# TASTE:
# Text-Aligned Speech Tokenization and Embedding for Spoken Language Modeling

Liang-Hsuan Tseng, Yi-Chang Chen, Kuan-Yi Lee,
Da-shan Shiu, Hung-yi Lee

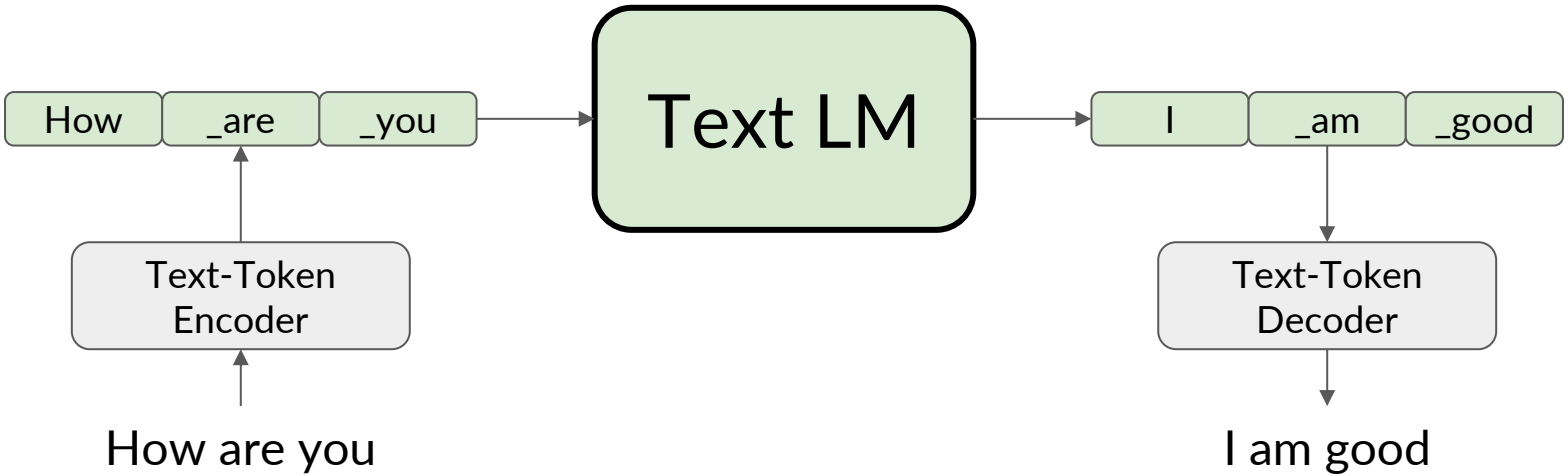
National Taiwan University

ICLR

MEDIATEK

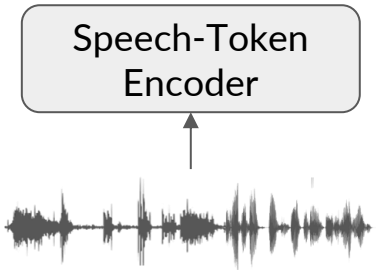
# **Background**: SLM

Inspired from the huge success of text language models, spoken language models aims at generating speech by **modeling speech tokens**.



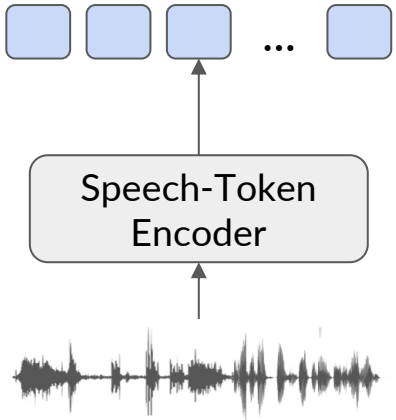
# **Background**: SLM

Inspired from the huge success of text language models, spoken language models (**SLM**) aim at generating speech by **modeling speech tokens**.



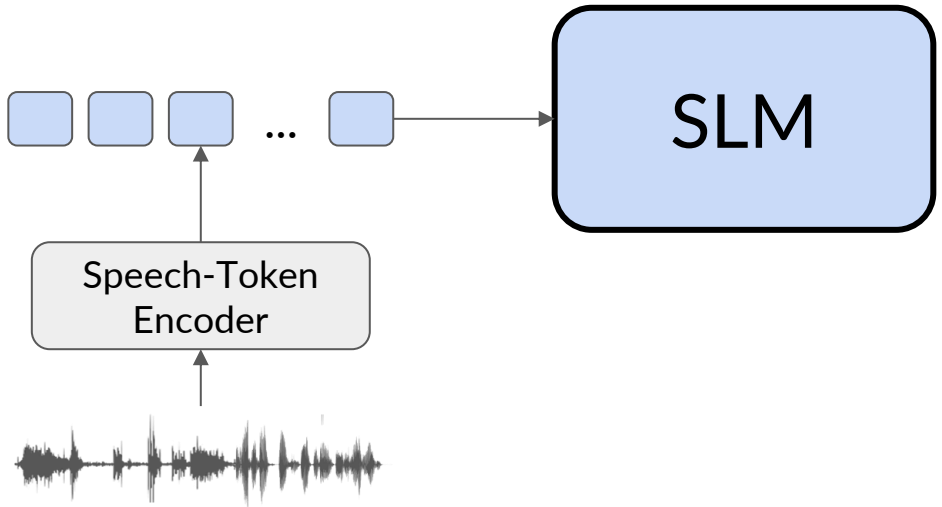
# **Background**: SLM

Inspired from the huge success of text language models, spoken language models (**SLM**) aim at generating speech by **modeling speech tokens**.



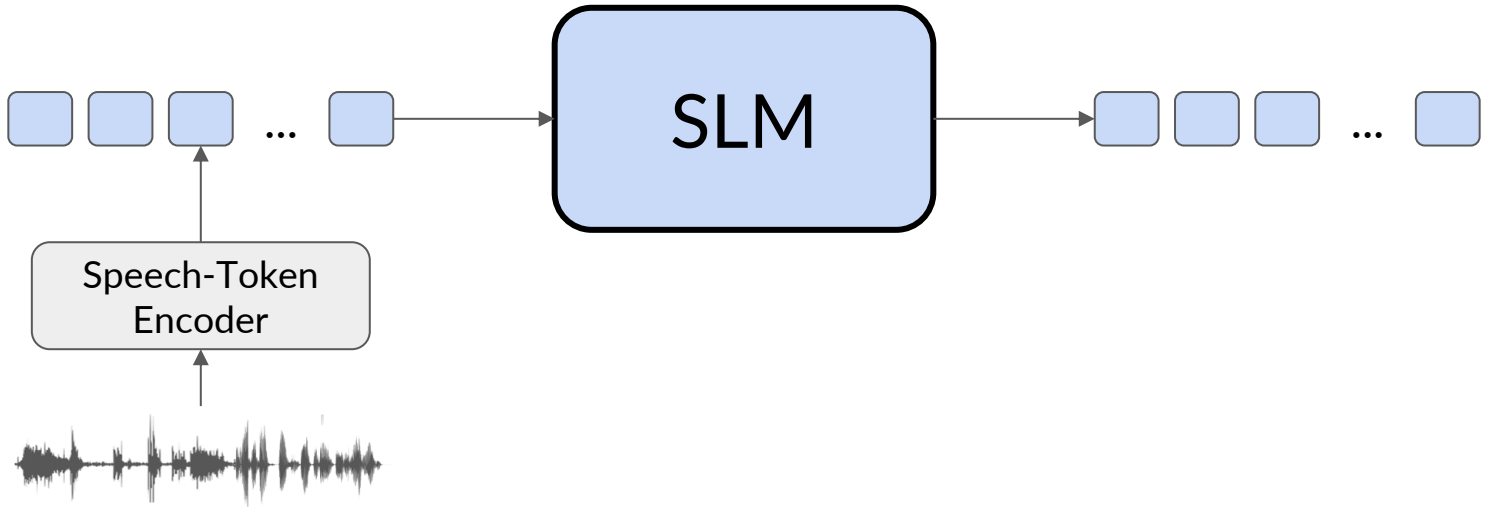
# **Background**: SLM

Inspired from the huge success of text language models, spoken language models (**SLM**) aim at generating speech by **modeling speech tokens**.



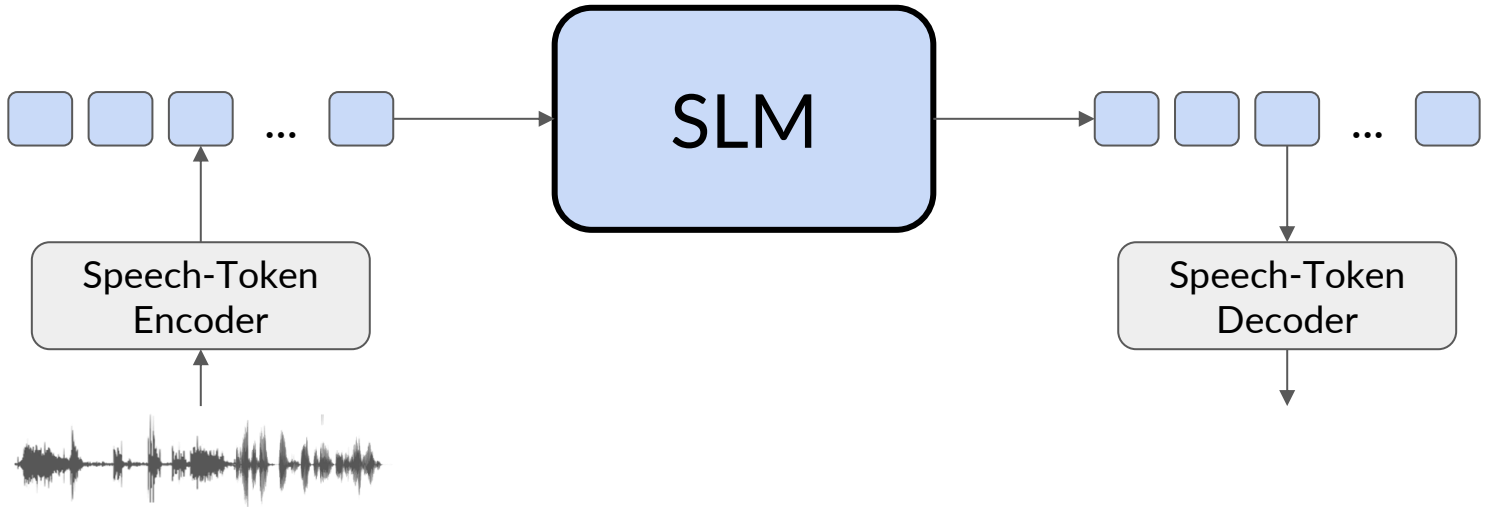
# **Background**: SLM

Inspired from the huge success of text language models, spoken language models (**SLM**) aim at generating speech by **modeling speech tokens**.



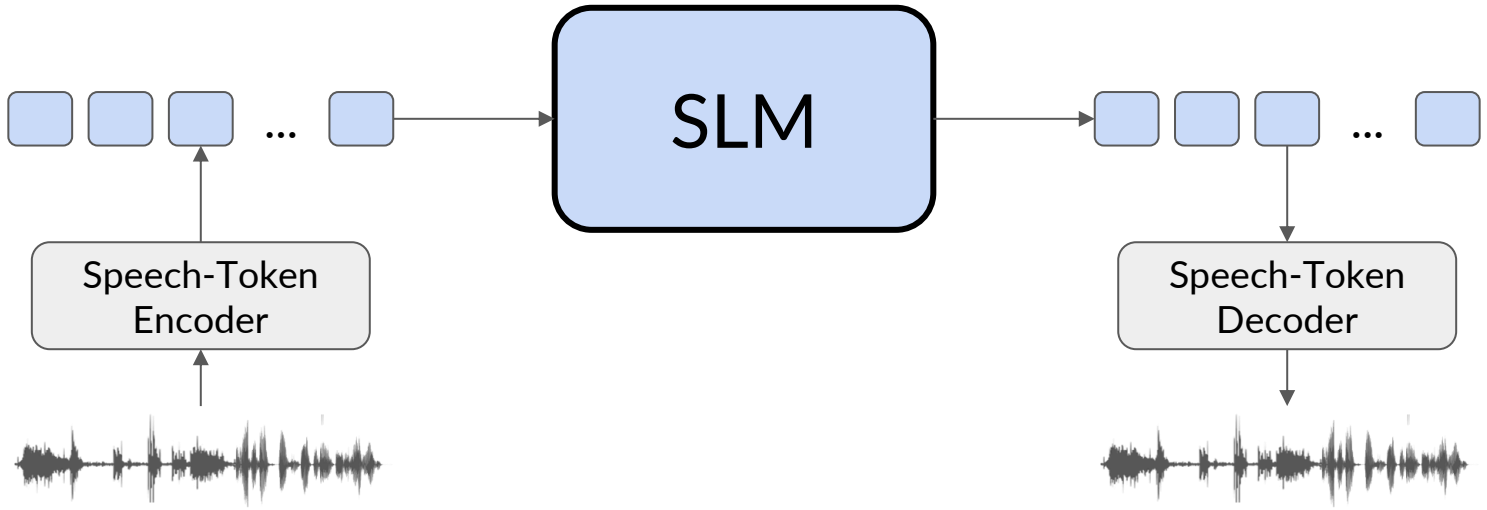
# **Background**: SLM

Inspired from the huge success of text language models, spoken language models (**SLM**) aim at generating speech by **modeling speech tokens**.



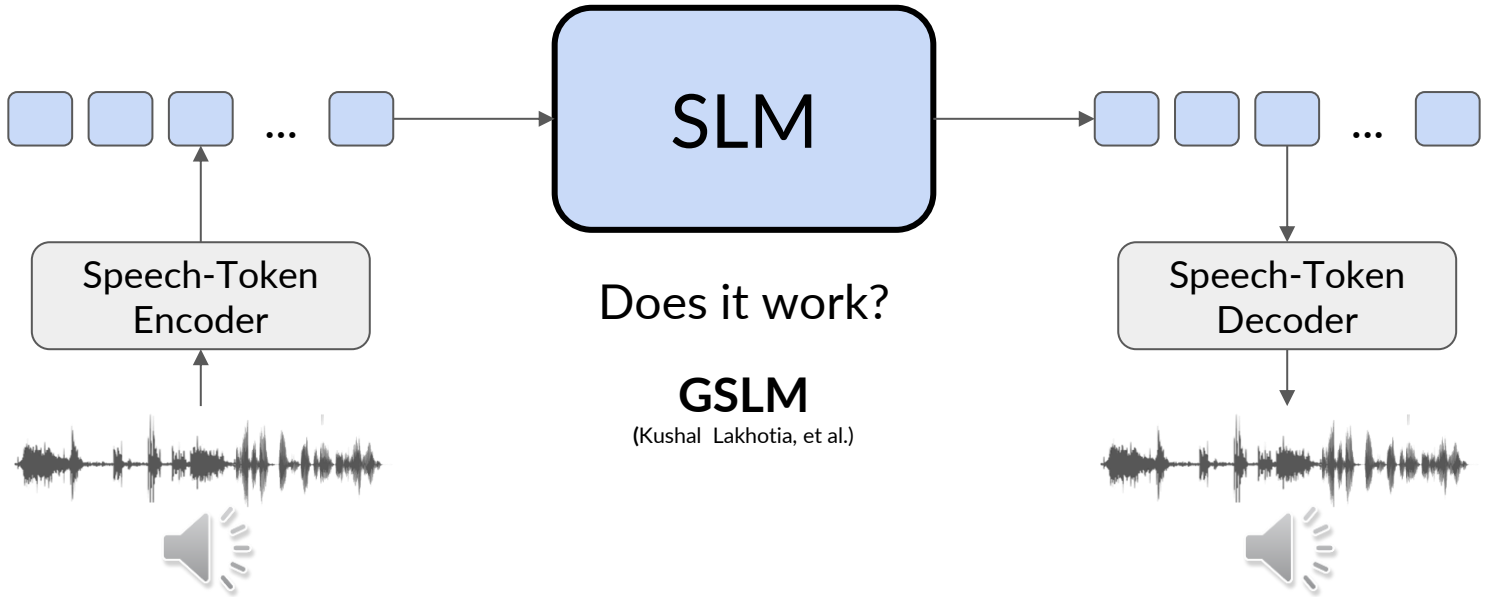
# **Background**: SLM

Inspired from the huge success of text language models, spoken language models (**SLM**) aim at generating speech by **modeling speech tokens**.



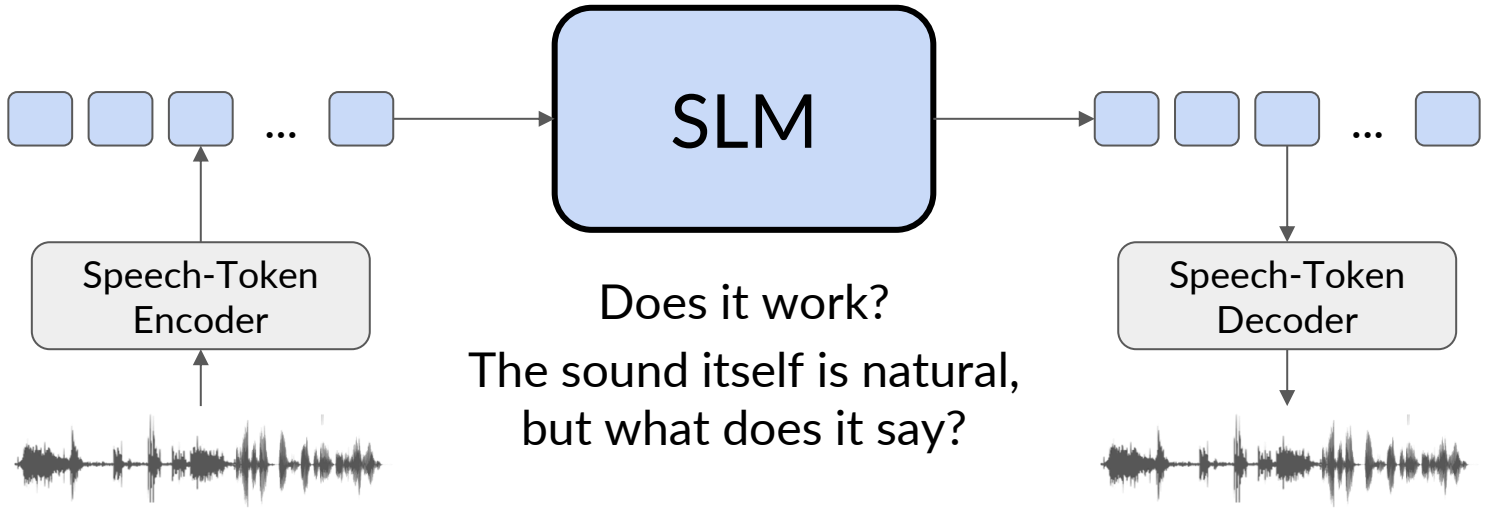
# **Background**: SLM

Inspired from the huge success of text language models, spoken language models (**SLM**) aim at generating speech by **modeling speech tokens**.



