T2V2
A Unified Non-Autoregressive Model for Speech Recognition and Synthesis via Multitask Learning

Nabarun Goswami[1], Hanqin Wang[1], Tatsuya Harada[1,2]

[1]The University of Tokyo, Japan

[2]RIKEN, Japan

5 April 2025

Acknowledgement

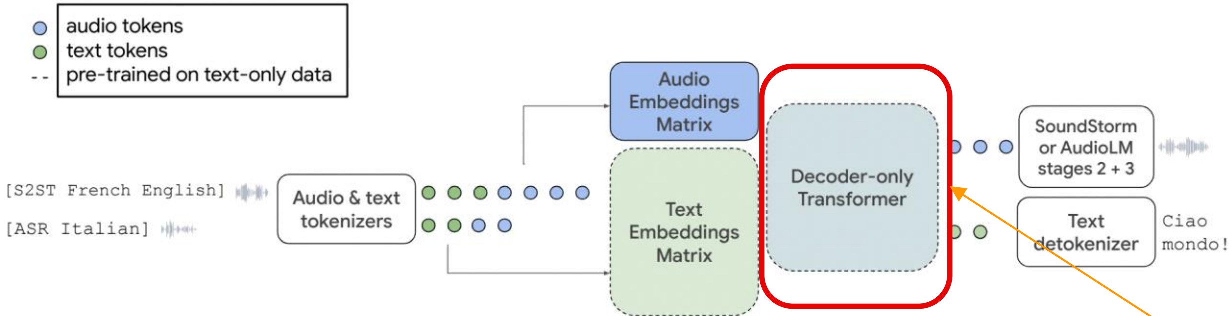
This work was partially supported by:

- JST Moonshot R&D Grant Number JPMJPS2011
- CREST Grant Number JPMJCR2015
- Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo.

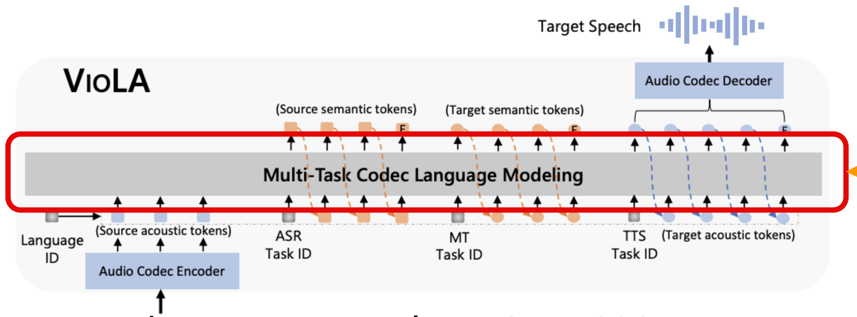
Motivation

- **Challenges:**
  - High latency in autoregressive (AR) models
  - External alignment tools increase complexity in non-autoregressive (NAR) models
  - Lack of unified representation limits cross-task improvements
- **Benefits of Unified ASR-TTS with Discrete Tokens:**
  - Shared discrete representations improve efficiency and scalability
  - Discrete tokens enable efficient storage, transmission, and improved sequence modeling
  - Single efficient training process (both tasks typically trained on the same data)
  - Dual-task modeling allows tasks to mutually aid and enhance each other's performance

