VioLA: Unified Codec Language Models for Speech Recognition, Synthesis, and Translation

- Arxiv (23.05.23)
- Microsoft
- Paper

Overivew

"Is one decoder-only generative model all you need for speech recognition, synthesis, and translation?"

⇒ **A multi-lingual, multi-modal auto-regressive Transformer model**

- **all the speech utterances to discrete tokens**
  - **STT, TTS, Speech-to-Speech, Text-to-Text are converted to token-based sequence conversion problem**
  - **Task ID, Language ID**
  - **Embedding Module following Encodec**

- **ASR 성능: PER 11.36**
  - **Cf) Encoder-Decoder Model 9.47, Decoder Model 9.61)**
- **MT 성능: BLEU 55.85**
  - **cf) ASR->MT: 55.98)**
- **Zero-shot TTS 성능: Speaker Similarity 0.54, WER 4.97**
  - **cf) VALL-E X : Speaker Similarity 0.53, WER 5.81**

Model Structure

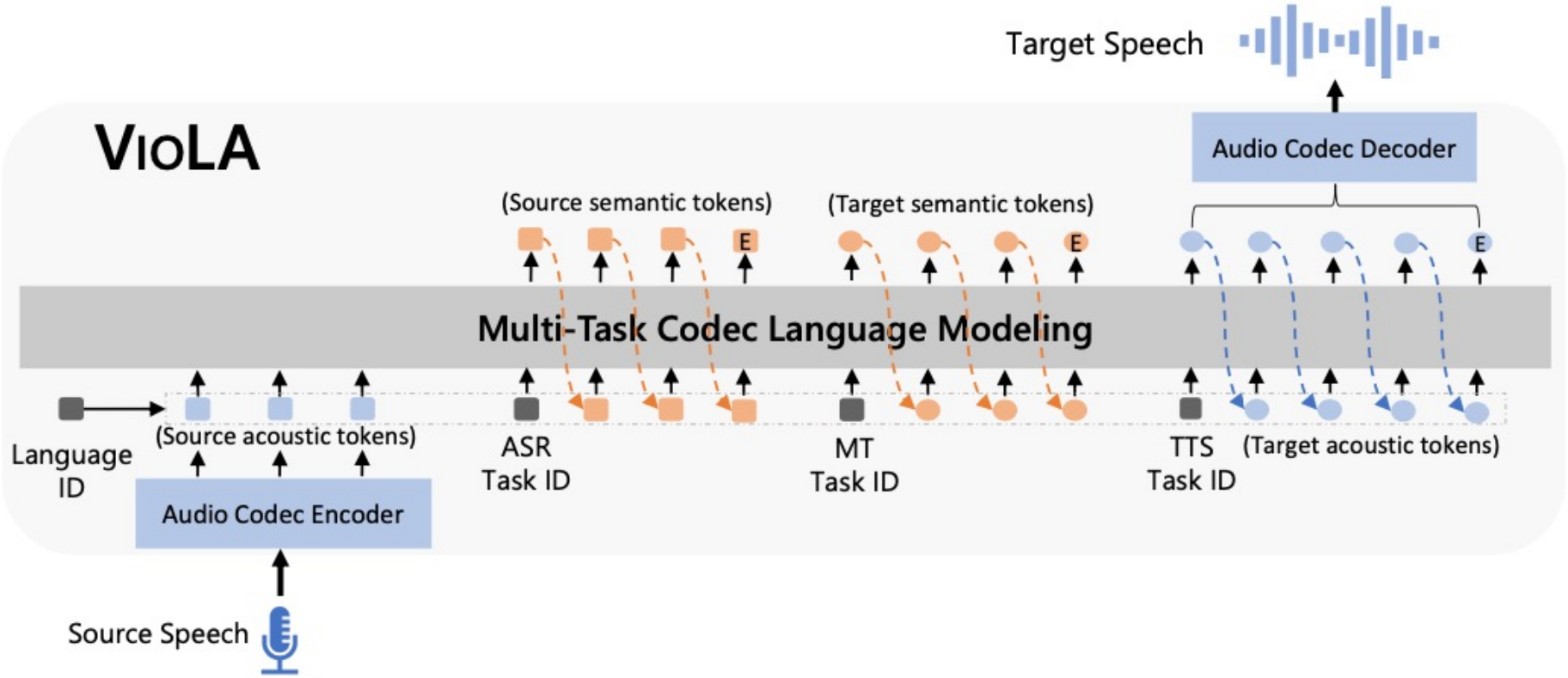
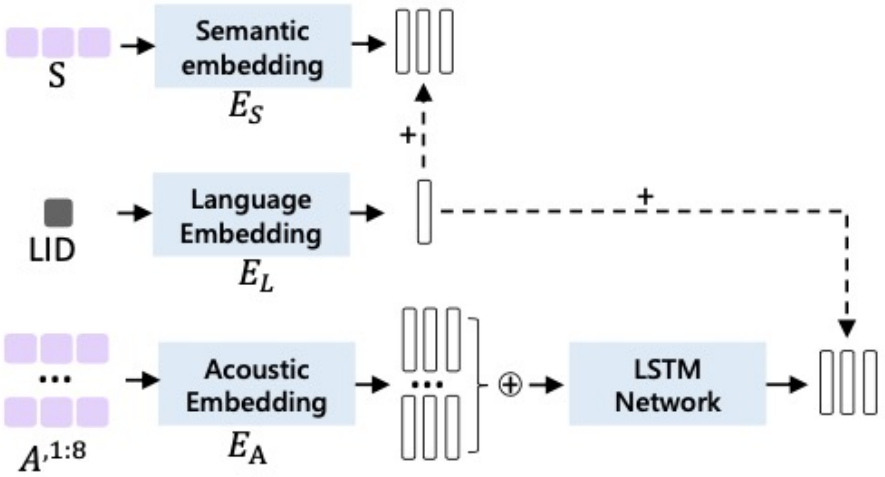