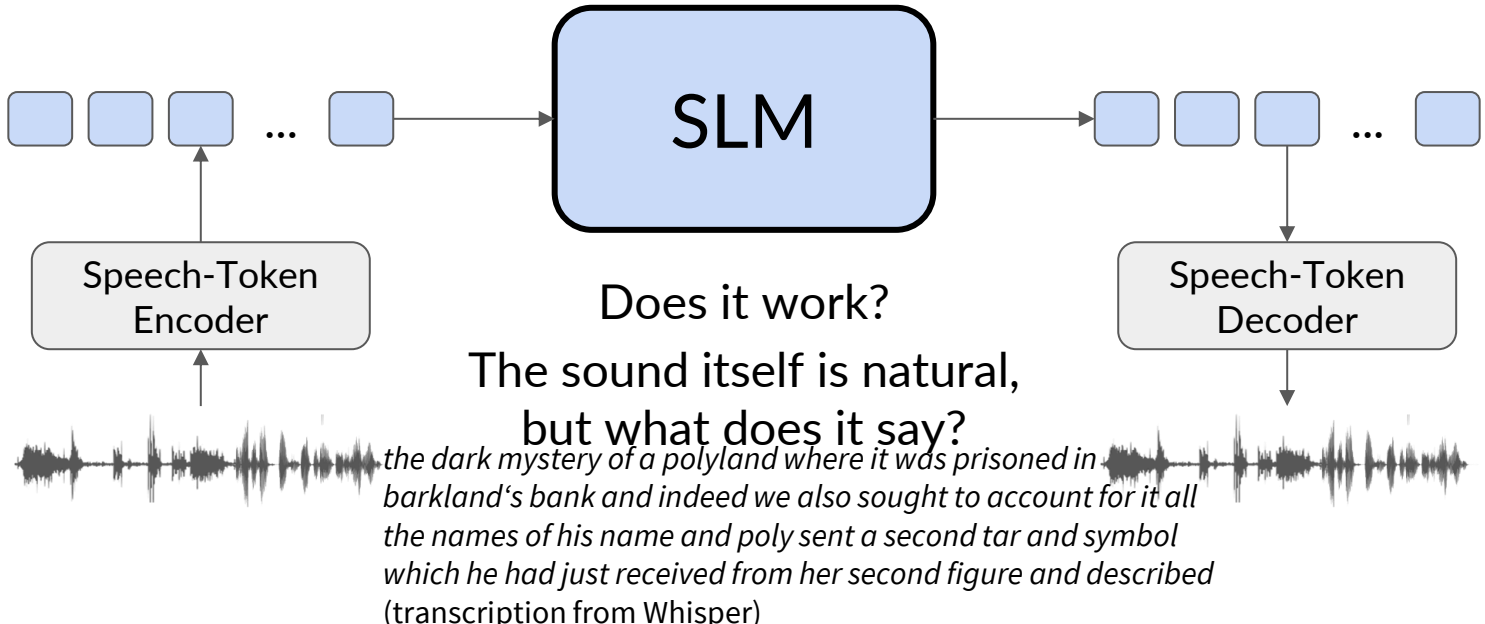
# **Background**: SLM

Inspired from the huge success of text language models, spoken language models (**SLM**) aim at generating speech by **modeling speech tokens**.



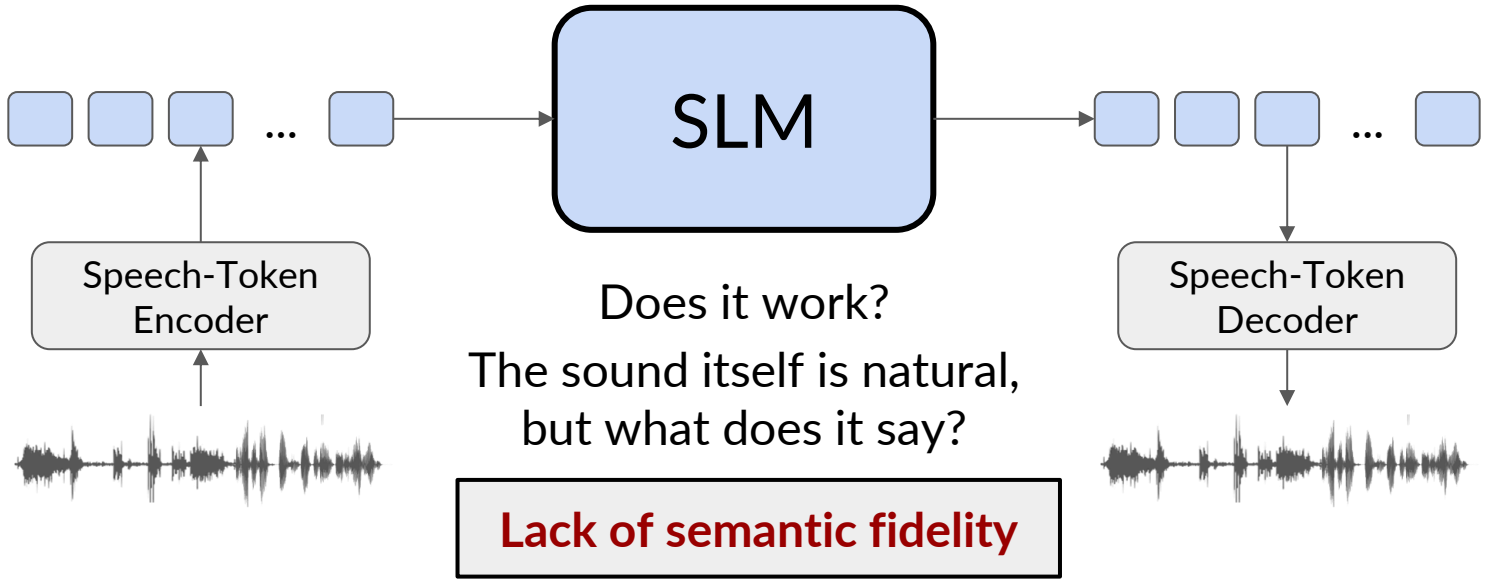
# **Background**: SLM

Inspired from the huge success of text language models, spoken language models (**SLM**) aim at generating speech by **modeling speech tokens**.



# **Background**: SLM

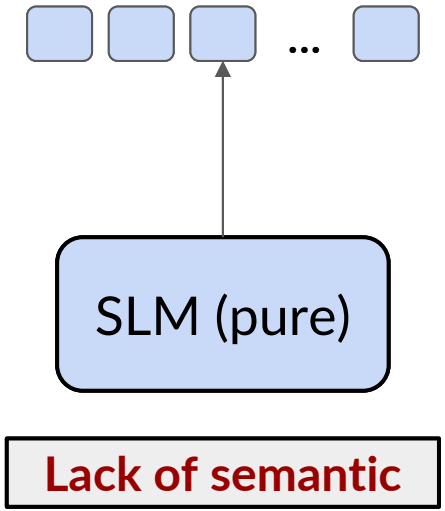
Inspired from the huge success of text language models, spoken language models (**SLM**) aim at generating speech by **modeling speech tokens**.



# **Background**: Improving Semantic

To improve semantic coherence, previous works have proposed **leveraging text**...

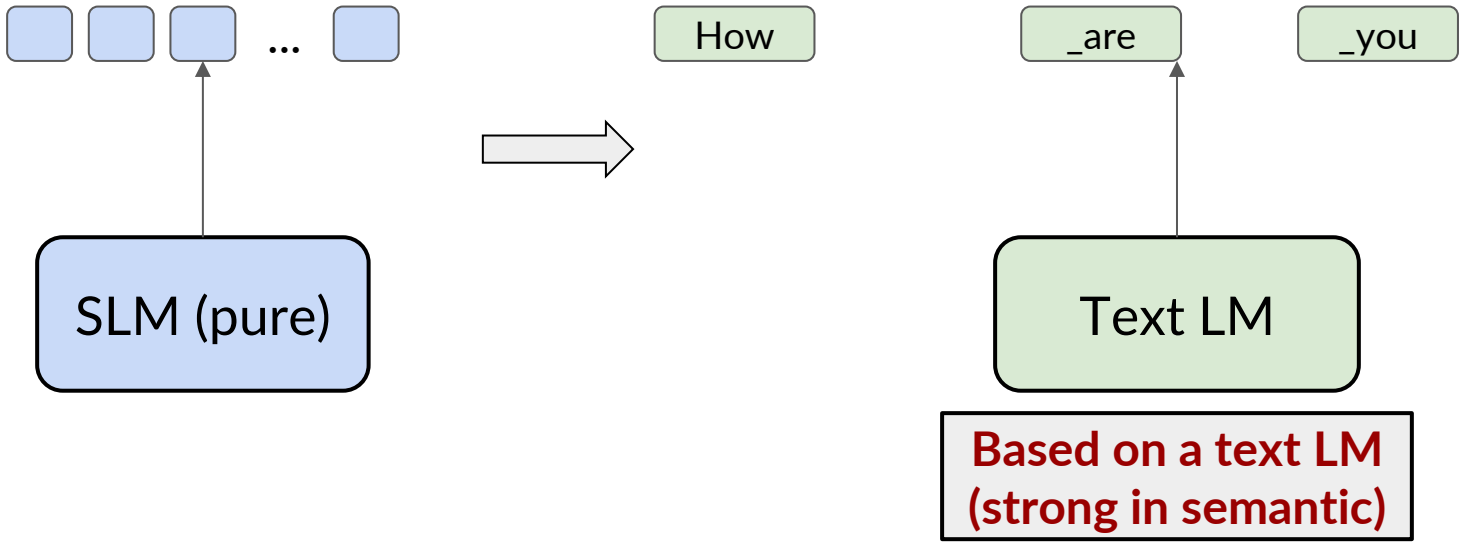
- Hybrid Text-Speech Generation (Mini-Omni; LLaMA-Omni; Moshi;...)



# **Background**: Improving Semantic

To improve semantic coherence, previous works have proposed **leveraging text**...

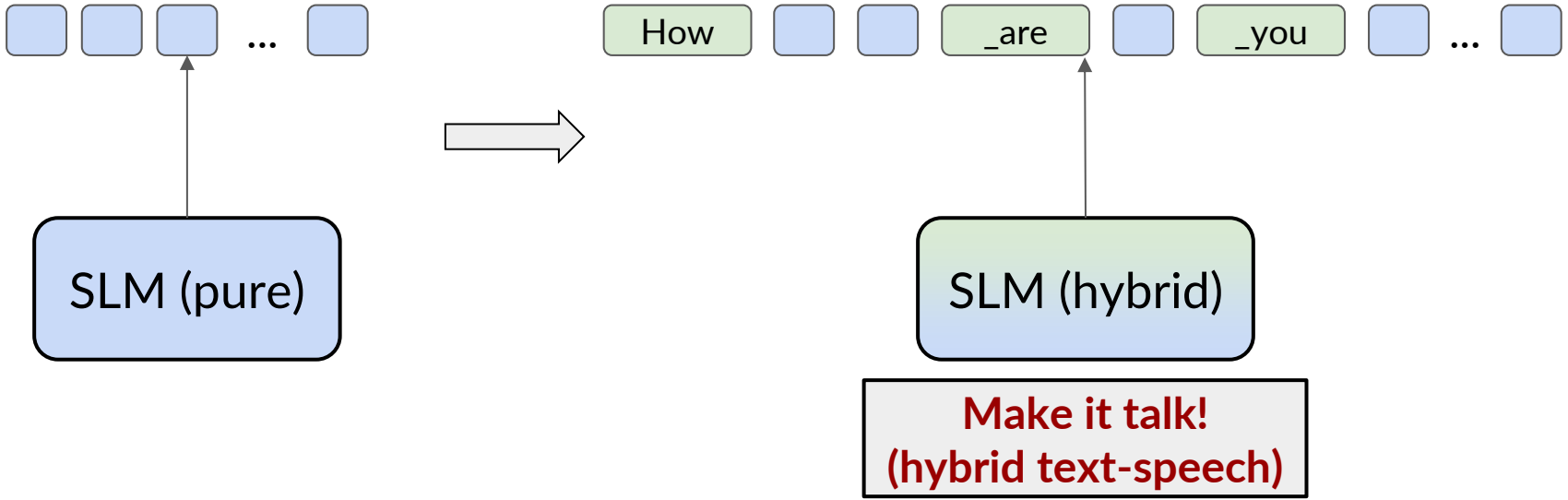
- Hybrid Text-Speech Generation (Mini-Omni; LLaMA-Omni; Moshi;...)



# **Background**: Improving Semantic

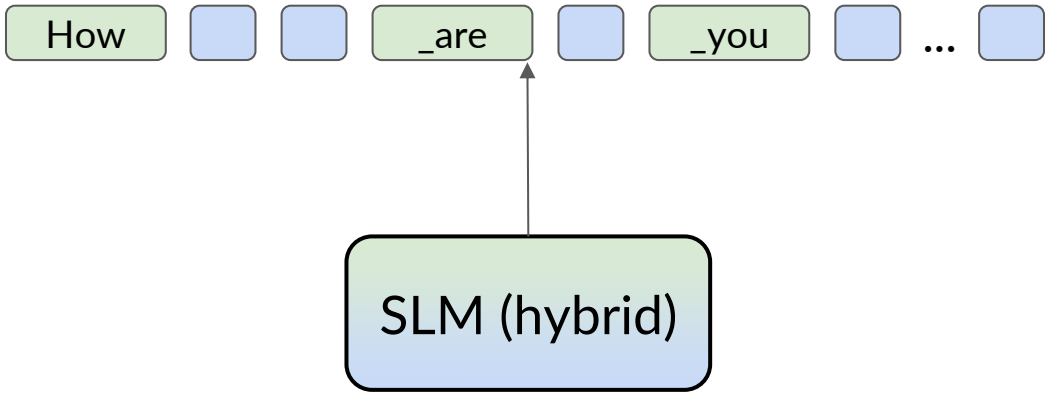
To improve semantic coherence, previous works have proposed **leveraging text**...

- Hybrid Text-Speech Generation (Mini-Omni; LLaMA-Omni; Moshi;...)



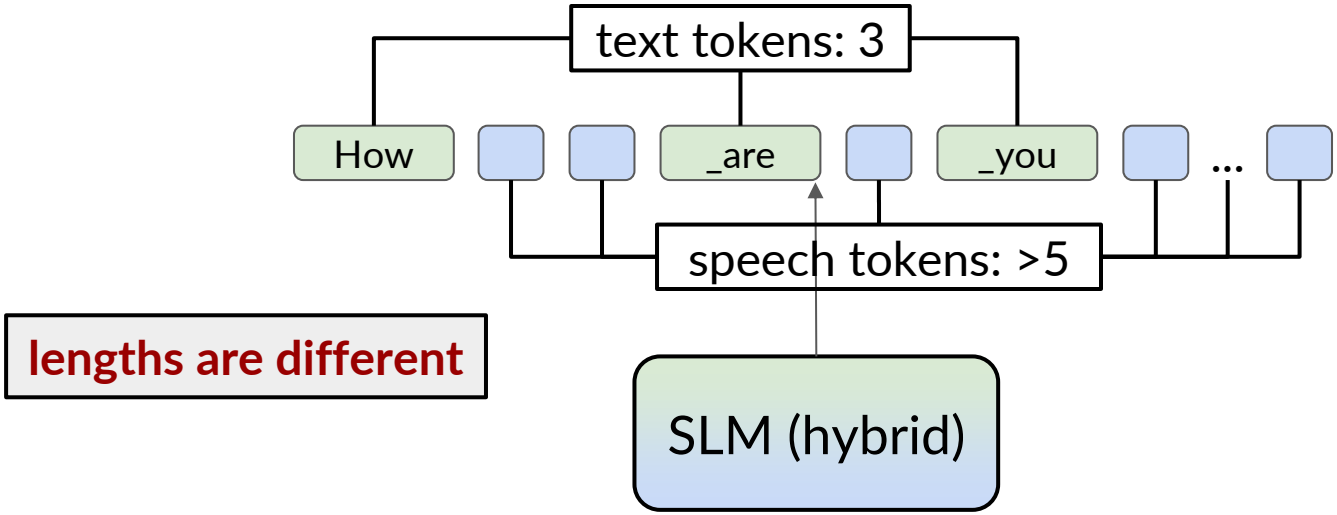
# **Background**: Hybrid Text-Speech Generation

But hybrid generation needs to tackle the modality mismatch problem.



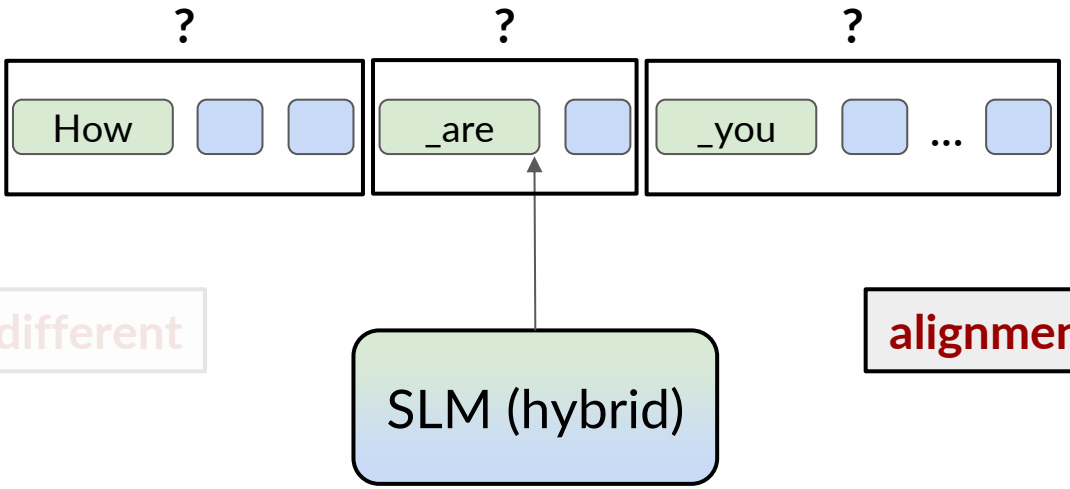
# **Background**: Hybrid Text-Speech Generation

But hybrid generation needs to tackle the modality mismatch problem.