Related Works in Discrete Unified ASR-TTS

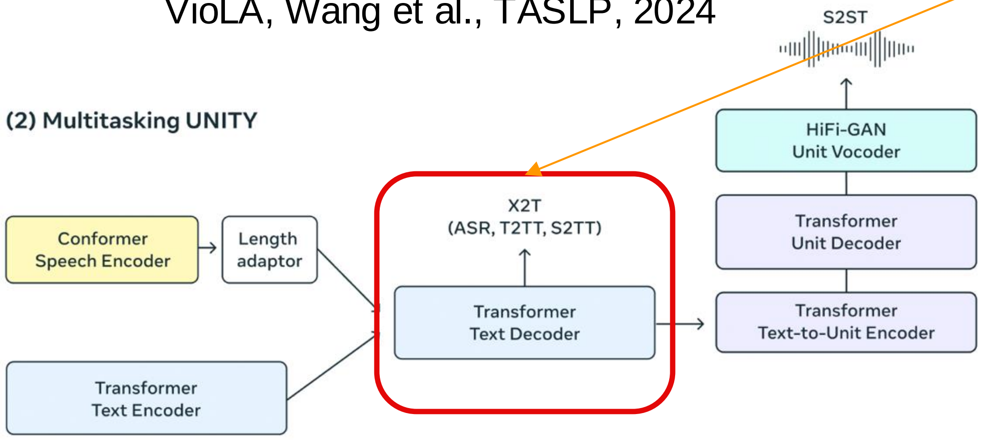


AudioPALM, Rubenstein et al., Technical Report, 2023



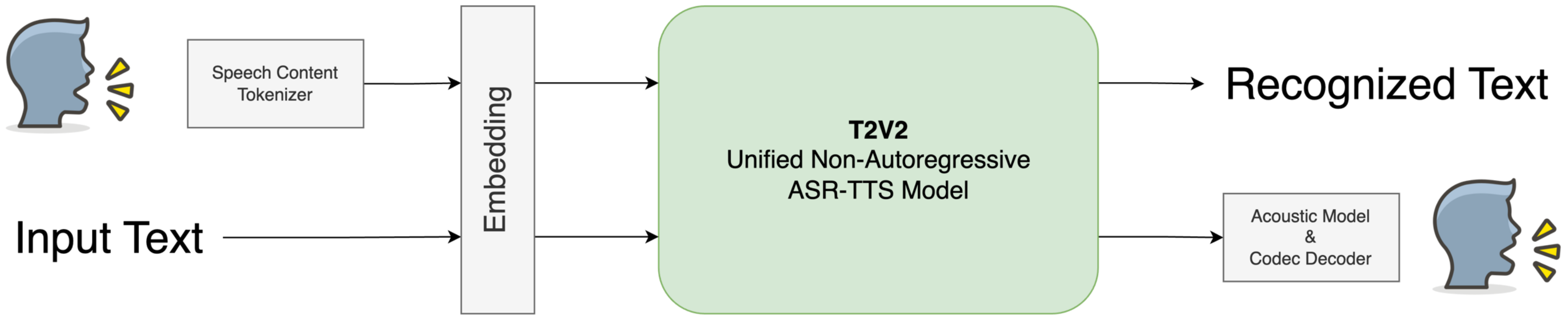
VioLA, Wang et al., TASLP, 2024

Autoregressive Components



SeamlessM4T, Barrault et al., Technical Report, 2023

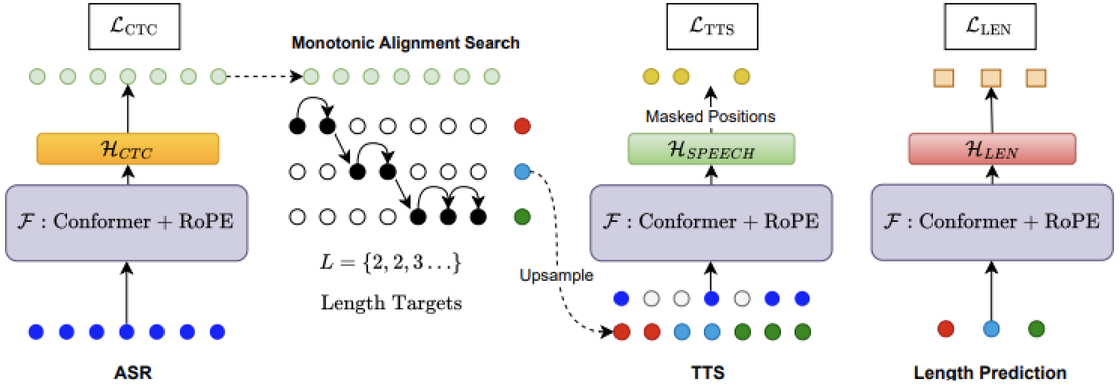
Overall Pipeline of T2V2



T2V2: Task Details

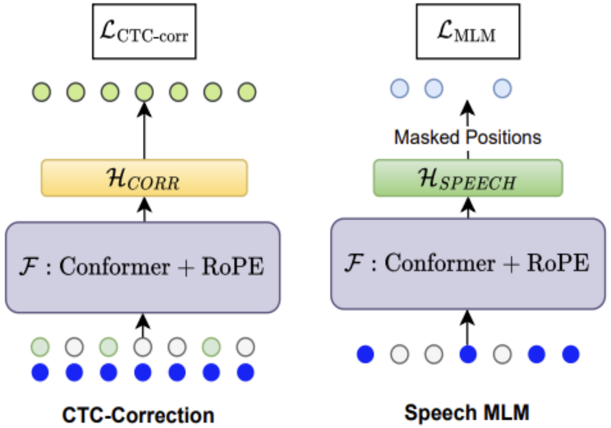
Core Tasks:

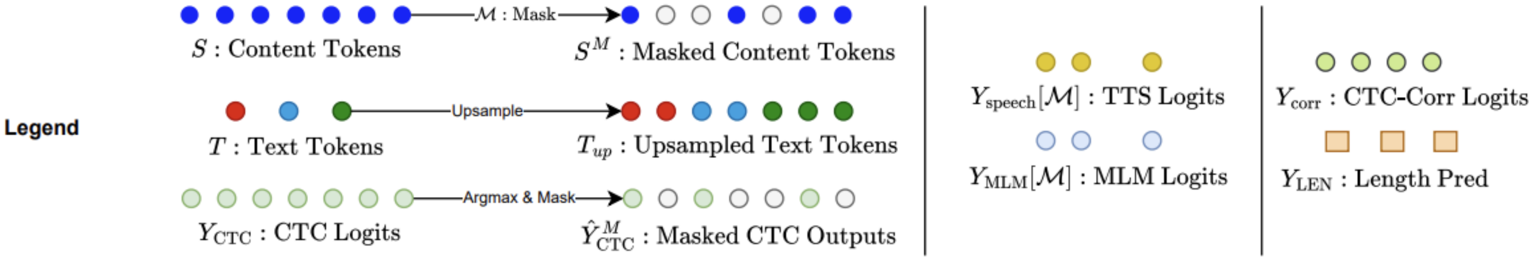
- ASR: CTC-based training
- TTS: Masked language modeling (MLM) with Monotonic Alignment Search (MAS) with intermediate CTC outputs.



Auxiliary Tasks:

- CTC Error Correction: Refines ASR outputs
- Unconditional Speech MLM: Enables classifier-free guidance for TTS





Loss Functions:

$$\mathcal{L}_{\text{CTC}} = -\log \sum_{\mathbf{a} \in \mathcal{A}(\boldsymbol{T})} P(\mathbf{a} | \mathbf{Y}_{\text{CTC}})$$

$$\mathcal{L}_{\text{TTS}} = -\sum_{i \in \mathcal{M}} \boldsymbol{S}_i \log P(\mathcal{H}_{\text{SPEECH}}(\mathcal{F}(\boldsymbol{X}_{\text{TTS}}))_i)$$

$$\mathcal{L}_{\text{LEN}} = \sum_i |\mathcal{H}_{\text{LEN}}(\mathcal{F}(\boldsymbol{T}))_i - \log(\boldsymbol{L}_i)|$$

$$\mathcal{L}_{\text{CTC-corr}} = -\log \sum_{\mathbf{a} \in \mathcal{A}(\boldsymbol{T})} P(\mathbf{a} | \mathbf{Y}_{\text{corr}})$$