Figure 1: The overall framework of VIOLA, which regards various speech processing tasks as a conditional codec language model task. The model training is conducted on a multi-task learning framework with ASR, MT, and TTS tasks, and the model is capable of performing speech-to-text recognition and translation, text-to-text translation, text-to-speech synthesis, and speech-to-speech translation tasks.

Model Structure

- **Embedding Module**



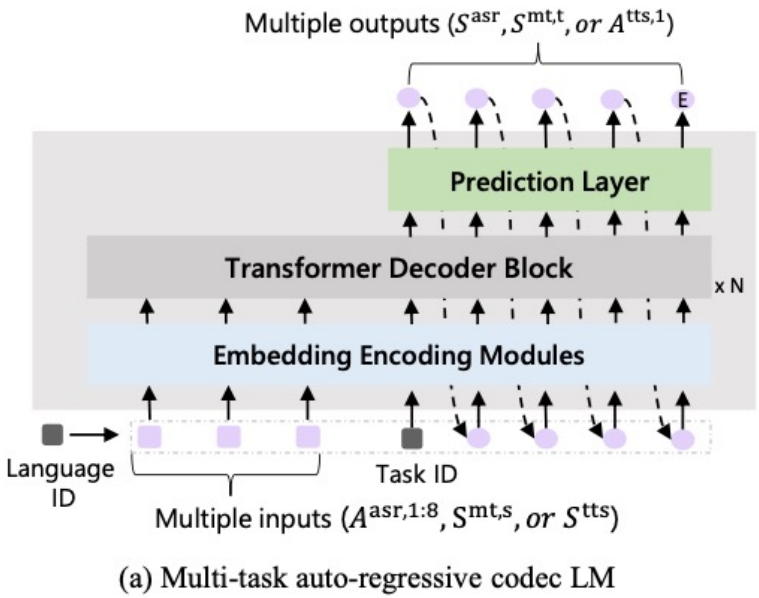
(b) Embedding encoding modules

- **Semantic tokens**
  - **G2P**
  - **Random initialized embedding**
- **Language Embedding**
  - **Language ID**
- **Acoustic Embedding**
  - **8-layer acoustic tokens from Encodec**
  - **Layer별로 embedding 구성한 다음 average**
  - **Unidirectional LSTM에 넣음**

