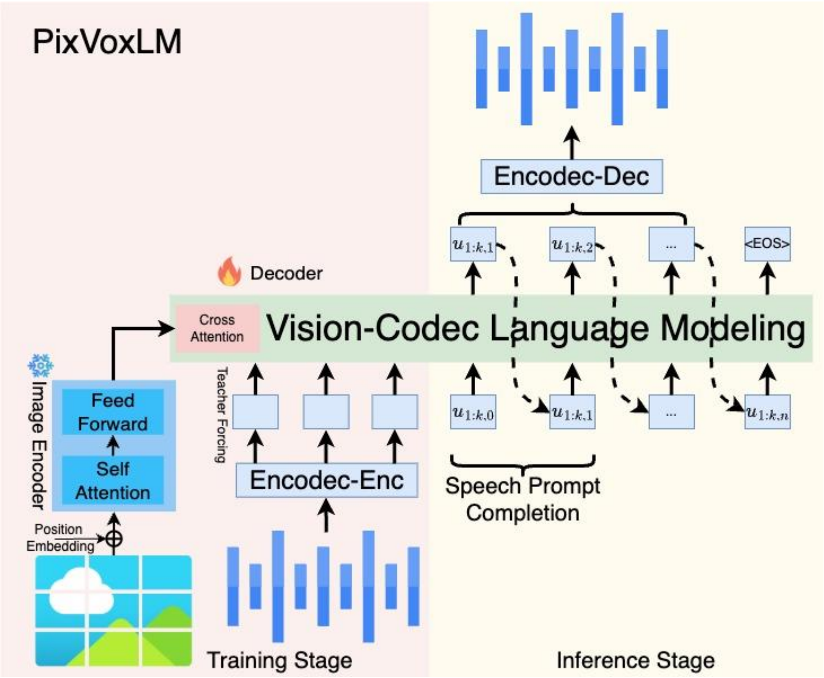
$$U = Encodec - Enc(S)$$

$$\hat{S} = Encodec - Dec(U)$$

- Step 2: Image-2-Unit (I2U) Mapping

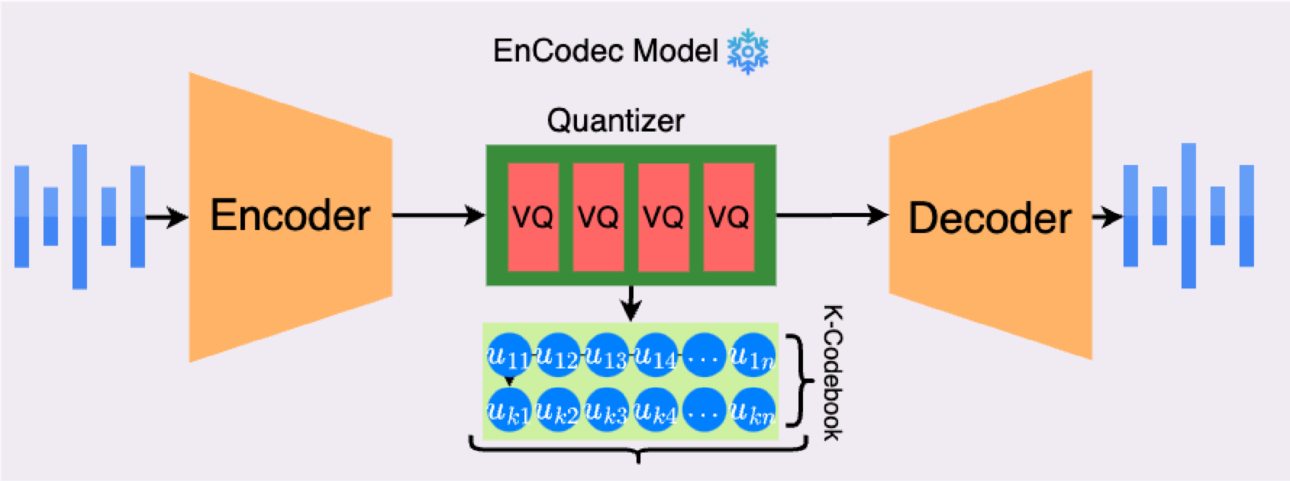
$$\hat{U} = I2U(I)$$

Train only I2U