Model Architecture: Shared Conformer with RoPE + Task-specific Heads

Key Introductions and Contributions

- **Unified Multitask Learning:**
  - First NAR unified model for ASR and TTS achieved via Multitask Learning
- **Monotonic Alignment Search with Intermediate CTC outputs:**
  - Self-contained alignment method, removing dependence on external tools
- **CTC Error Correction:**
  - Addresses CTC independence limitation
- **Classifier-Free Guidance:**
  - Improves robustness in TTS

Experimental Setup

- **Model Architecture:** 6-layer Conformer (D=384, H=8, FF=1536, Ks=7)
- **Additional Modules (pre-trained on LibriLight (60K hours)):**
  - **Content Tokenizer:** HuBERT-Kmeans (1024 clusters, @50Hz)
  - **Codec:** Descript Audio Codec (12-layer RVQ @50Hz)
  - **Acoustic Model (content → acoustic) :** SoundStorm
- **Datasets:**
  - **Train:** LibriHeavy (small: 500 hours, large: 50K hours)
  - **Test**:
    - **ASR**: Librispeech *test-clean*
    - **TTS**: 40 sentences from LibriSpeech test-clean, 20 speaker prompts from DAPS

Zero-Shot TTS Results

Table 4: Zero-shot TTS performance comparison. Methods with * indicate multilingual models. UD refers to Unpaired Data while PD refers to Paired Data in hours.

	UD	PD	UTMOS	CER	SECS	IR-e2e (s)	IR-t2c (s)
Large scale paired data							
HierSpeech++*	500k	2.8k	**4.46 ± 0.02**	0.88	**0.94 ± 0.01**	0.16 ± 0.00	-
XTTS*	-	27k	4.12 ± 0.07	0.78	0.93 ± 0.01	2.60 ± 0.03	-
WhisperSpeech	60k	60k	3.95 ± 0.11	0.66	0.93 ± 0.01	17.91 ± 0.04	2.84 ± 0.01
Small scale paired data							
YourTTS*	-	689	3.69 ± 0.08	2.02	0.90 ± 0.02	**0.11 ± 0.00**	-
StyleTTS2	94k	245	4.43 ± 0.03	1.59	0.91 ± 0.02	0.27 ± 0.00	-
Ours	60k	500	4.43 ± 0.02	**0.55**	**0.94 ± 0.01**	0.57 ± 0.00	**0.06 ± 0.00**

Table 5: Comparative MOS for Speech Quality (CMOS) and Speaker Similarity (SCMOS) on a scale $\{-2, +2\}$. p-value ≤ 0.05 indicate statistical significance.

	CMOS (p-value)	SCMOS (p-value)
HierSpeech++	**+0.10 ± 0.25 (0.337)**	**+0.12 ± 0.26 (0.287)**
XTTS	**-0.13 ± 0.28 (0.418)**	-0.30 ± 0.22 (0.007)
StyleTTS2	**+0.16 ± 0.25 (0.271)**	**+0.14 ± 0.24 (0.201)**
WhisperSpeech	**-0.11 ± 0.27 (0.490)**	-0.63 ± 0.21 ($1.5e^{-7}$)
Ours	**0.00**	**0.00**

State-of-the-Art UTMOS, CER, SECS

Significantly faster than AR baselines

State-of-the-Art CMOS, SCMOS

Zero-Shot TTS Samples:

Text input: Rodolfo meanwhile having returned home, and having missed the crucifix, guessed who had taken it, but gave himself no concern about it.

Text input: The railroads had not reached Jackson county, and wild game was plentiful on my father's farm on Big Creek near Lee's Summit.

Speaker Prompt

Our Output

Speaker Prompt

Our Output

Discrete ASR Result

Table 9: ASR results for models trained with punctuation and casing. The publicly released models for Zipformer-Transducer are used for the evaluation, while Conformer-CTC is trained by us.