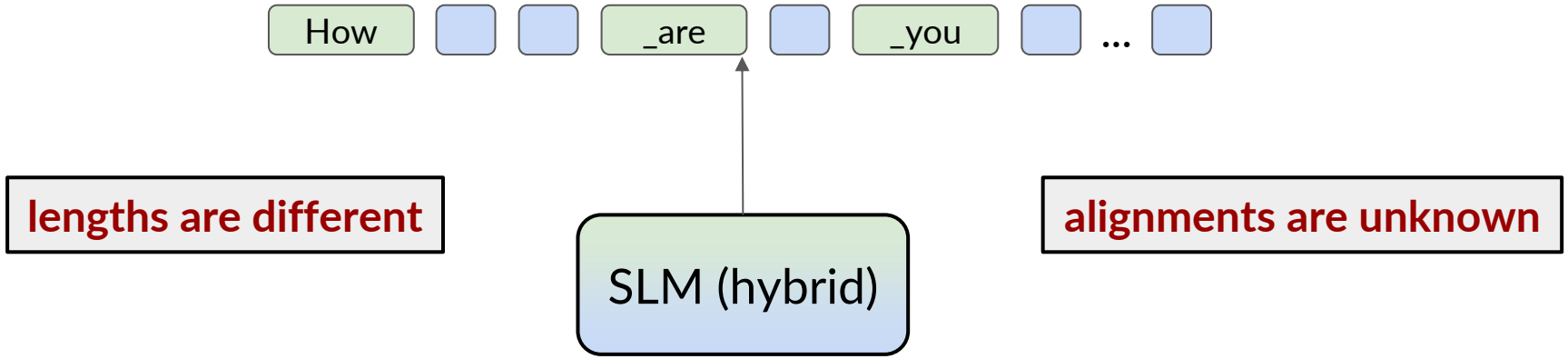
# **Background**: Hybrid Text-Speech Generation

But hybrid generation needs to tackle the modality mismatch problem.



# **Background**: Hybrid Text-Speech Generation

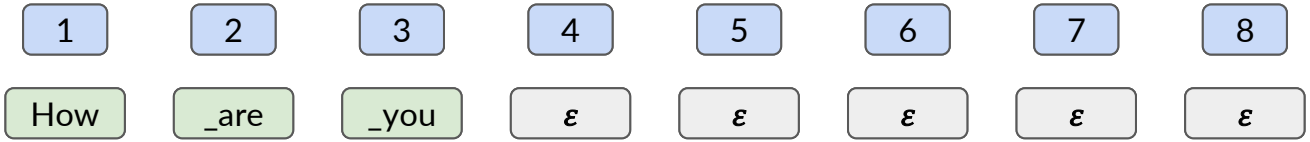
But hybrid generation needs to tackle the modality mismatch problem.



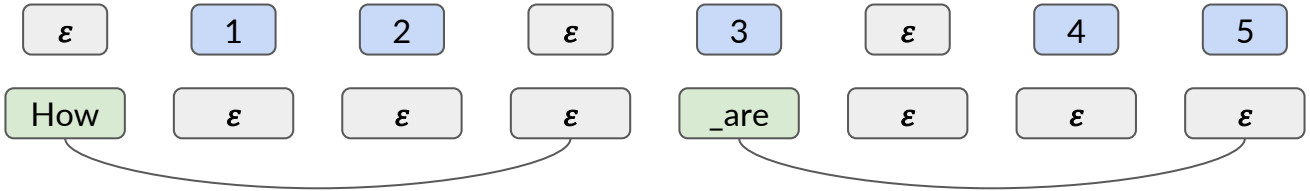
# **Background**: Hybrid Text-Speech Generation

Previous works may mitigate the problem by **adding paddings** ( ε )

- Mini-Omni (Zhifei Xie and Changqiao Wu)



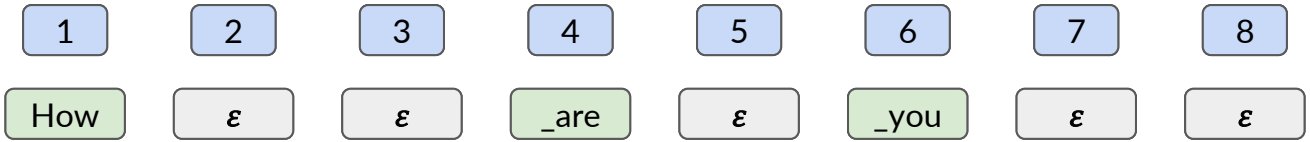

- LLaMA-Omni (Qingkai Fang, et al.)




CTC loss

fixed ratio (1:3)

- Moshi (Alexandre Défossez, et al.)



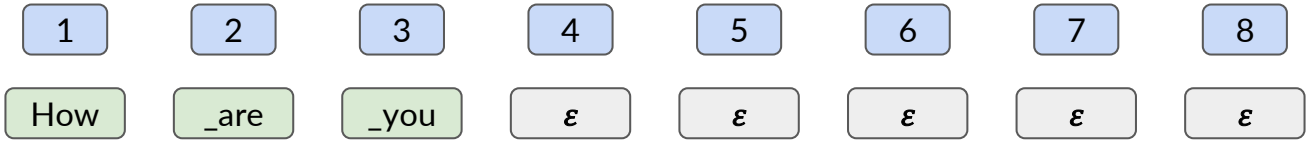

modeling the duration

ref: Prof. Hung-yi Lee's slides

# **Background**: Hybrid Text-Speech Generation

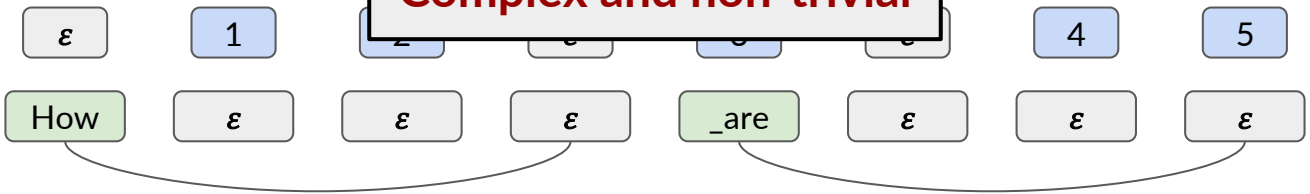
Previous works may mitigate the problem by **adding paddings** ( ε )

- Mini-Omni (Zhifei Xie and Changqiao Wu)




- LLaMA-Omni (Qingkai Fan)

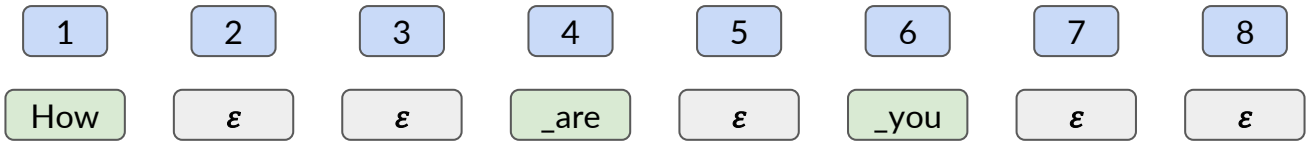
**Complex and non-trivial**




CTC loss

fixed ratio (1:3)

- Moshi (Alexandre Défossez, et al.)

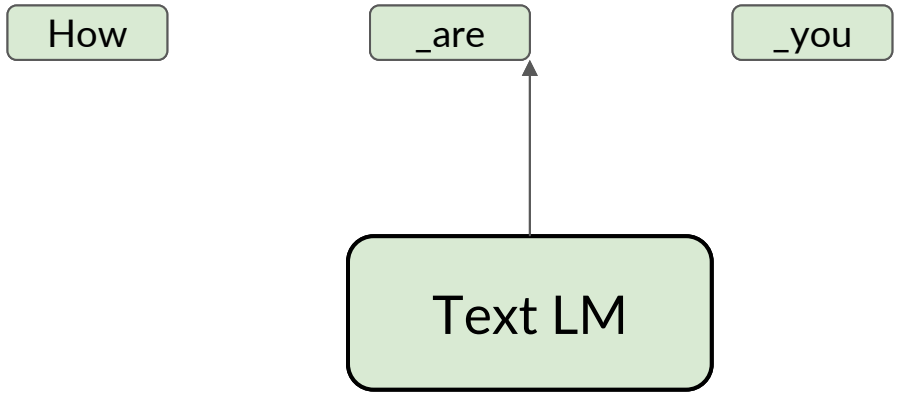



modeling the duration

ref: Prof. Hung-yi Lee's slides

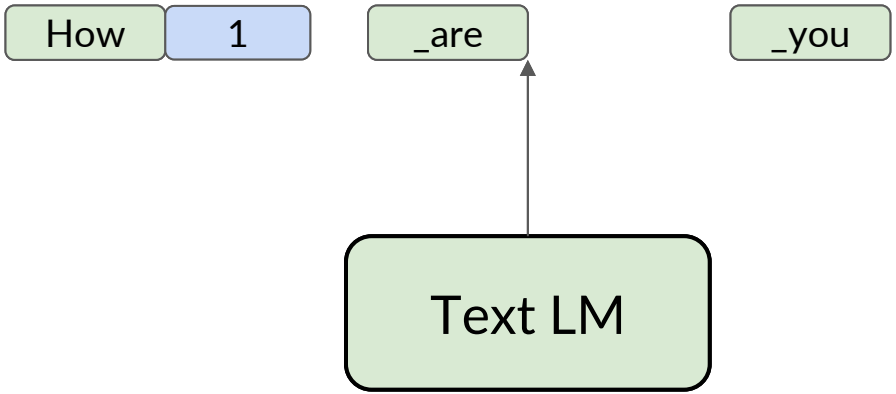
# **Introduction**: TASTE

Rethink that we are developing a hybrid SLM starting from a text LM



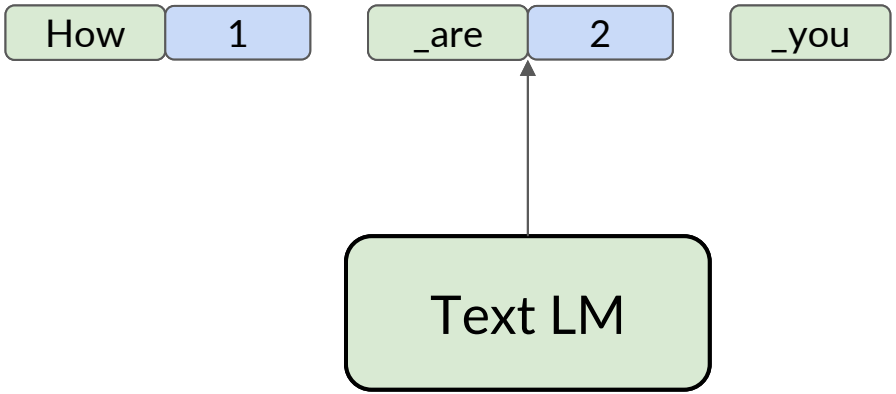
# **Introduction**: TASTE

Rethink that we are developing a hybrid SLM starting from a text LM
⇒ Can we teach the text LM how to speak out each word by using the additional speech tokens attached?



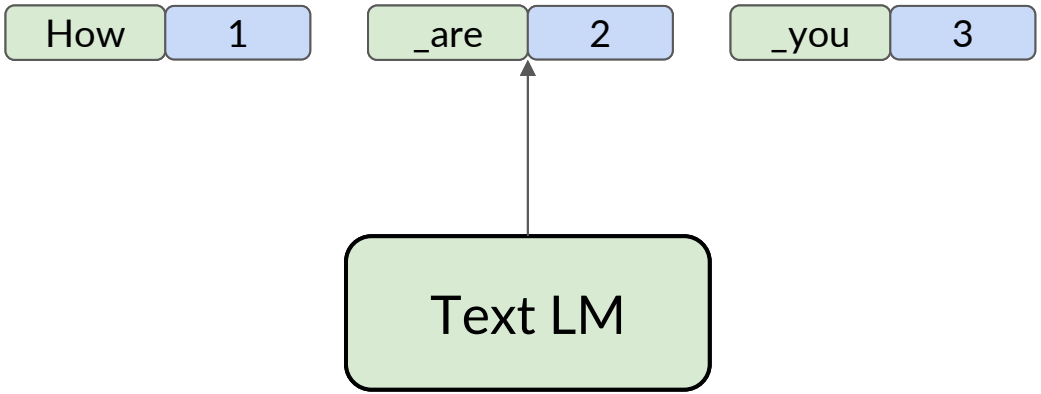
# **Introduction**: TASTE

Rethink that we are developing a hybrid SLM starting from a text LM
⇒ Can we teach the text LM how to speak out each word by using the additional speech tokens attached?



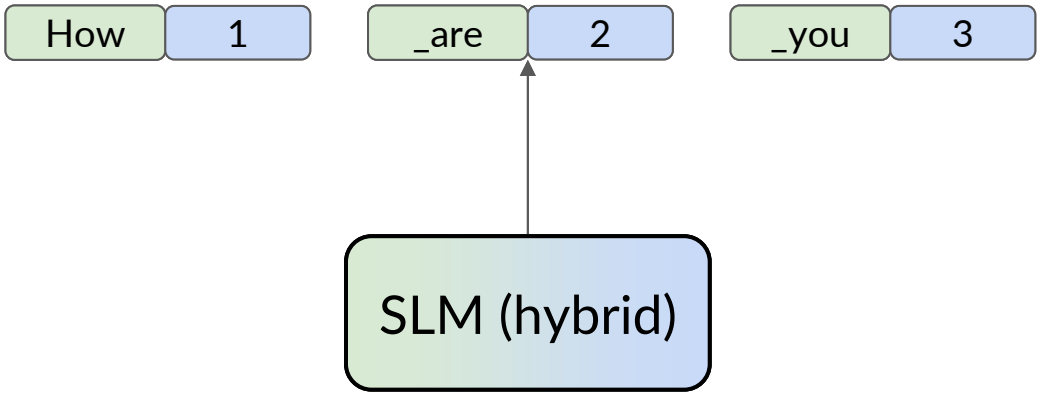
# **Introduction**: TASTE

Rethink that we are developing a hybrid SLM starting from a text LM
⇒ Can we teach the text LM how to speak out each word by using the additional speech tokens attached?



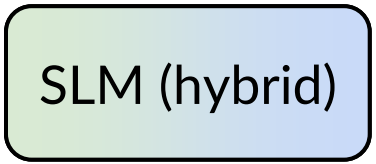
# **Introduction**: TASTE

Rethink that we are developing a hybrid SLM starting from a text LM
⇒ Can we teach the text LM how to speak out each word by using the additional speech tokens attached?



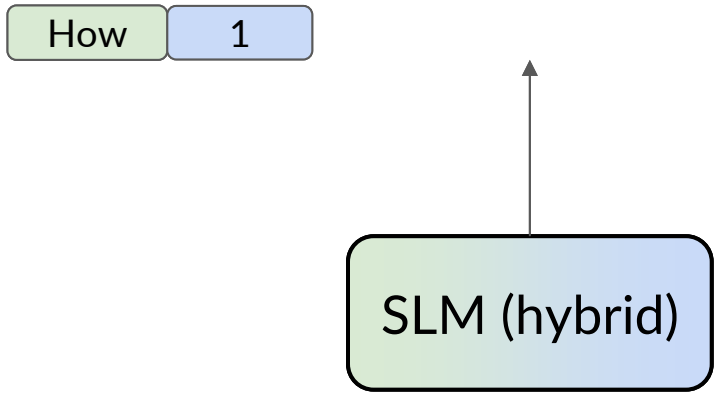
# **Introduction**: TASTE

The hybrid generation and modeling is now trivial and simple.



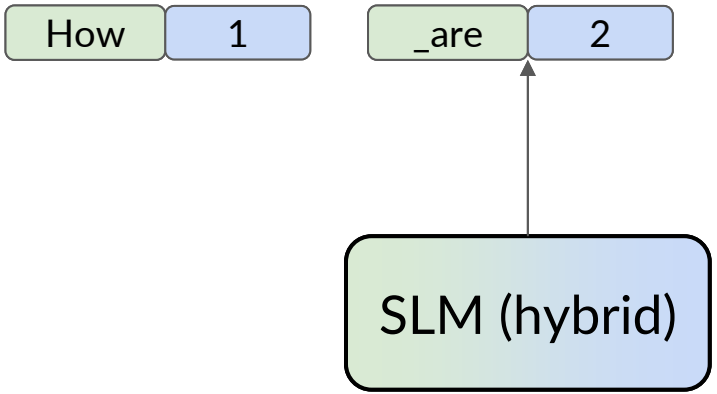
# **Introduction**: TASTE

The hybrid generation and modeling is now trivial and simple.



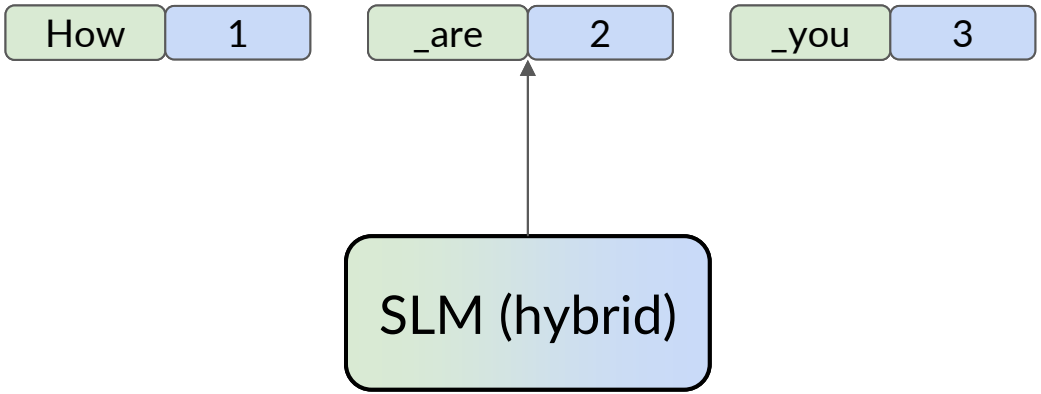
# **Introduction**: TASTE

The hybrid generation and modeling is now trivial and simple.