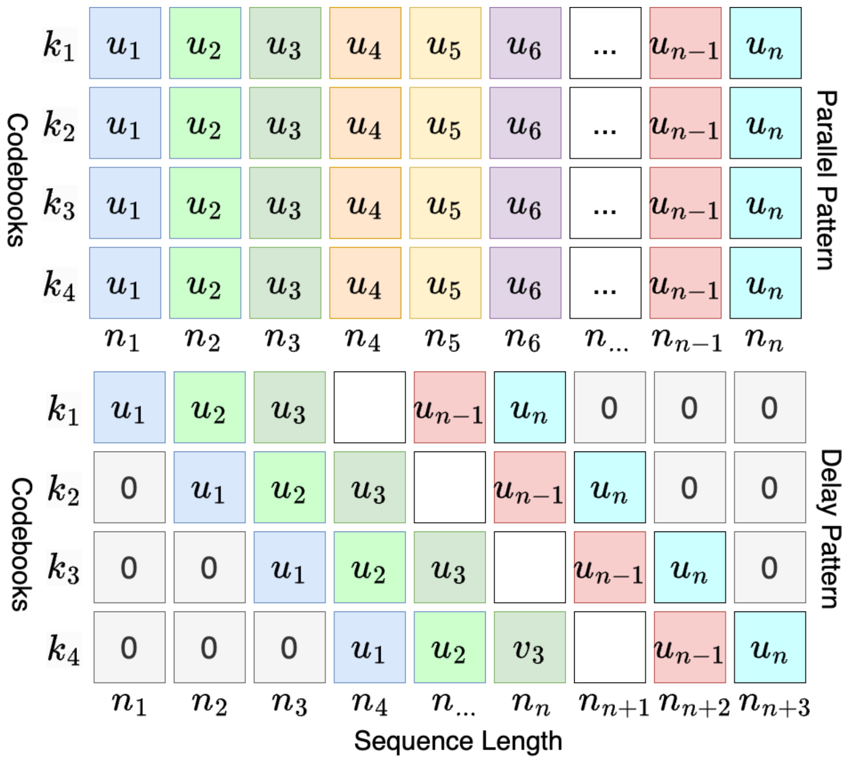
Neural Encodec

- A model that can convert audio signals into discrete representations, and vice versa



Neural Encodec

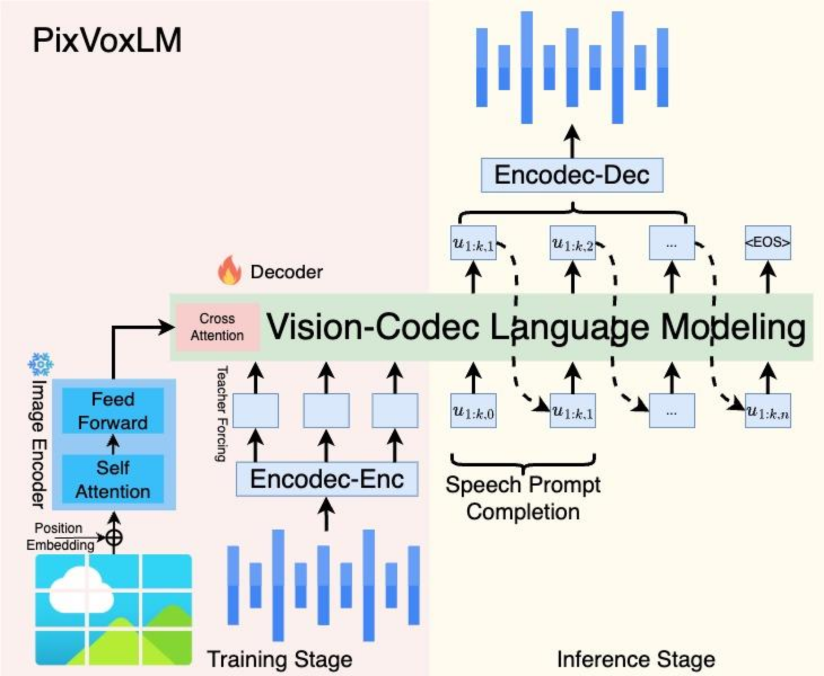
- Parallel Pattern [9]
- Dellay Pattern   [9]