  ⇒ **1024-dimensional embedding**

Model Structure

- **Training**



(a) Multi-task auto-regressive codec LM

- **ASR**
  - **Input: acoustic token (8-layer)**
  - **Output: semantic token**
- **MT**
  - **Input: phoneme -> semantic token (source lang)**
  - **Output: semantic token (target lang)**
- **TTS**
  - **Input: Phoneme -> semantic token**
  - **Output: acoustic token (1-layer)**

- **Training Loss**
  - **3가지 Task를 한꺼번에 훈련**

$$\mathcal{L} = \mathcal{L}_{asr} + \mathcal{L}_{mt} + \mathcal{L}_{tts}$$

$$\mathcal{L}_{asr} = p\left(\boldsymbol{S}^{\text{asr}} | \boldsymbol{A}^{\text{asr},1:8}, b_{\text{asr}}; \theta\right) = \prod_{t=0}^{T_{\text{S}}} p\left(S_t^{\text{asr}} | \boldsymbol{A}^{\text{asr},1:8}, b_{\text{asr}}, \boldsymbol{S}_{<t}^{\text{asr}}, \theta\right)$$

$$\mathcal{L}_{mt} = p\left(\boldsymbol{S}^{\text{mt,t}} | \boldsymbol{S}^{\text{mt, s}}, b_{\text{mt}}; \theta\right) = \prod_{t=0}^{T_{\text{S}}} p\left(S_t^{\text{mt, t}} | \boldsymbol{S}^{\text{mt,s}}, b_{\text{mt}}, \boldsymbol{S}_{<t}^{\text{mt,t}}, \theta\right)$$

$$\mathcal{L}_{tts} = p\left(\boldsymbol{A}^{\text{tts},1} | \boldsymbol{S}^{\text{tts}}, b_{\text{tts}}; \theta\right) = \prod_{t=0}^{T_{\text{A}}} p\left(A_n^{\text{tts},1} | \boldsymbol{S}^{\text{tts}}, b_{\text{tts}}, \boldsymbol{A}_{<t}^{\text{tts},1}, \theta\right)$$

Model Structure

- Inference
  - ASR
    - Beam search decoding
  - MT
    - Beam search decoding
  - TTS
    - Sampling decoding
    - Strategy 1: 5번 결과 내서 Speaker Similarity가 가장 좋은 것 채택
    - Strategy 2: 5번 결과 내서 Speaker Similarity + WER 가 가장 좋은 것 채택
  - Speech-to-text translation
    - ASR -> MT로 수행
  - Speech-to-speech translation
    - ASR->MT->TTS로 수행

Experiment

- **Training data**
  - **ASR, TTS**
    - **중국어 data WenetSpeech (10,000 시간)**
    - **영어 data Librilight (60,000 시간)**
      - Librispeech를 학습한 ASR 모델로 transcript 생성해서 학습에 사용
  - **MT**
    - **중국어-영어 63M sentence pair (AI Chanllenger + WMT2020)**
- **평가 데이터**
  - **ASR: WenetSpeech의 dev set**
  - **MT: WMT2020의 test set**
  - **Chinese-to-English S2TT, zero-shot English TTS prompted by Chinese speech, zero-short Chinese-to-English S2ST**
    - **EMIME : native Chinese speaker의 영어, 중국어 발화 음성**
  - **zero-shot En- glish TTS prompted by English speech**
    - **Librispeech dev-clean and test- clean sets**

Result

- **Baseline**
  - **Input: Fbank**
  - **AED (Attention-based Encoder Decoder) : 6 encoder 6 decoder cross attention**
  - **LM (decoder-only model)**
- **ASR**

Method	Input	Param.(M)	PER ↓
AED (enc-dec)	Fbank	246.3	**9.47**
LM (decoder)		**150.8**	9.61
LM (decoder)	Codec	**177.5**	11.71
VIOLA (12L)		178.5	12.97
VIOLA (18L)		250.6	**11.36**

Table 1: Comparison of different models on speech recognition task.