	Libriheavy Subset	CER	WER	IR (s)
Non-discrete ASR (BPE encoding)				
Zipformer-Transducer	small	2.01	5.33	1.49 ± 0.07
Zipformer-Transducer	large	0.66	1.99	1.51 ± 0.14
Discrete ASR (Byte Encoding)				
Conformer-CTC	small	**2.69**	8.28	0.32 ± 0.02
Ours	small	2.71	**8.27**	0.47 ± 0.03
Conformer-CTC	large	1.53	4.36	0.34 ± 0.02
Ours	large	**1.31**	**4.09**	0.55 ± 0.02

State-of-the-Art Discrete NAR ASR

Significantly faster than AR baselines

Performance gap with continuous AR baseline

Table 8: Individual error type improvements.

	Sub	Ins	Del
w/o CORR	1.300	0.090	0.140
w CORR	**1.255** ↓3.46%	**0.082** ↓8.89%	**0.135** ↓3.57%

CTC Error Correction improves all types of errors

Ablation Study (TTS)

Table 1: Zero-shot TTS ablation study for different tasks.

Task Setting	UTMOS	CER	SECS
w SMLM, w CORR	4.39 ± 0.04	0.95	0.94 ± 0.01
w SMLM, w/o CORR	4.41 ± 0.04	1.08	0.94 ± 0.01
w/o SMLM, w/o CORR	4.39 ± 0.03	0.82	0.94 ± 0.01

Table 2: Zero-shot TTS ablation study for different number of iterations.

Iters	UTMOS	CER	SECS
1	4.39 ± 0.04	**0.95**	0.94 ± 0.01
4	**4.43 ± 0.03**	1.12	0.94 ± 0.01
8	4.41 ± 0.03	1.23	0.94 ± 0.01

Table 3: TTS ablation study for CFG weight λ.

λ	UTMOS	CER	SECS
0.0	**4.43 ± 0.03**	1.12	0.94 ± 0.01
1.0	**4.43 ± 0.02**	**0.55**	0.94 ± 0.01
1.5	4.40 ± 0.04	0.95	0.94 ± 0.01
2.0	4.42 ± 0.02	0.69	0.94 ± 0.01

- Auxiliary tasks do not hamperTTS performance
- Increasing number of iterations increases quality but quickly saturates at 4 iterations
- CFG significantly improves robustness

Ablation Study (ASR)

Table 6: ASR ablation study for different tasks.

	CER	WER
w SMLM, w CORR	**2.732**	**8.651**
w SMLM, w/o CORR	2.949	9.428
w/o SMLM, w/o CORR	2.886	9.120

Table 7: ASR ablation study for different correction thresholds and iterations.

Corr. Thresh	Iters	CER	WER	IR(s)
w/o CORR	-	2.73	8.65	0.32 ± 0.02
0.8	1	2.73	8.44	0.40 ± 0.03
0.8	4	2.72	8.37	0.39 ± 0.03
0.8	8	2.72	8.33	0.42 ± 0.03
0.7	8	2.72	8.29	0.42 ± 0.03
0.7	16	**2.71**	**8.27**	0.47 ± 0.03
0.7	32	**2.71**	**8.27**	0.60 ± 0.03

- CTC-Correction task helps improve ASR performance

- Increasing number of iterations leads to improvement in ASR performance

Conclusion & Future Work

Conclusion:

T2V2 effectively integrates ASR & TTS, leveraging multitask learning and discrete tokens.

Limitations:

Slightly underperforming continuous feature-based ASR, separate content-acoustic token translation for TTS.

Ethical Considerations:

High-quality synthetic speech achievable with short samples poses risks of misuse; we verified synthetic speech detectability by third-party detectors (e.g. https://detect.resemble.ai/)

Future Directions:

Extend framework to multi-lingual and code-switching scenarios, improve discrete ASR performance.

Thank you!