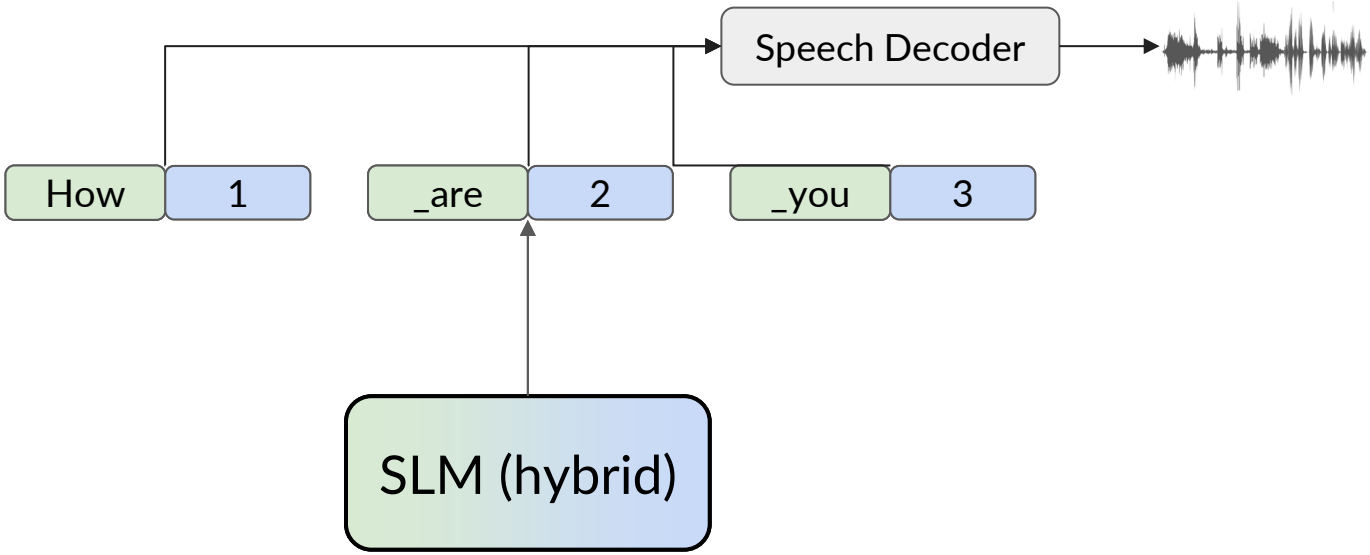
# **Introduction**: TASTE

The hybrid generation and modeling is now trivial and simple.



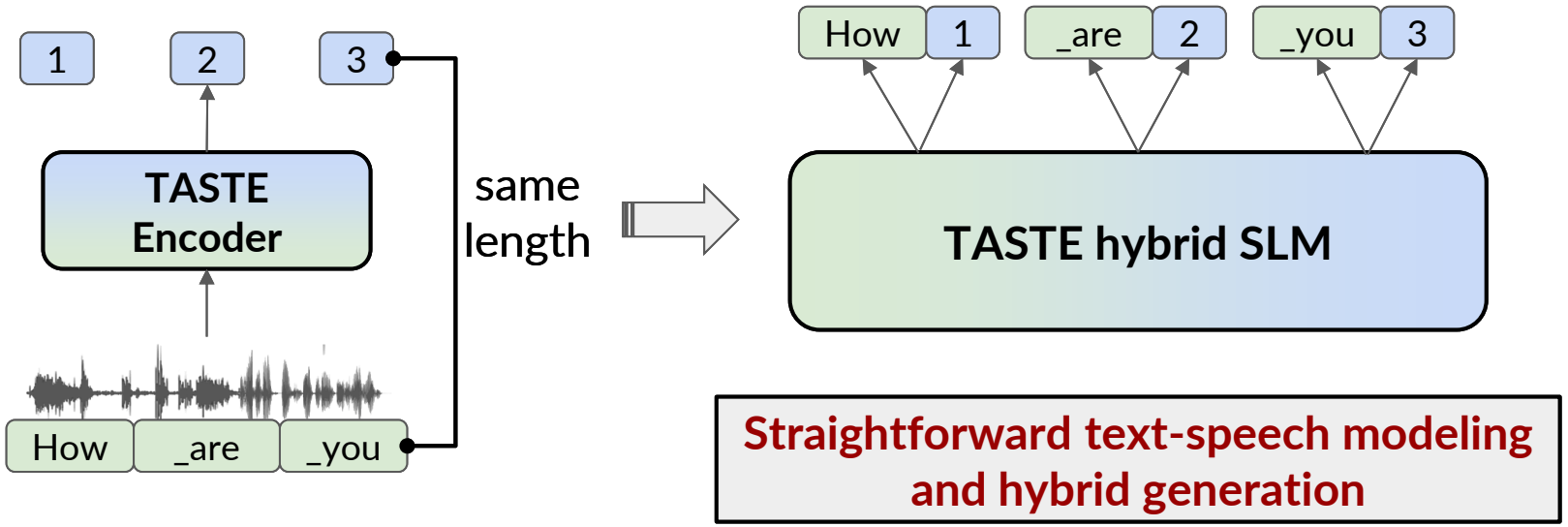
# **Introduction**: TASTE

The hybrid generation and modeling is now trivial and simple.



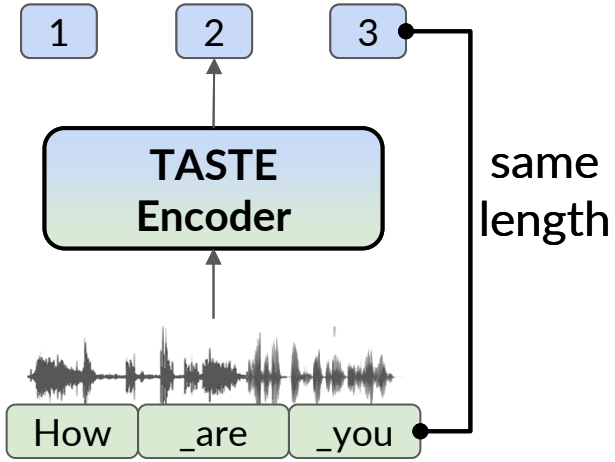
# **Introduction**: TASTE

We propose **T**ext-**A**ligned **S**peech **T**okenization and **E**mbedding, a specialized speech tokenization tailored for the hybrid text-speech SLM.



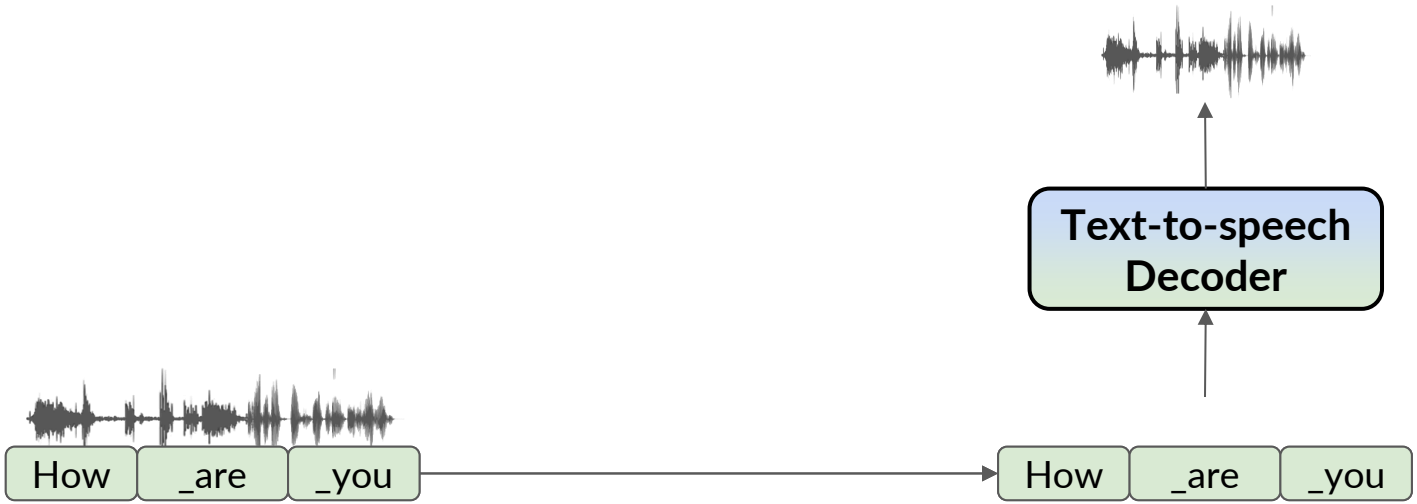
# **Introduction**: TASTE

We propose **T**ext-**A**ligned **S**peech **T**okenization and **E**mbedding, a specialized speech tokenization tailored for the hybrid text-speech SLM.



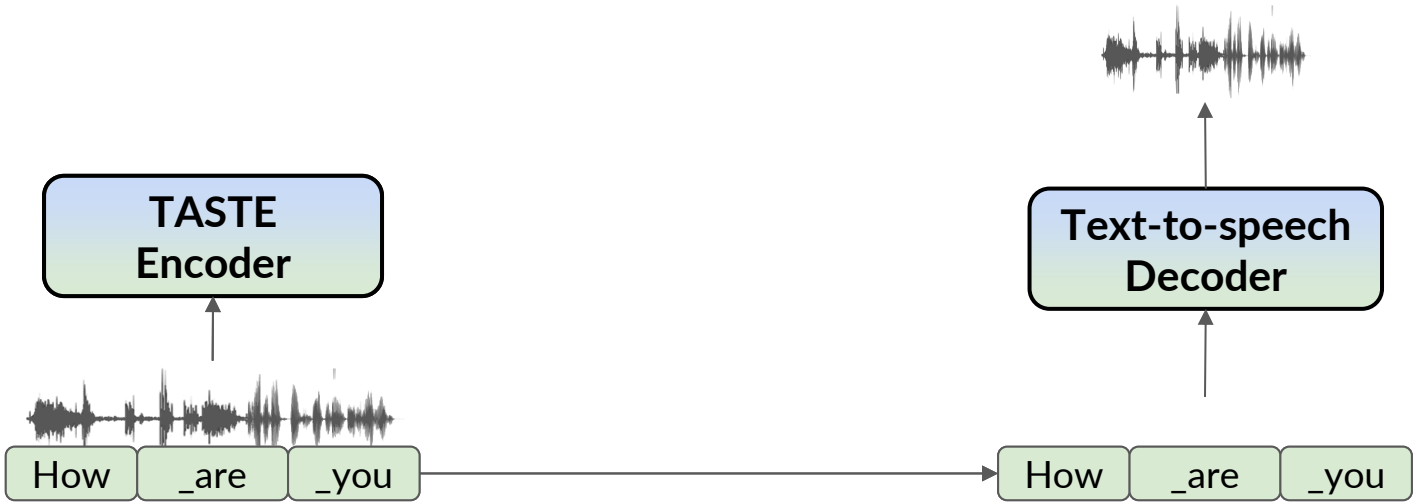
# **Method**: Reconstruction as Learning Objective

To make the TASTE token convey paralinguistic information for generative modeling, we adopt **speech reconstruction** as the learning objective.



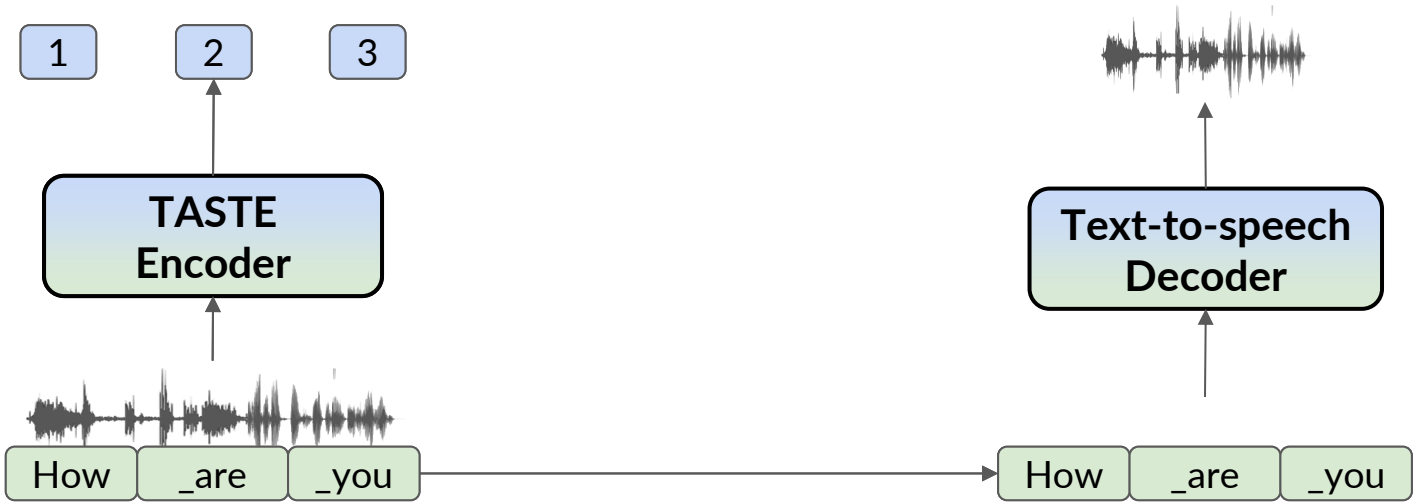
# **Method**: Reconstruction as Learning Objective

To make the TASTE token convey paralinguistic information for generative modeling, we adopt **speech reconstruction** as the learning objective.



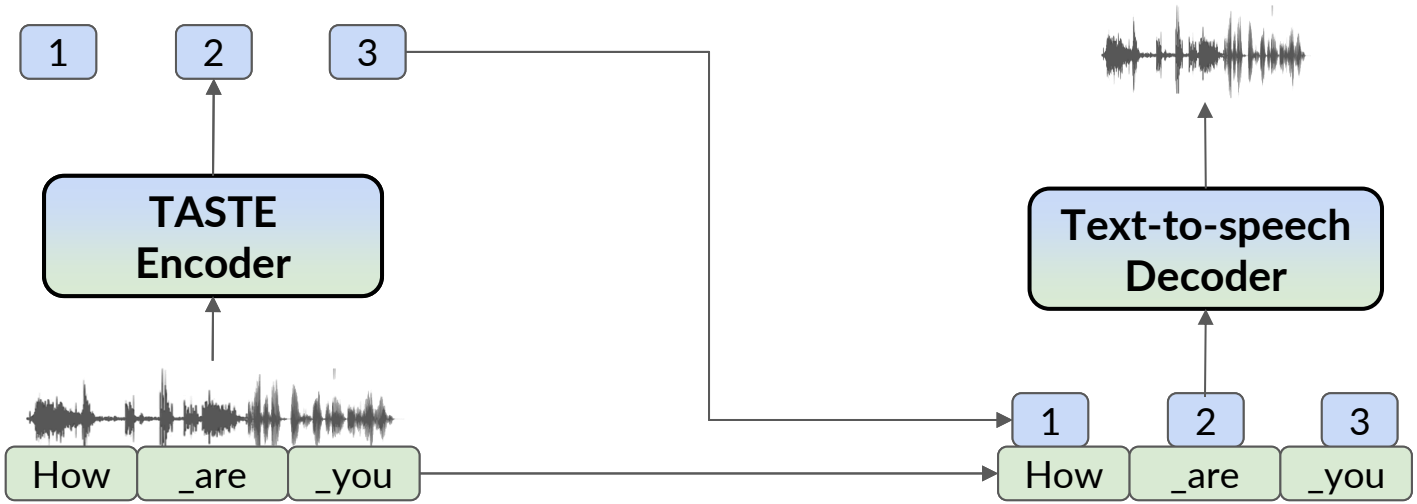
# **Method**: Reconstruction as Learning Objective

To make the TASTE token convey paralinguistic information for generative modeling, we adopt **speech reconstruction** as the learning objective.



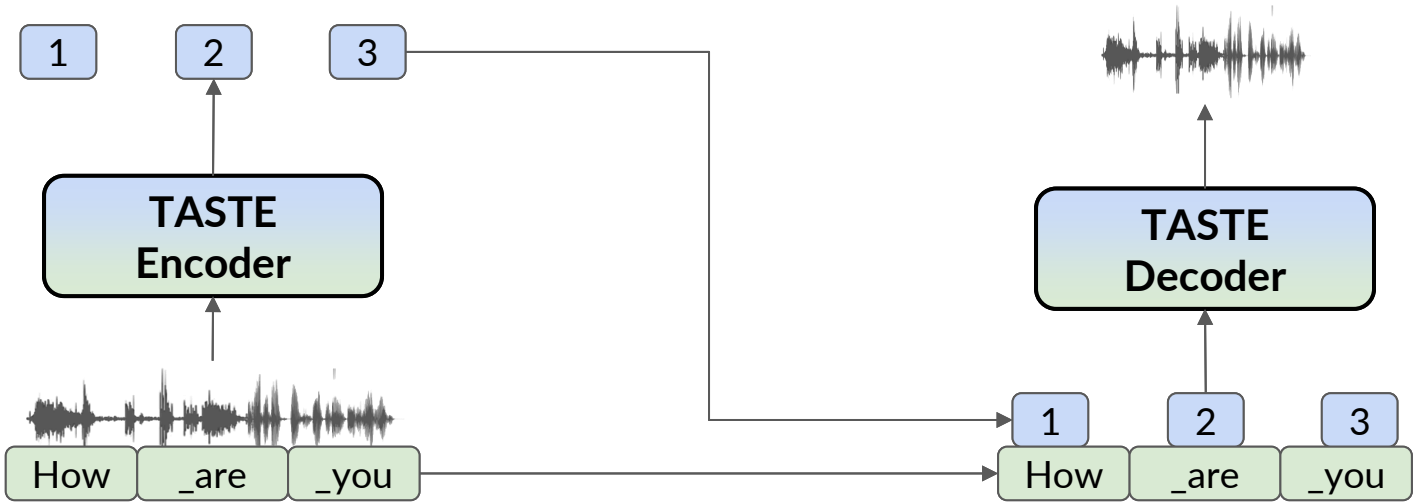
# **Method**: Reconstruction as Learning Objective

To make the TASTE token convey paralinguistic information for generative modeling, we adopt **speech reconstruction** as the learning objective.



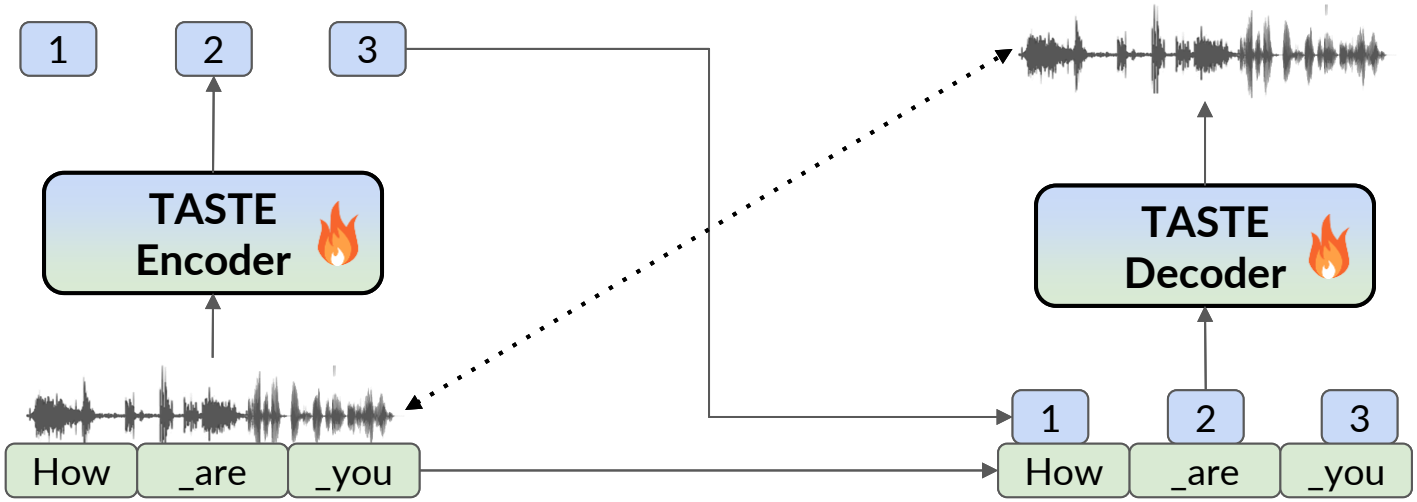
# **Method**: Reconstruction as Learning Objective

To make the TASTE token convey paralinguistic information for generative modeling, we adopt **speech reconstruction** as the learning objective.