- **MT**

Method	Param.(M)	BLEU ↑
AED (enc-dec)	242.5	56.83
LM (decoder)	**145.4**	56.81
VIOLA (12L)	178.5	54.44
VIOLA (18L)	250.6	**56.97**

Table 2: Comparison of different models on machine translation task.

Method	Input	Total Param.(M)	BLEU ↑
$AED_{ASR} \rightarrow AED_{MT}$	Fbank	488.8	**55.98**
$LM_{ASR} \rightarrow LM_{MT}$	Codec	322.9	55.70
VIOLA (12L)		**178.5**	53.16
VIOLA (18L)		250.6	**55.85**

Table 3: Comparison of different models on speech-to-text translation task.

Result

- **Zero-Shot Text-to-Speech Synthesis**

Method	Strategy I			Strategy II			AVG		
	SS	WER	SN	SS	WER	SN	SS ↑	WER ↓	SN ↑
Ground Truth Audio	-	-	-	-	-	-	0.67	1.98	3.81
VALL-E X	0.53	5.81	3.20	0.49	3.38	3.20	0.51	4.60	3.20
VIOLA (12L)	0.52	6.13	3.20	0.47	3.75	3.19	0.50	4.94	3.20
VIOLA (18L)	**0.54**	**4.97**	**3.22**	**0.50**	**2.89**	**3.21**	**0.52**	**3.93**	**3.22**

Table 4: Comparison of different models on zero-shot text-to-speech task. SS means speaker similarity, SN means speech naturalness, AVG means the average scores of Strategy I and Strategy II.

- **Zero-Shot Cross-Lingual TTS / zero-shot speech-to-speech**

Method	Input	Strategy I			Strategy II			AVG		
		SS	BLEU	SN	SS	BLEU	SN	SS ↑	BLEU ↑	SN ↑
Ground Truth Audio	-	-	-	-	-	-	-	0.58	93.31	3.82
Zero-shot cross-lingual text-to-speech										
Target text → VALL-E X	Text	0.49	69.37	3.32	0.44	83.99	3.35	0.47	76.68	3.34
Target text → VIOLA (12L)		0.49	65.22	3.32	0.43	82.42	3.31	0.46	73.82	3.32
Target text → VIOLA (18L)		**0.50**	**72.55**	**3.33**	**0.45**	**85.21**	**3.36**	**0.48**	**78.88**	**3.35**
Zero-shot speech-to-speech										
AED_{ASR} → AED_{MT} → VALL-E X	Fbank	0.50	42.00	3.52	0.48	49.30	3.53	0.49	45.65	3.53
LM_{ASR} → LM_{MT} → VALL-E X		0.50	41.05	3.51	0.48	48.74	3.52	0.49	44.90	3.52
VIOLA (12L)	Codec	0.49	39.26	3.49	0.47	45.86	3.52	0.48	42.56	3.51
VIOLA (18L)		**0.51**	**43.96**	**3.54**	**0.49**	**51.57**	**3.56**	**0.50**	**47.77**	**3.55**

Table 5: Comparison on zero-shot cross-lingual text-to-speech and speech-to-speech translation tasks.

Conclusion & Limitation

- **Conclusion**
  - **Speech -> Codec-based token**
  - **Token conversion problem**
  - **Multi-lingual, multi-task**

- **Limitation**
  - **TTS에서만 In-context learning**
  - **Cascaded inference for speech-to-text, speech-to-speech translation tasks**