[9] Jade Copet et al, Simple and Controllable Music Generation

Vision-Codec Language Model

- Image Encoder
- Unit Decoder
- Objective function:
  - → $L = \sum_{i=1}^{k} \sum_{n=1}^{N} \hat{u} \log p(\hat{u}|u)$



Results

Experiment setup

- Dataset: Flickr8k (8000 images)
  - → 6000 for training, 1000 for validation and 1000 for test
  - → Each image has five spoken captions
- Experiment setup
  - → Vision-Codec Language Model: BLIP model
  - → The trainable parameters 125M out of 211M
  - → Learning Rate: 5e-5, batchsize=60

Result

- Delay Pattern performs better than Parallel Pattern
- PixVoxLM outperforms the end-to-end SAS model
- PixVoxLM (delay pattern) is better than SAT and SAT-FT in the M and C metrics.

TABLE I
PERFORMANCE COMPARISON OF PIXVOXLM WITH EXISTING I2S MODELS ACROSS VARIOUS EVALUATION METRICS

Methods	B1↑	B2↑	B3↑	B4↑	M↑	R↑	C↑
Multiple-Model Training							
SAT [12]	-	-	-	11.60	14.10	39.00	23.20
SAT-FT [12]	-	-	-	12.60	14.50	39.10	24.20
E-I2S [14]	-	-	-	14.78	17.40	45.75	32.89
Single-Model Training							
SAS [18]	29.60	14.70	7.20	3.50	11.30	23.20	8.00
PixVoxLM-Parallel	34.52	18.75	10.65	6.22	10.51	26.30	9.43
PixVoxLM-Delay	48.08	30.59	18.92	11.49	15.19	35.76	25.54

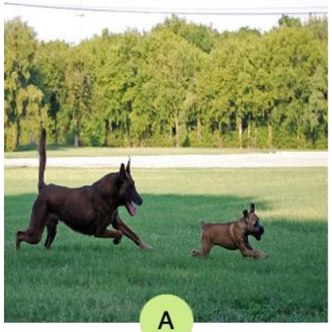
Visual-guided speech completion

- Use image and partial speech inputs for more accurate and context-aware completions.
- Delay pattern have better result than Parallel

TABLE II
SPEECH PROMPT COMPLETION AT VARIOUS INFORMATION LEVELS

PixVoxLM	Prompt	B1↑	B2↑	B3↑	B4↑	M↑	R↑	C↑
Parallel	0%	34.52	18.75	10.65	6.22	10.51	26.30	9.43
	25%	37.11	23.30	14.58	8.87	13.05	29.71	15.24
	50%	46.26	34.00	26.25	20.42	20.54	40.83	37.64
Delay	0%	48.08	30.59	18.92	11.49	15.19	35.76	25.54
	25%	49.76	34.18	23.04	14.90	18.00	39.04	32.68
	50%	57.31	44.3	35.31	28.11	24.19	48.35	60.10

Example

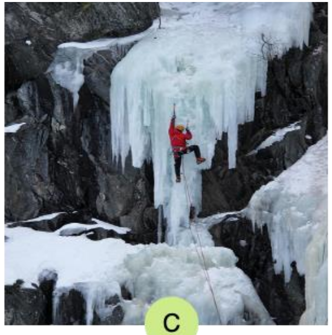


A

GT: Two dogs play in the grass
ASR: Two dogs running in grass

B

GT: Three children playing in sand at beach
ASR: Thre children playing in the sand



C

GT: A man climbs icy rocks
ASR: Clamber or climbing a neste

Conclusion

Conclusion & Future work

- Conclusion
  - PixVoxLM offers a simple and efficient solution for generating speech directly from images
  - PixVoxLM outperform the recent end-to-end SAS model
- Future work
  - Subjective evaluations highlight several issues
  - Need to improve the performance

Thank for your attention