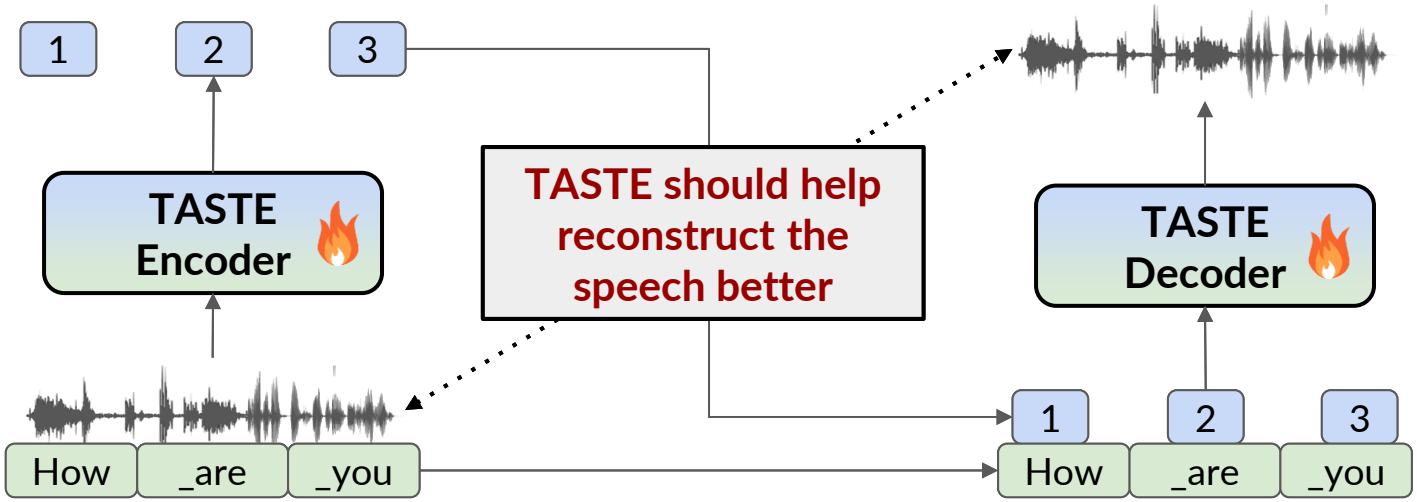
# **Method**: Reconstruction as Learning Objective

To make the TASTE token convey paralinguistic information for generative modeling, we adopt **speech reconstruction** as the learning objective.



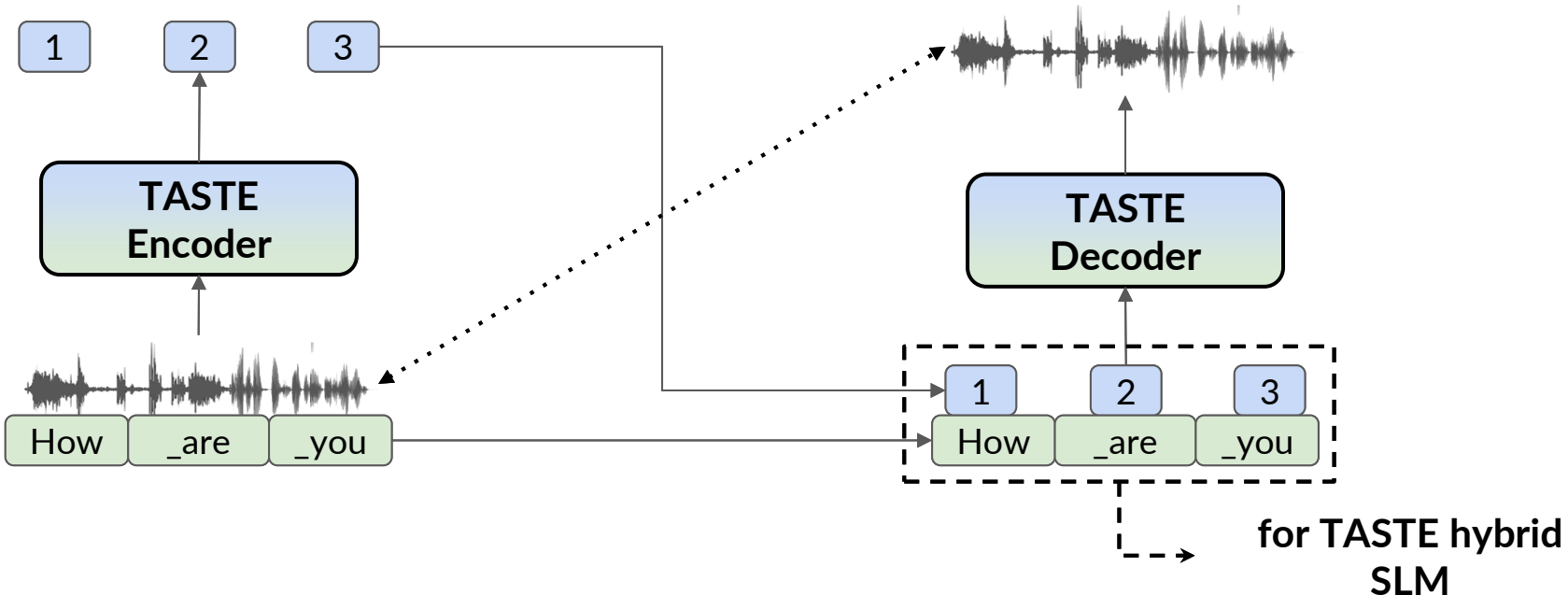
# **Method**: Reconstruction as Learning Objective

To make the TASTE token convey paralinguistic information for generative modeling, we adopt **speech reconstruction** as the learning objective.



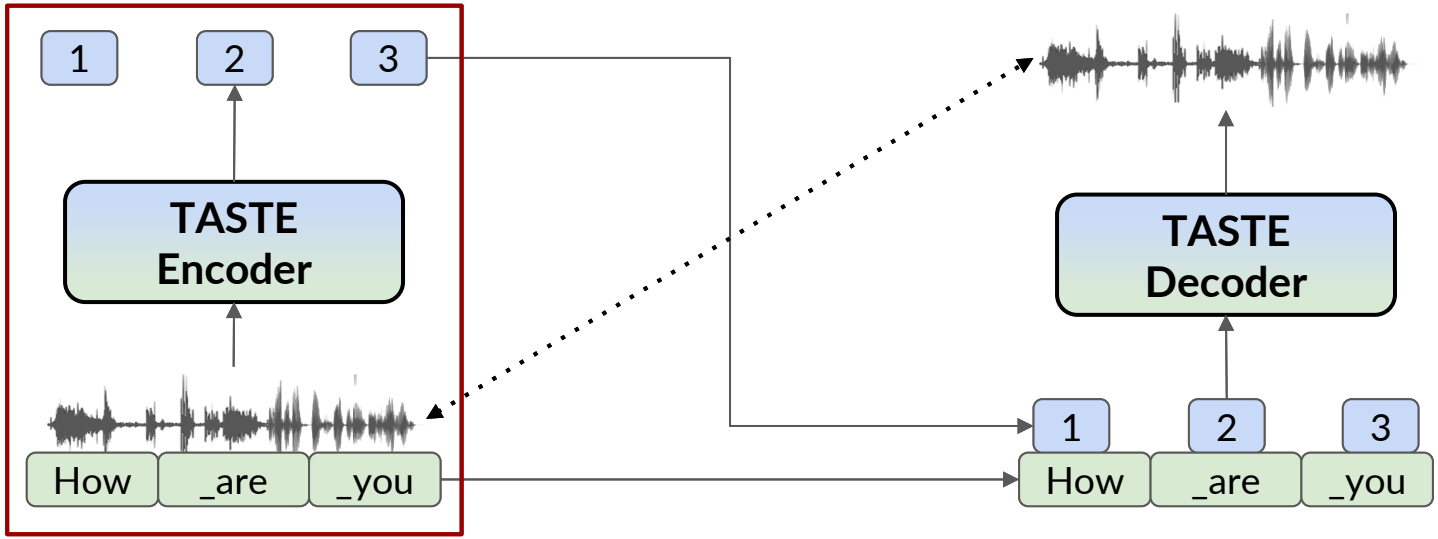
# **Method**: Reconstruction as Learning Objective

To make the TASTE token convey paralinguistic information for generative modeling, we adopt **speech reconstruction** as the learning objective.



# **Method**: Reconstruction as Learning Objective

To make the TASTE token convey paralinguistic information for generative modeling, we adopt **speech reconstruction** as the learning objective.



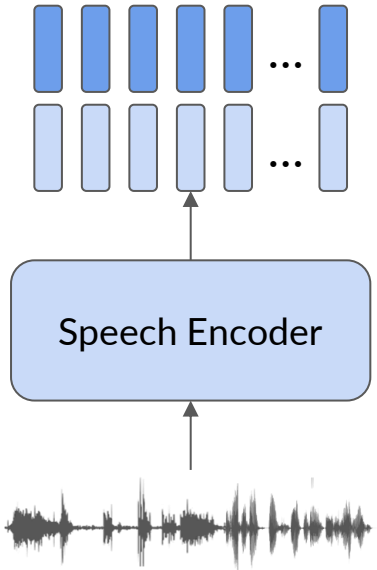

**How to extract the speech tokens that match the text's length?**

# **Method**: Cross-Attention for Lengths Matching

How to match the length of the speech features towards the text tokens' length?
→ by applying **cross-attention**!

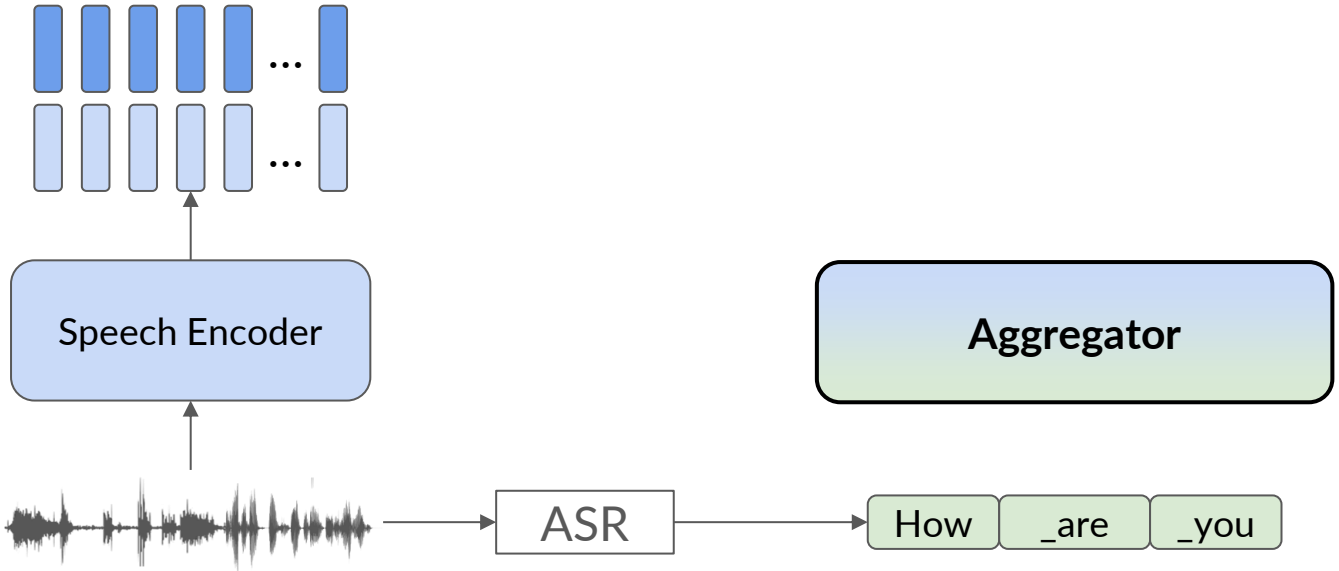
# **Method**: Cross-Attention for Lengths Matching

How to match the length of the speech features towards the text tokens' length?
→ by applying **cross-attention**!



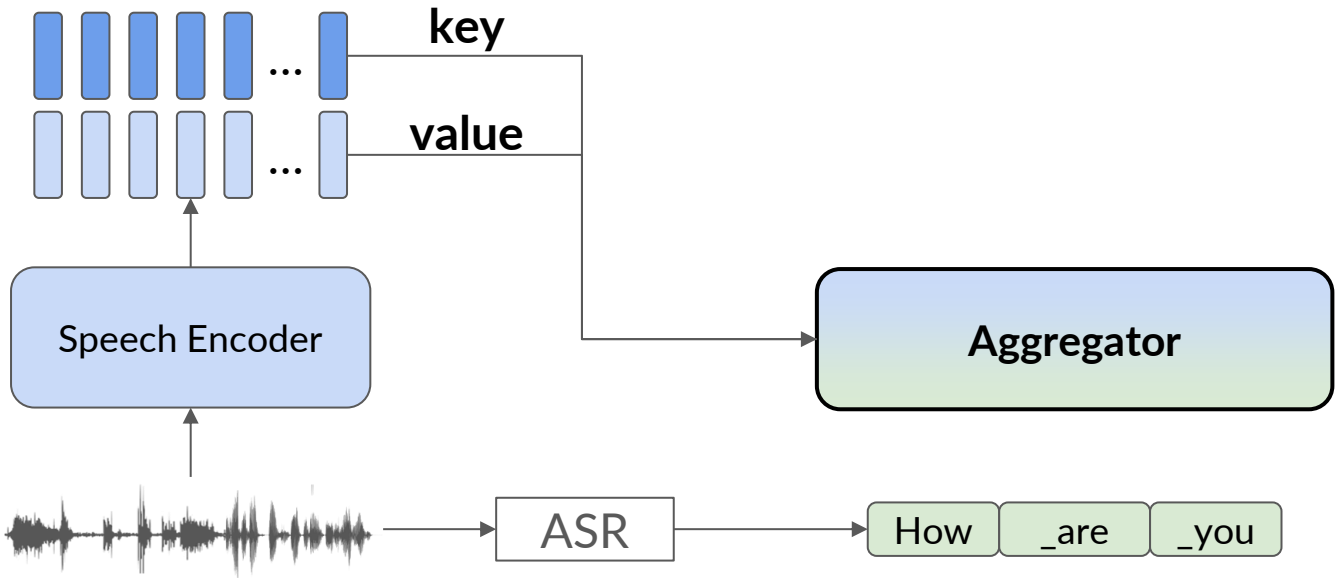

# **Method**: Cross-Attention for Lengths Matching

How to match the length of the speech features towards the text tokens' length?
→ by applying **cross-attention**!



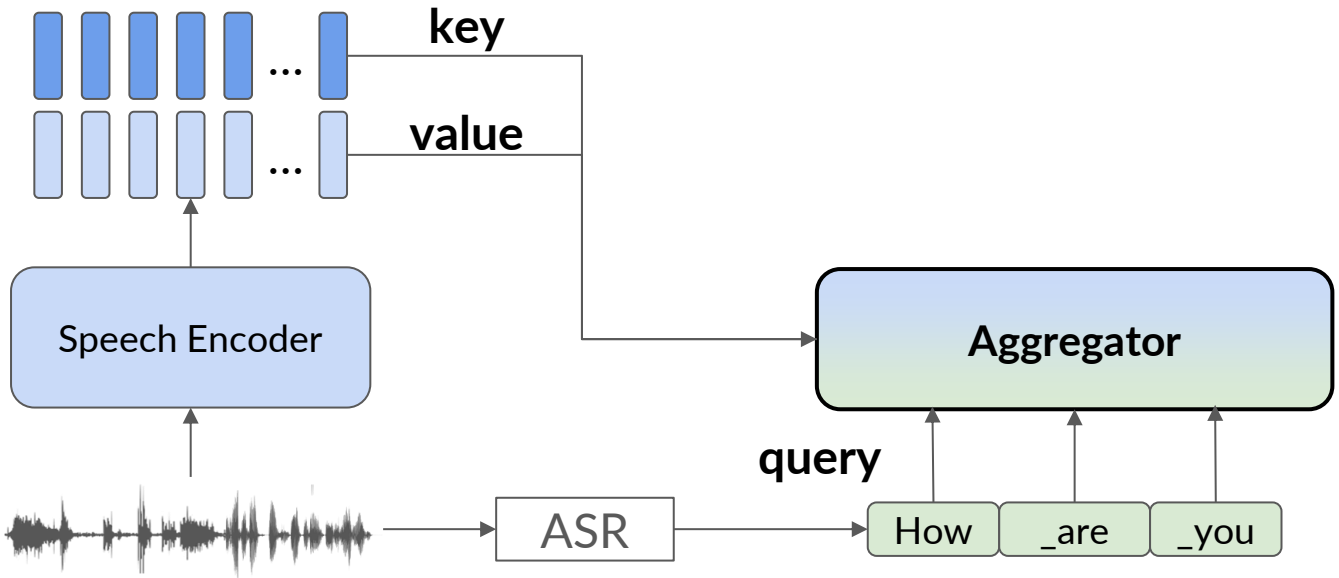
# **Method**: Cross-Attention for Lengths Matching

How to match the length of the speech features towards the text tokens' length?
→ by applying **cross-attention**!



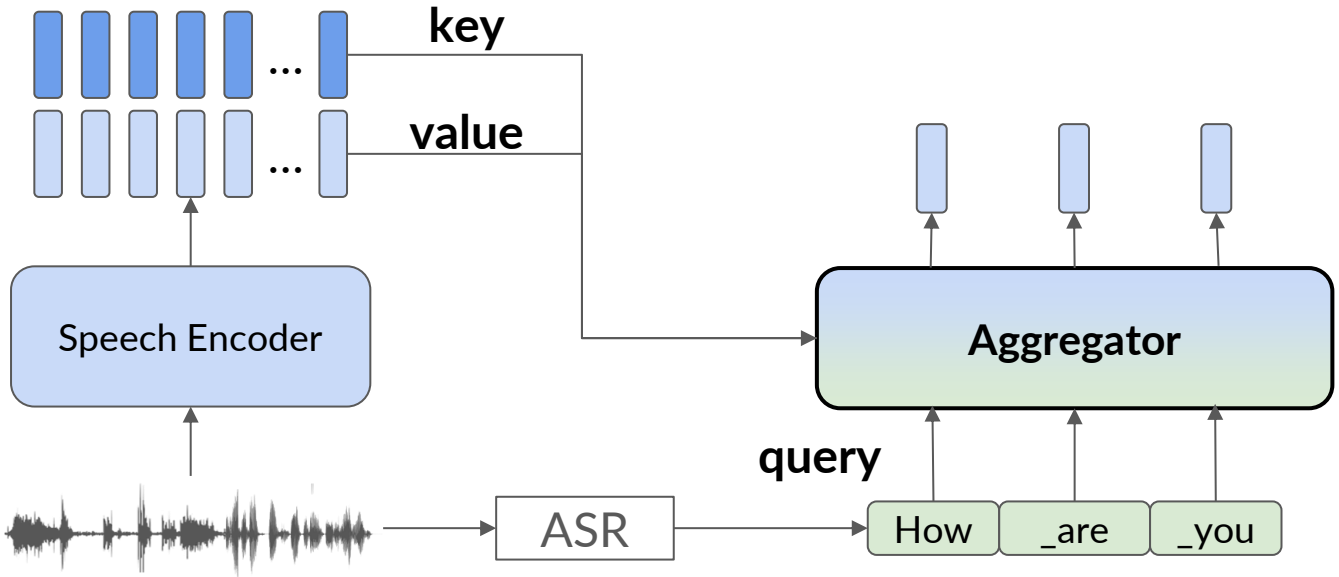
# **Method**: Cross-Attention for Lengths Matching

How to match the length of the speech features towards the text tokens' length?
→ by applying **cross-attention**!



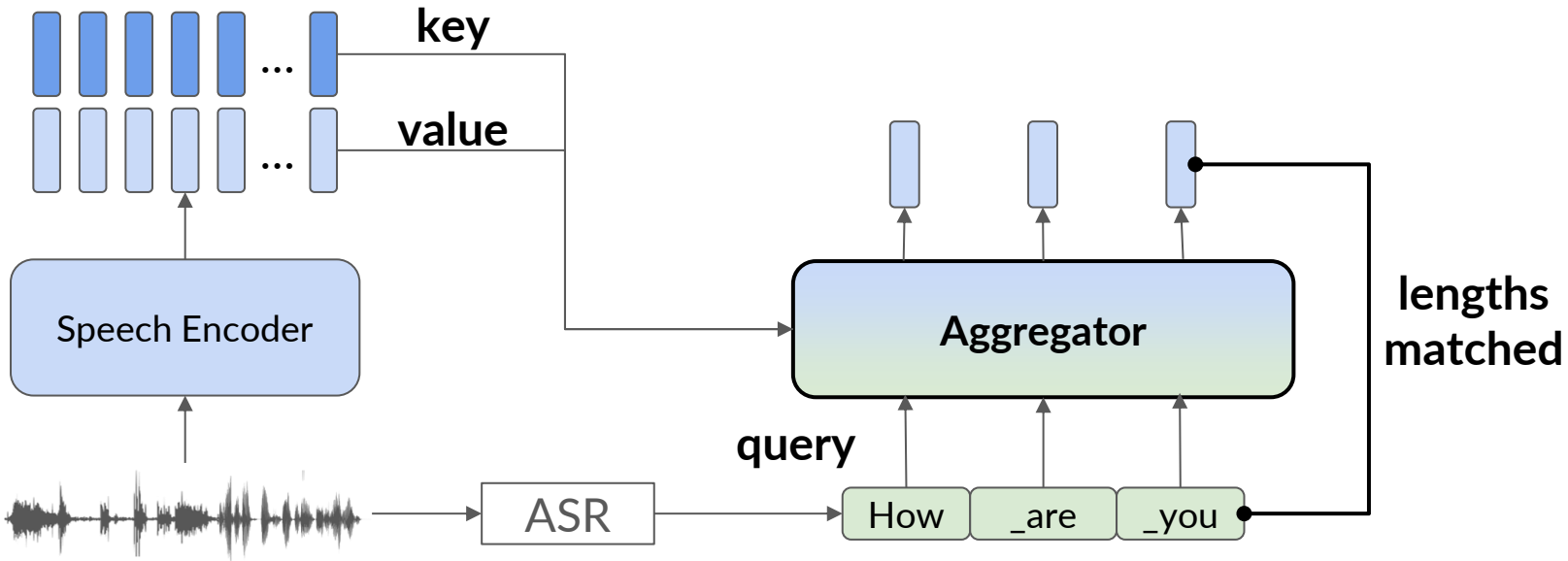
# **Method**: Cross-Attention for Lengths Matching

How to match the length of the speech features towards the text tokens' length?
→ by applying **cross-attention**!



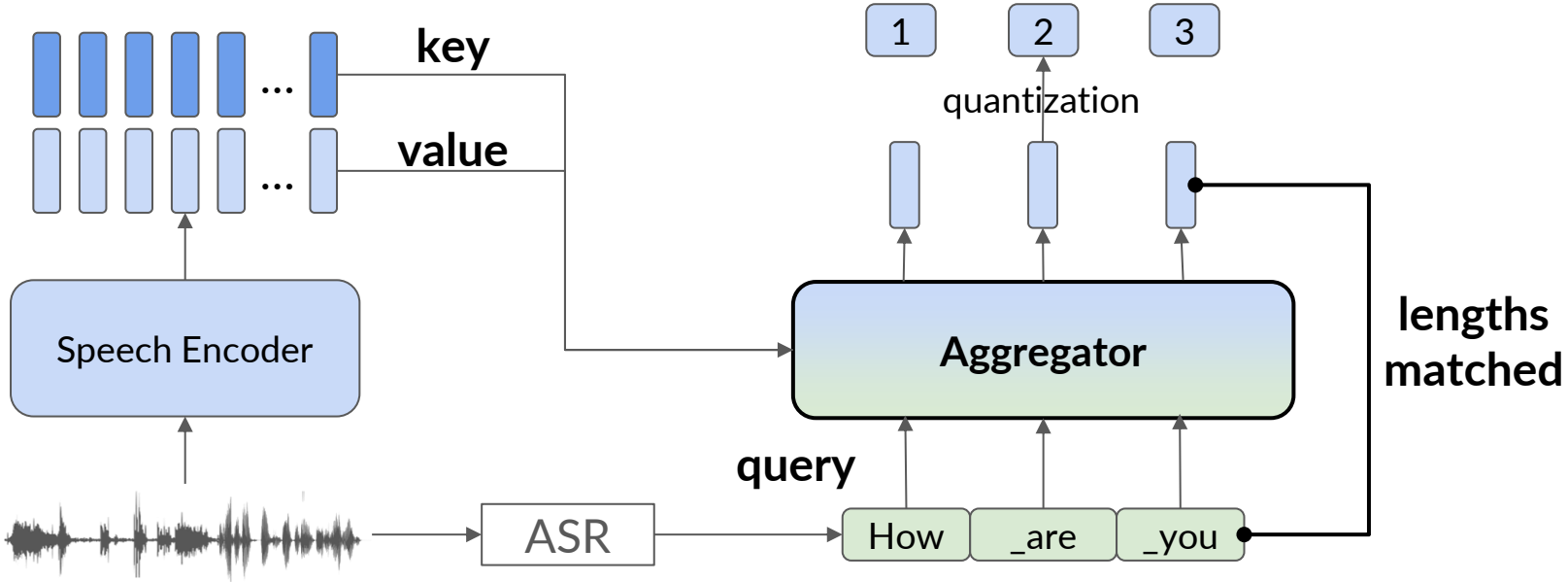
# **Method**: Cross-Attention for Lengths Matching

How to match the length of the speech features towards the text tokens' length?
→ by applying **cross-attention**!



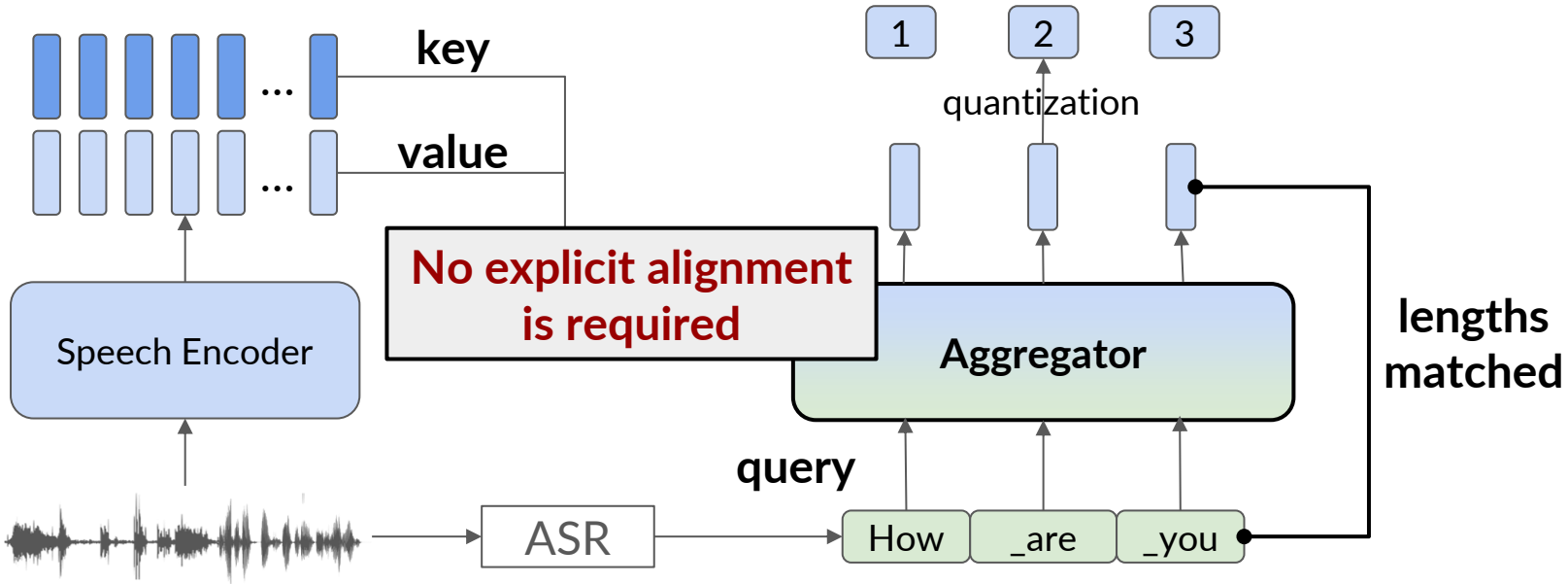
# **Method**: Cross-Attention for Lengths Matching

How to match the length of the speech features towards the text tokens' length?
→ by applying **cross-attention**!



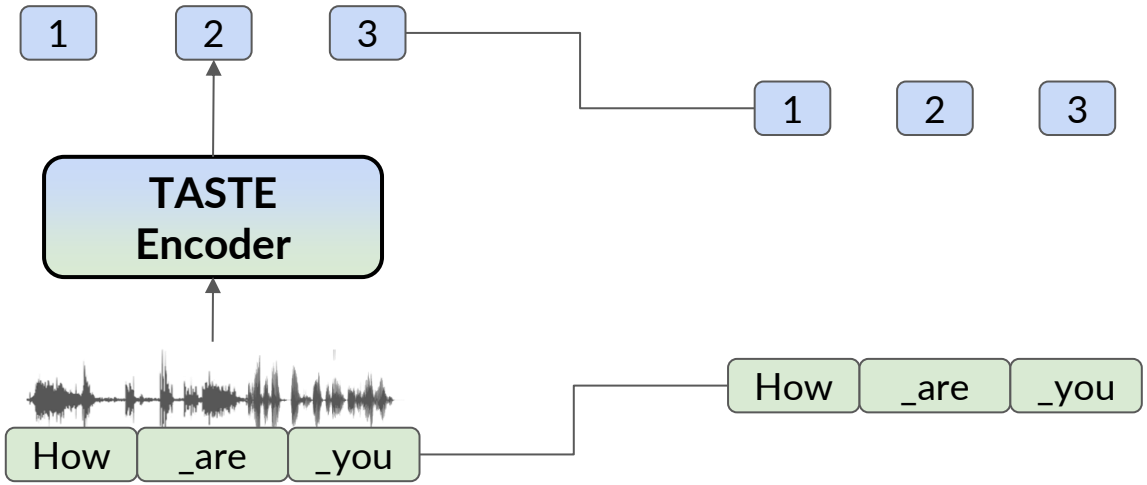
# **Method**: Cross-Attention for Lengths Matching

How to match the length of the speech features towards the text tokens' length?
→ by applying **cross-attention**!



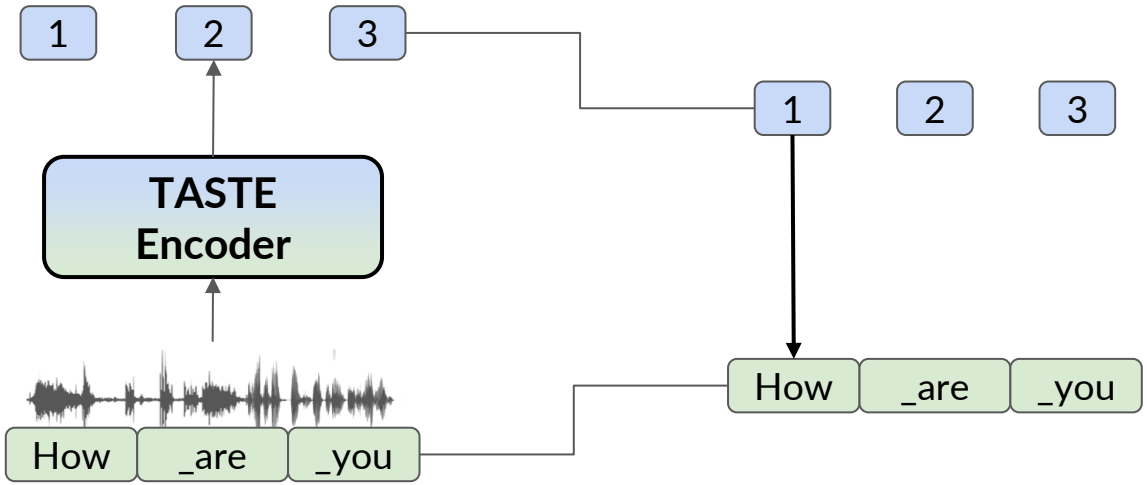
# **Result**: A Study on TASTE Tokenization

Does TASTE speech tokens really attached to their corresponding text tokens?



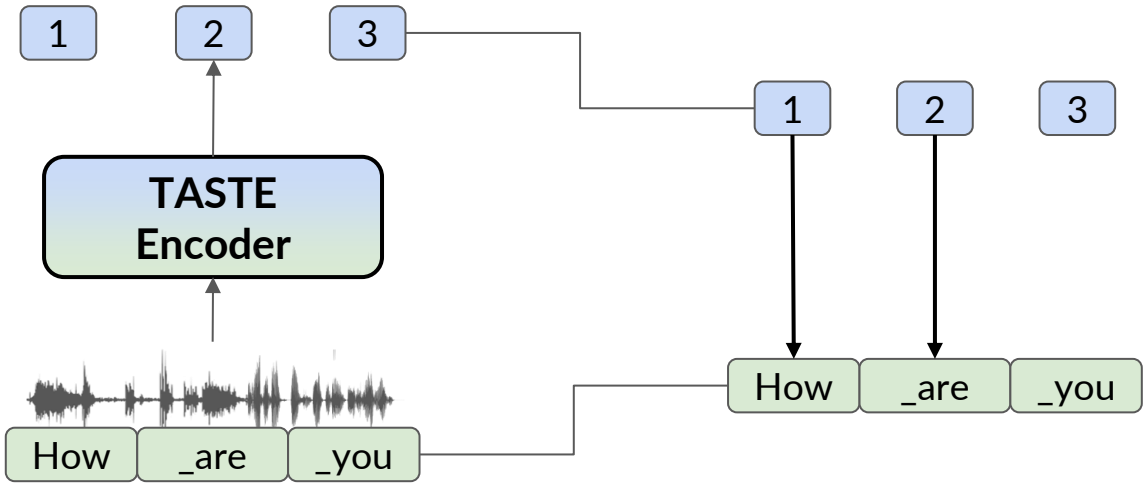
# **Result**: A Study on TASTE Tokenization

Does TASTE speech tokens really attached to their corresponding text tokens?



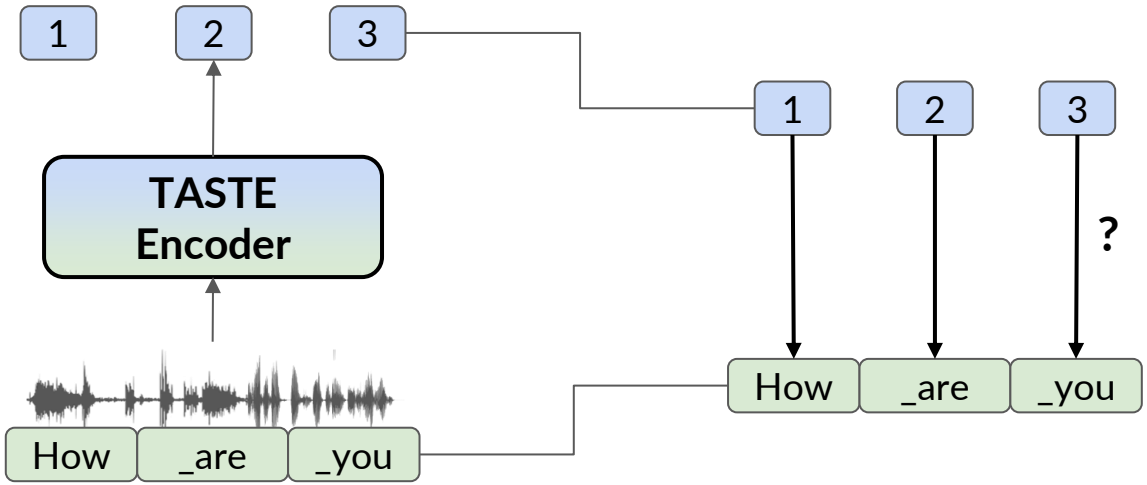
# **Result**: A Study on TASTE Tokenization

Does TASTE speech tokens really attached to their corresponding text tokens?



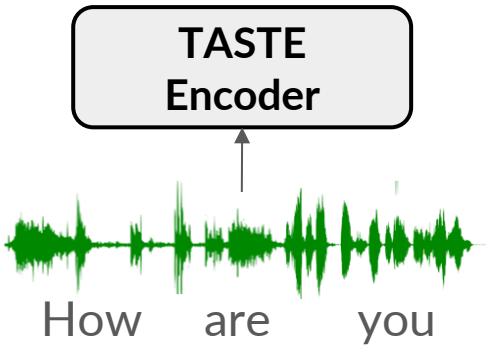
# **Result**: A Study on TASTE Tokenization

Does TASTE speech tokens really attached to their corresponding text tokens?



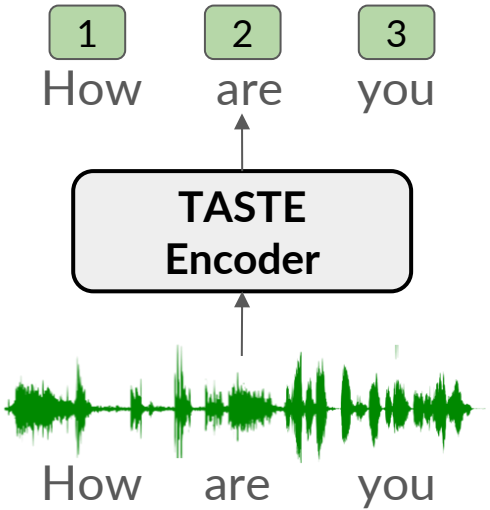
# **Result**: A Study on TASTE Tokenization

We conduct token swap to verify the attached behavior



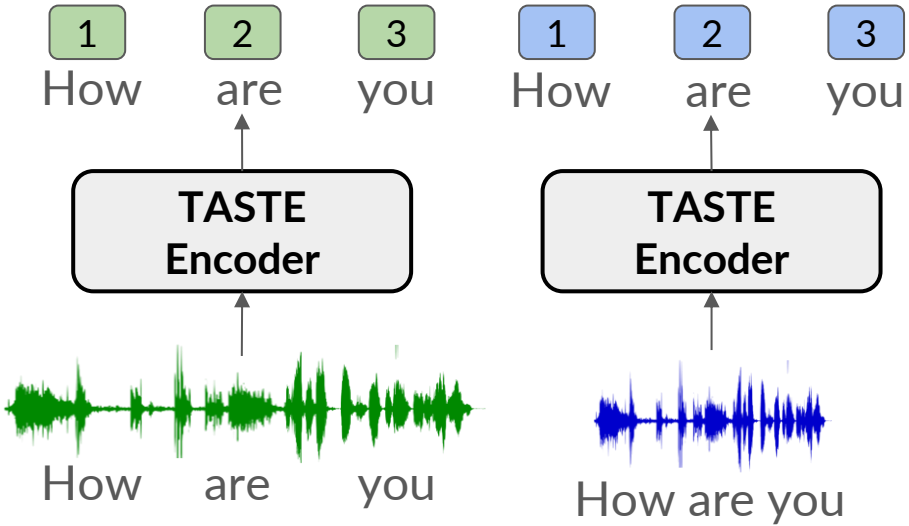
# **Result**: A Study on TASTE Tokenization

We conduct token swap to verify the attached behavior



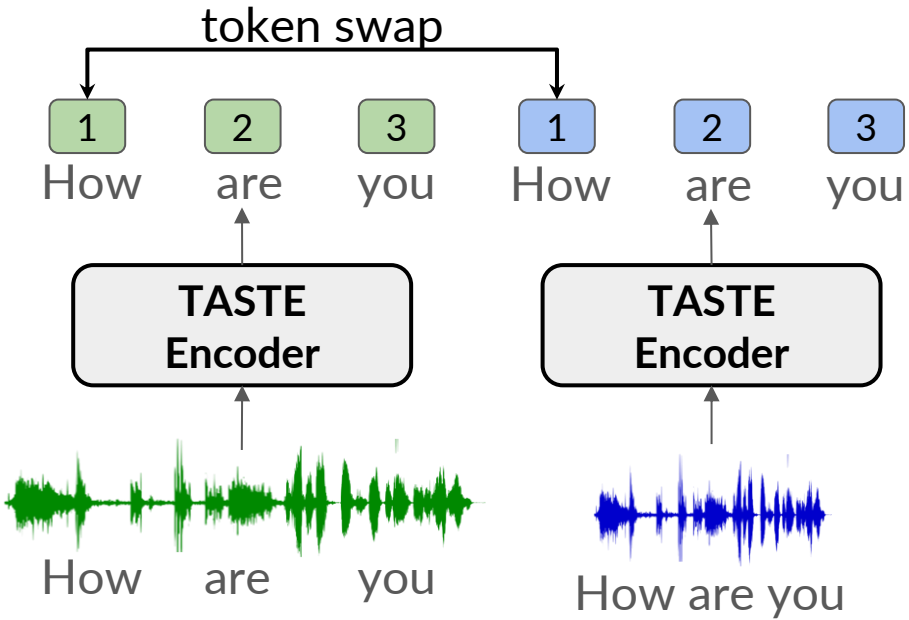
# **Result**: A Study on TASTE Tokenization

We conduct token swap to verify the attached behavior



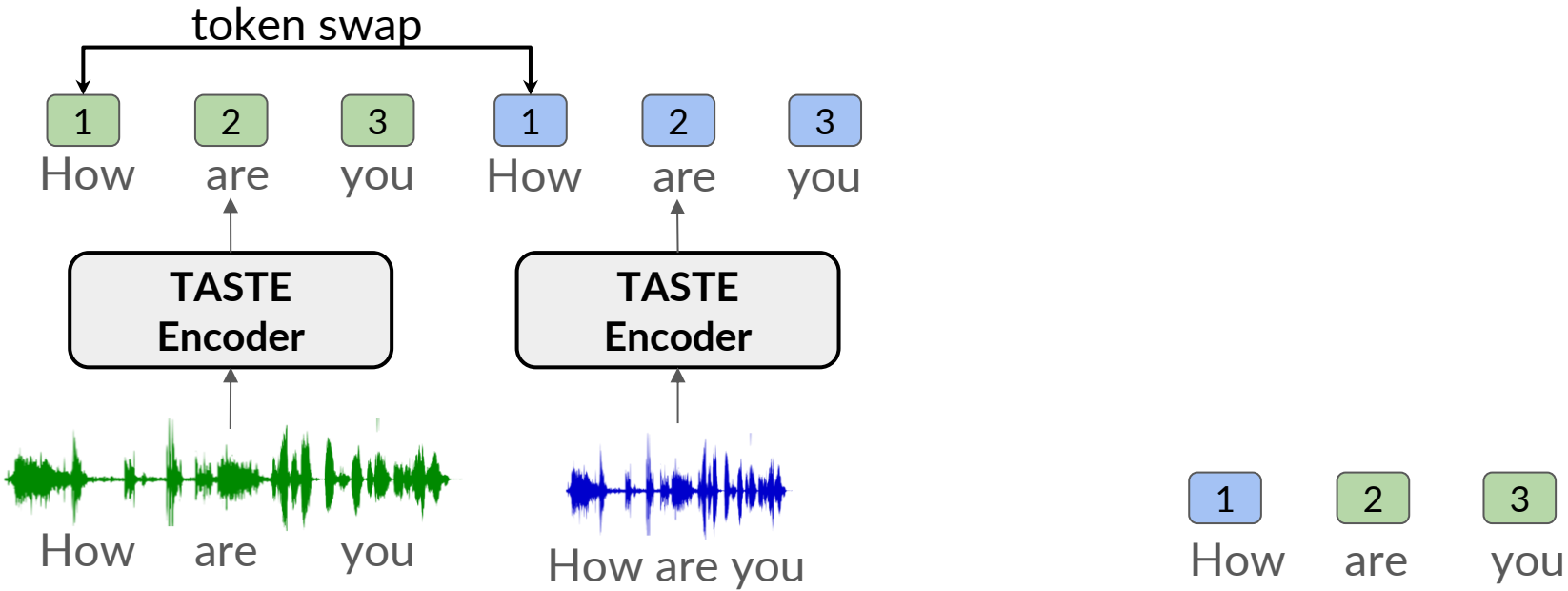
# **Result**: A Study on TASTE Tokenization

We conduct token swap to verify the attached behavior



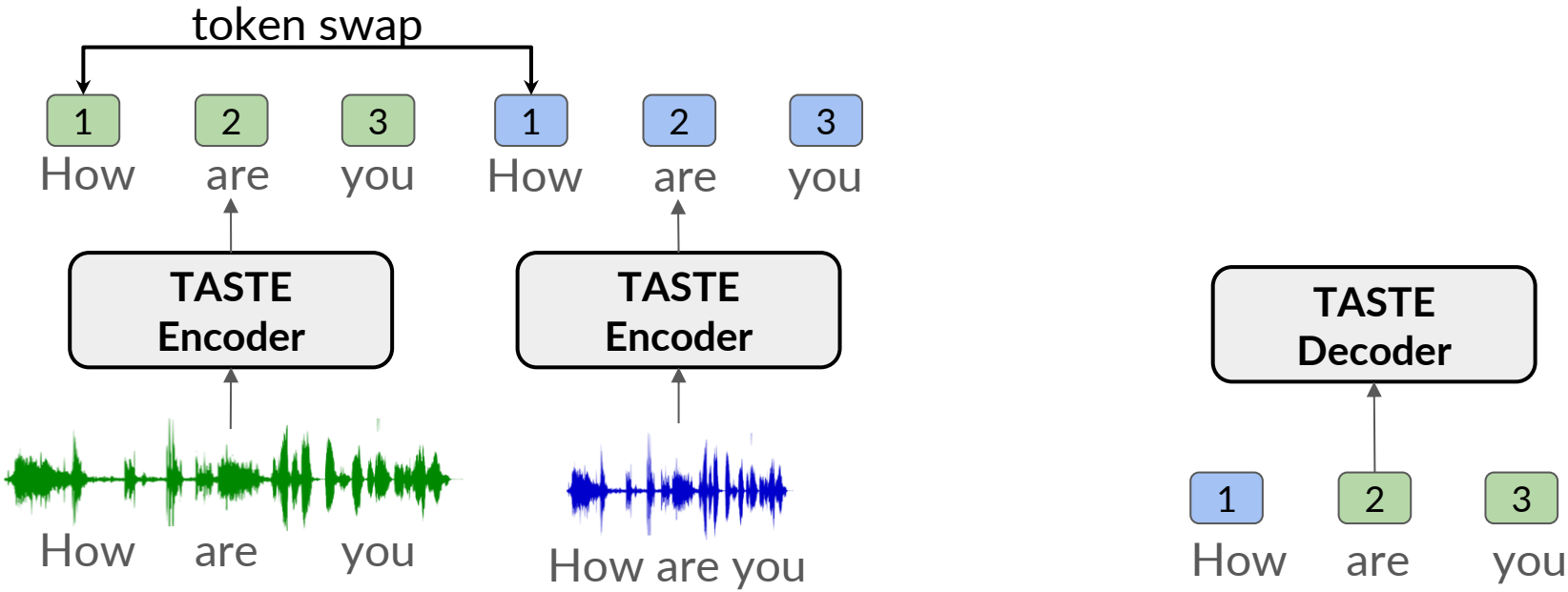
# **Result**: A Study on TASTE Tokenization

We conduct token swap to verify the attached behavior



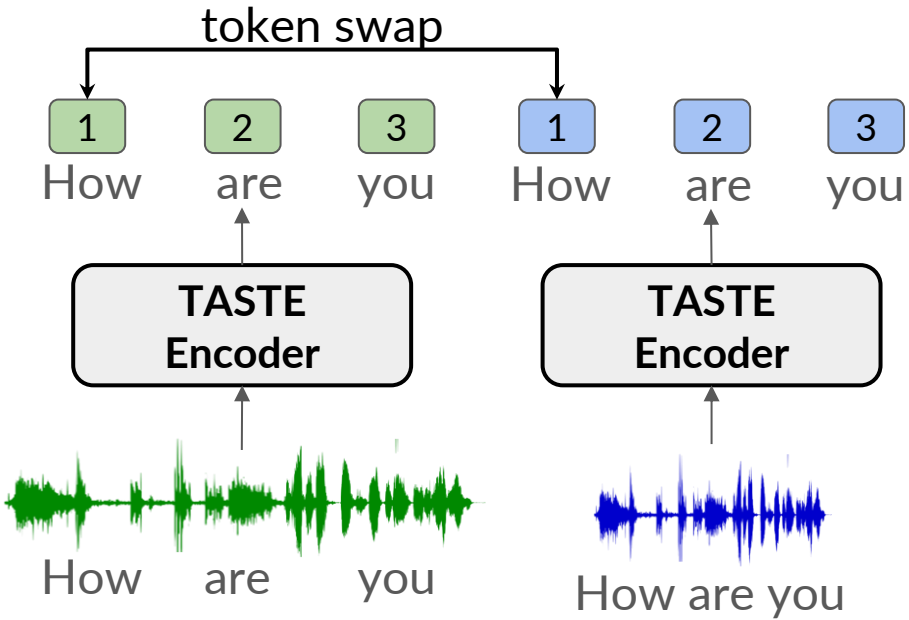
# **Result**: A Study on TASTE Tokenization

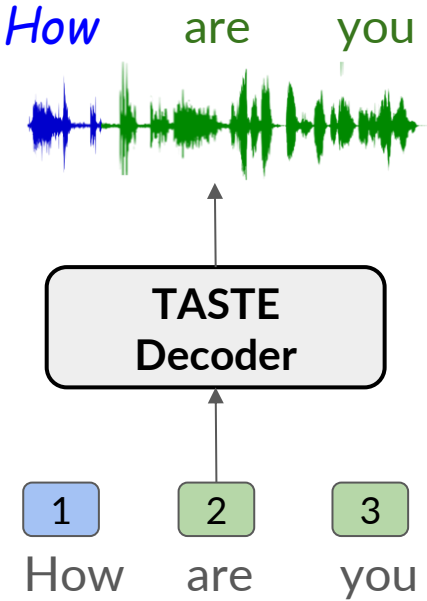
We conduct token swap to verify the attached behavior



# **Result**: A Study on TASTE Tokenization

We conduct token swap to verify the attached behavior



# **Result**: A Study on TASTE Tokenization

results of TASTE token swap

***fast***

*The captain's face had been buried in a pile of papers, but now Murdoch came around to stare at the gang leader.*

**slow**

The captain's face had been buried in a pile of papers, but now Murdoch came around to stare at the gang leader.

# **Result**: A Study on TASTE Tokenization

results of TASTE token swap

***fast***

***The captain's face*** *had been buried in a pile of papers, but now Murdoch came around to stare at the gang leader.*

**swap**

**slow**

**The captain's face** had been buried in a pile of papers, but now Murdoch came around to stare at the gang leader.

# **Result**: A Study on TASTE Tokenization

results of TASTE token swap

***The captain's face*** *had been buried in a pile of papers, but now Murdoch came around to stare at the gang leader.*

**swap**

**The captain's face** had been buried in a pile of papers, but now Murdoch came around to stare at the gang leader.

**slow** ***fast***

**The captain's face** *had been buried in a pile of papers, but now Murdoch came around to stare at the gang leader.*

# **Result**: A Study on TASTE Tokenization

results of TASTE token swap

***The captain's face*** *had been buried in a pile of papers, but now Murdoch came around to stare at the gang leader.*

**swap**

**The captain's face** had been buried in a pile of papers, but now Murdoch came around to stare at the gang leader.

**slow** ***fast***

**The captain's face** *had been buried in a pile of papers, but now Murdoch came around to stare at the gang leader.*

***fast*** **slow**

***The captain's face*** had been buried in a pile of papers, but now Murdoch came around to stare at the gang leader.

# **Result**: A Study on TASTE Tokenization

results of TASTE token swap

*The captain's face had been buried in a pile of papers, but now Murdoch **came around to** stare at the gang leader.*

**swap**

The captain's face had been buried in a pile of papers, but now Murdoch **came around to** stare at the gang leader.

***fast*** **slow** ***fast***

*The captain's face had been buried in a pile of papers, but now Murdoch **came around to** stare at the gang leader.*

**slow** ***fast*** **slow**

The captain's face had been buried in a pile of papers, but now Murdoch ***came around to*** stare at the gang leader.

# **Result**: TASTE Hybrid SLM

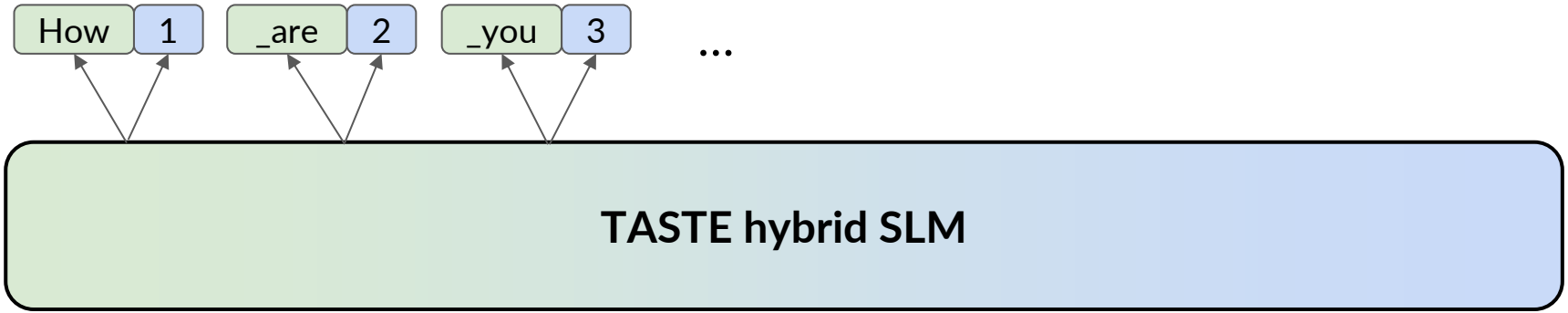
Recall that we model the text-speech tokens with next-prediction

# **Result**: TASTE Hybrid SLM

Recall that we model the text-speech tokens with next-prediction

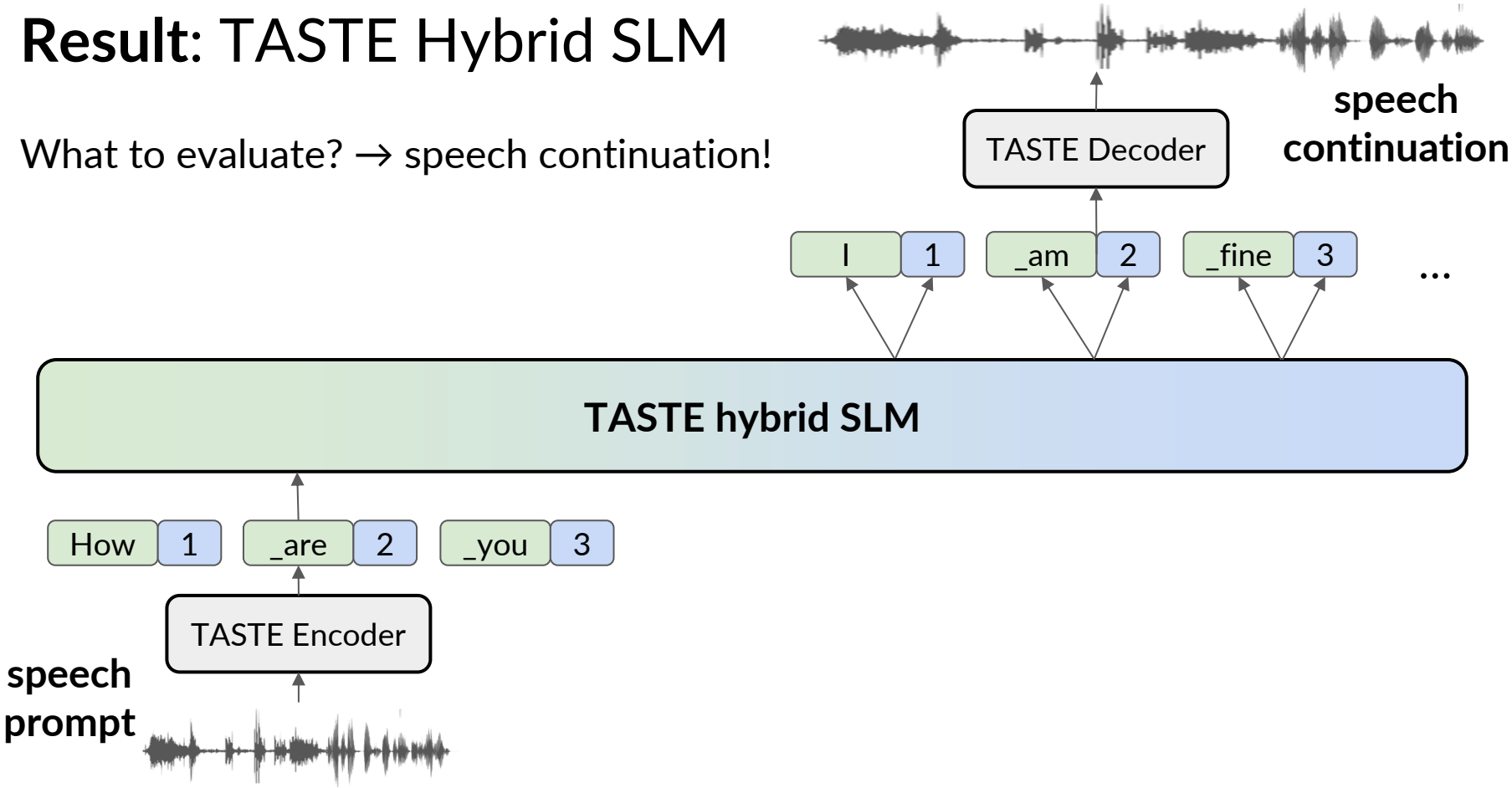
# **Result**: TASTE Hybrid SLM

Recall that we model the text-speech tokens with next-prediction



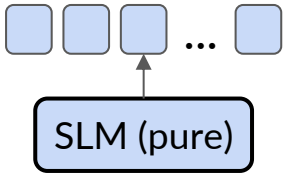
# **Result**: TASTE Hybrid SLM

What to evaluate? → speech continuation!



# **Result**: TASTE Hybrid SLM

Some examples of speech continuation




**GSLM**
(Kushal Lakhotia, et al.)

**speech prompt**

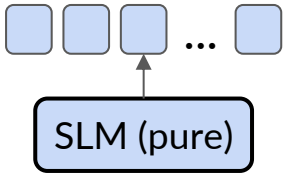
*The dark mystery of*

**speech continuation**

*the dark mystery of a polyland where it was prisoned in barkland's bank and indeed we also sought to account for it all the names of his name and poly sent a second tar and symbol which he had just received from her second figure and described*
(transcription from Whisper)

# **Result**: TASTE Hybrid SLM

Some examples of speech continuation



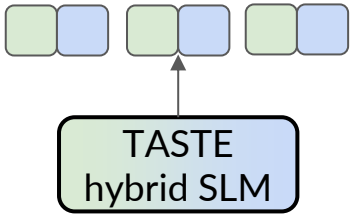

**GSLM**
(Kushal Lakhotia, et al.)

**speech prompt**

*The dark mystery of*

**speech continuation**

*the dark mystery of a polyland where it was prisoned in barkland's bank and indeed we also …*




**Ours**

*The dark mystery of*

*The dark mystery of the lost city of Atlantis has fascinated mankind for centuries. It is said that the city was located in the Atlantic Ocean, off the coast of modern-day Portugal, and was home to a civilization that flourished for thousands of years.* (self-generated)

# **Result**: TASTE Hybrid SLM

Some examples of speech continuation

| speech prompt | speech continuation |
| --- | --- |
| *I've managed to save some* | *I've managed to save some money on my phone bill, but I'm still paying a lot for my internet. I'm not sure if I should get a new phone or just get a new internet plan.* |
| *I reserve your services* | *I reserve your services for a long time. I am very happy with the result. I will definitely recommend you to my friends.* |
| *This powder must not* | *This powder must not be used in the presence of water or other liquids. It is a very fine powder and should be used with a fine powder applicator.* |
| *I'll take the armchair* | *I'll take the armchair in the corner and just sit there. It'll feel better. I'll try to not sleep.* |

# **Result**: TASTE Hybrid SLM

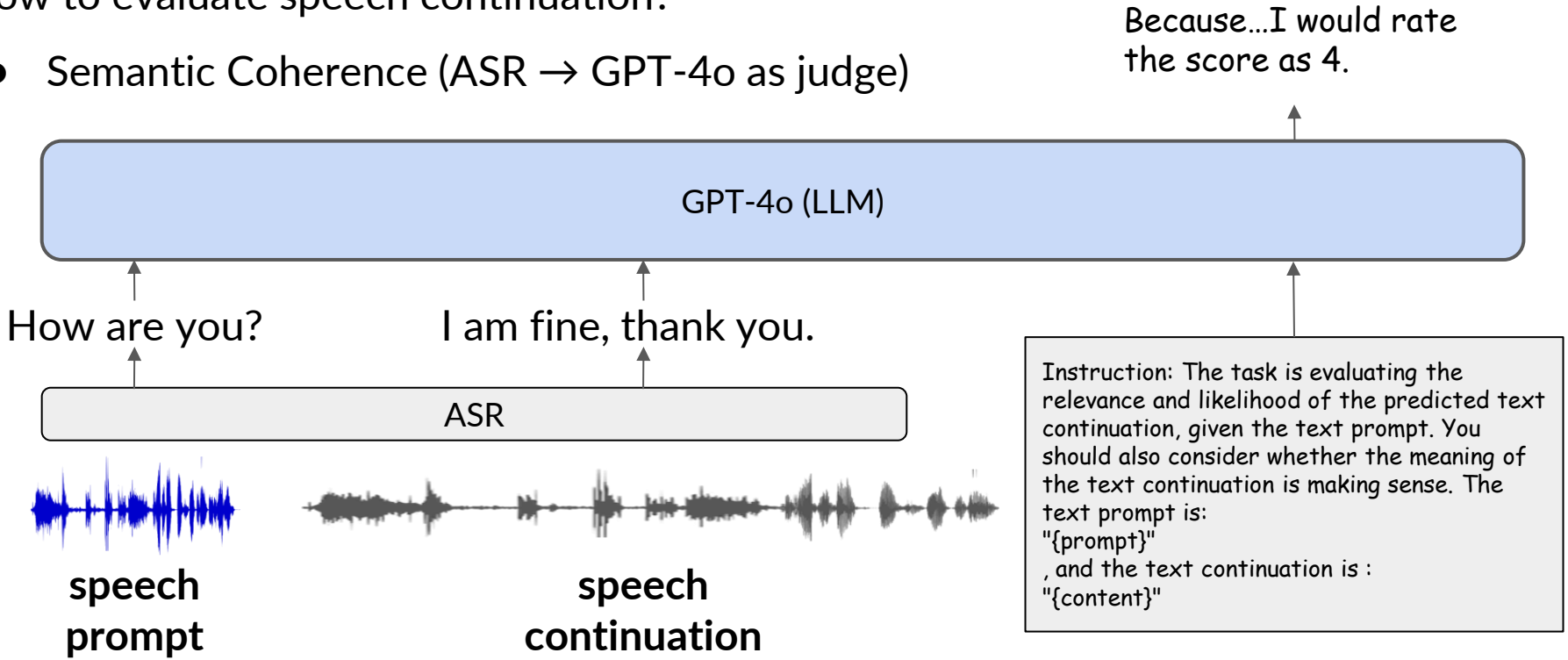
How to evaluate speech continuation?

- Semantic Coherence
- Audio Quality
- Overall Performance

# **Result**: TASTE Hybrid SLM

How to evaluate speech continuation?

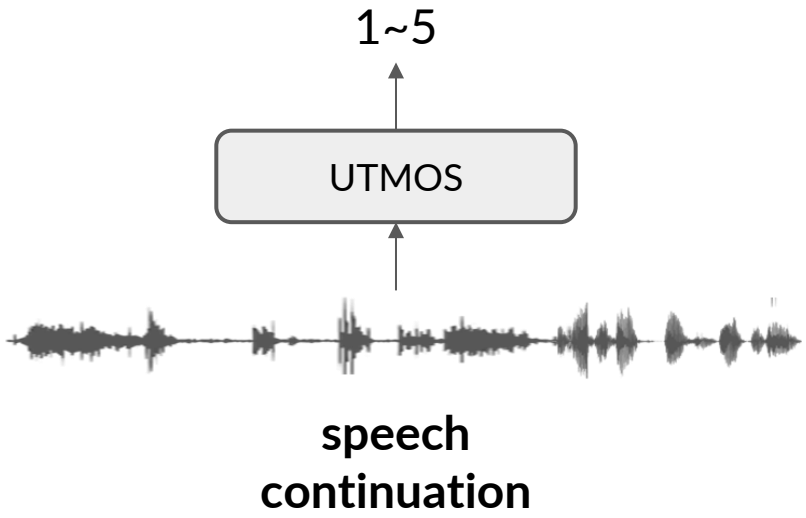
- Semantic Coherence (ASR → GPT-4o as judge)



# **Result**: TASTE Hybrid SLM

How to evaluate speech continuation?

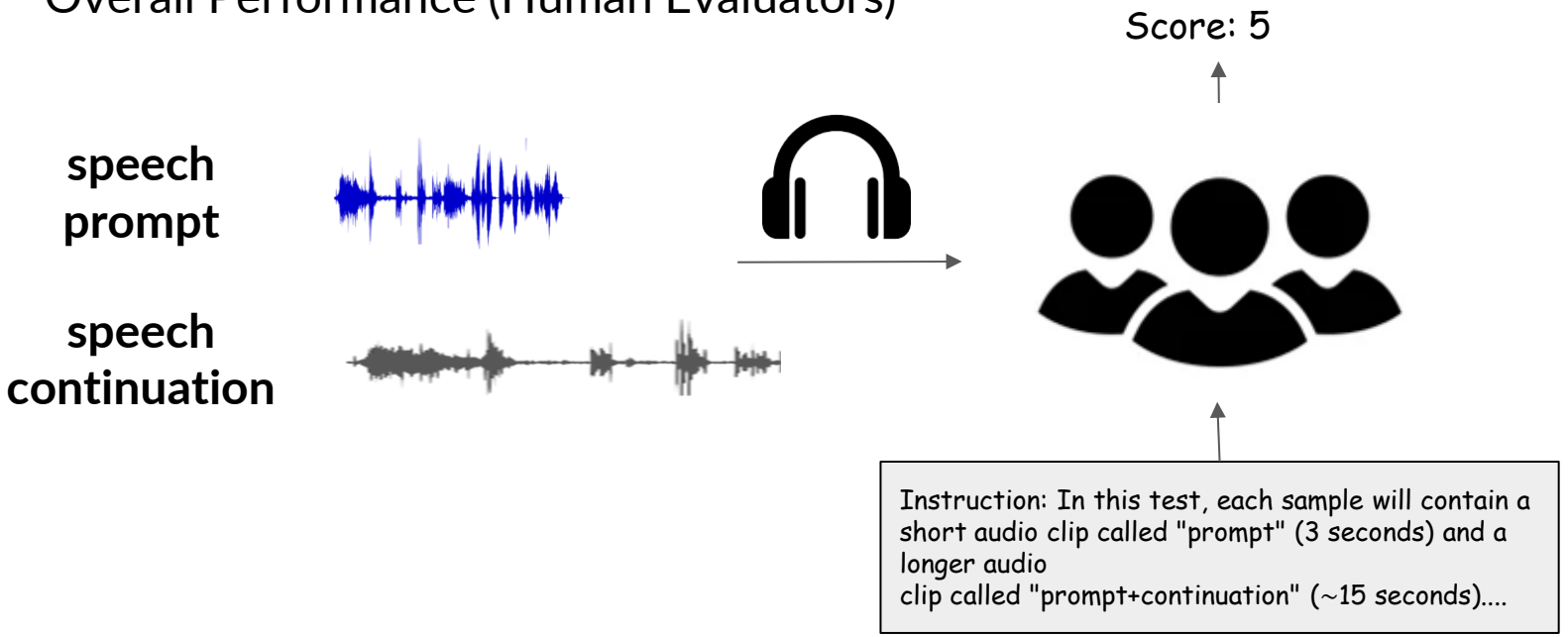
- Audio Quality (UTMOS)
  UTMOS is a widely-used audio MOS prediction model



# **Result**: TASTE Hybrid SLM

How to evaluate speech continuation?

- Overall Performance (Human Evaluators)



# **Result**: TASTE Hybrid SLM

Numerical results.

| Method | CONTINUATION | | |
|---|---|---|---|
| | GPT-4o | UTMOS | Human |
| *Cascade* | | | |
| Cascade (LLaMA3.2-1B$^{\alpha}$) | 3.15 | **4.25** | **4.00** |
| Cascade (LLaMA2-7B$^{\beta}$) | **3.43** | **4.25** | 3.98 |
| *Spoken LMs* | | | |
| TWIST 1.3B [11] | 1.48 | 3.25 | 1.95 |
| TWIST 7B [11] | 1.44 | 3.27 | 2.04 |
| Spirit LM [33] | 2.79 | 3.41 | 2.38 |
| Spirit LM Expr. [33] | 1.90 | 3.40 | 2.41 |
| Baseline (S3 token) | 1.37 | 4.04 | 2.84 |
| TASLM 1B (token) | 3.08 | 4.07 | 3.93 |
| TASLM 1B (embed.) | **3.16** | **4.22** | **4.16** |

# **Result**: TASTE Hybrid SLM

Semantic Coherence:
our method outperform other SLMs, and is the only one that is on par with the cascade system.

| Method | CONTINUATION | | |
|---|---|---|---|
| | GPT-4o | UTMOS | Human |
| *Cascade* | | | |
| Cascade (LLaMA3.2-1B$^{\alpha}$) | 3.15 | **4.25** | **4.00** |
| Cascade (LLaMA2-7B$^{\beta}$) | **3.43** | **4.25** | 3.98 |
| *Spoken LMs* | | | |
| TWIST 1.3B [11] | 1.48 | 3.25 | 1.95 |
| TWIST 7B [11] | 1.44 | 3.27 | 2.04 |
| Spirit LM [33] | 2.79 | 3.41 | 2.38 |
| Spirit LM Expr. [33] | 1.90 | 3.40 | 2.41 |
| Baseline (S3 token) | 1.37 | 4.04 | 2.84 |
| TASLM 1B (token) | 3.08 | 4.07 | 3.93 |
| TASLM 1B (embed.) | **3.16** | **4.22** | **4.16** |

# **Result**: TASTE Hybrid SLM

Audio Quality:
very well

| Method | CONTINUATION | | |
|---|---|---|---|
| | GPT-4o | UTMOS | Human |
| *Cascade* | | | |
| Cascade (LLaMA3.2-1B$^{\alpha}$) | 3.15 | **4.25** | **4.00** |
| Cascade (LLaMA2-7B$^{\beta}$) | **3.43** | **4.25** | 3.98 |
| *Spoken LMs* | | | |
| TWIST 1.3B [11] | 1.48 | 3.25 | 1.95 |
| TWIST 7B [11] | 1.44 | 3.27 | 2.04 |
| Spirit LM [33] | 2.79 | 3.41 | 2.38 |
| Spirit LM Expr. [33] | 1.90 | 3.40 | 2.41 |
| Baseline (S3 token) | 1.37 | 4.04 | 2.84 |
| TASLM 1B (token) | 3.08 | 4.07 | 3.93 |
| TASLM 1B (embed.) | **3.16** | **4.22** | **4.16** |

# **Result**: TASTE Hybrid SLM

Overall Performance:
From human perspective, our method is the best, even surpasses the cascade system

| Method | CONTINUATION | | |
|---|---|---|---|
| | GPT-4o | UTMOS | Human |
| *Cascade* | | | |
| Cascade (LLaMA3.2-1B$^{\alpha}$) | 3.15 | **4.25** | **4.00** |
| Cascade (LLaMA2-7B$^{\beta}$) | **3.43** | **4.25** | 3.98 |
| *Spoken LMs* | | | |
| TWIST 1.3B [11] | 1.48 | 3.25 | 1.95 |
| TWIST 7B [11] | 1.44 | 3.27 | 2.04 |
| Spirit LM [33] | 2.79 | 3.41 | 2.38 |
| Spirit LM Expr. [33] | 1.90 | 3.40 | 2.41 |
| Baseline (S3 token) | 1.37 | 4.04 | 2.84 |
| TASLM 1B (token) | 3.08 | 4.07 | 3.93 |
| TASLM 1B (embed.) | **3.16** | **4.22** | **4.16** |

# **Main Result**: Brief Summary

As speech tokenization

- TASTE successfully conveys paralinguistic information.

For spoken language modeling

- TASTE is **extremely effective for hybrid SLM**, yielding high quality speech continuation with strong semantic coherence.

# **Main Result**: Brief Summary

As speech tokenization

- TASTE successfully conveys paralinguistic information.

For spoken language modeling

- TASTE is **extremely effective for hybrid SLM**, yielding high quality speech continuation with strong semantic coherence.

We have successfully built a joint tokenization that facilitates joint SLM!

# Q & A

# Thank you!