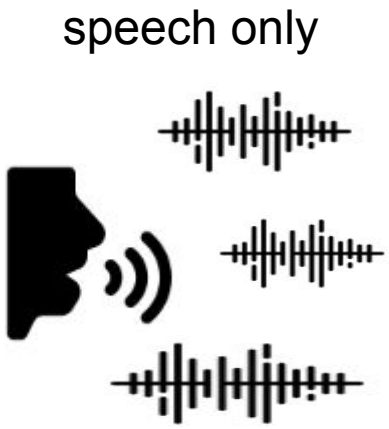
# REBORN: Reinforcement-Learned Boundary Segmentation with Iterative Training for Unsupervised ASR

Liang-Hsuan Tseng, En-Pei Hu, Cheng-Han Chiang, Yuan Tseng,
Hung-yi Lee, Lin-shan Lee, Shao-Hua Sun


# What is Unsupervised ASR (UASR)?

Learning a speech recognition system without any labeled data.



# What is Unsupervised ASR (UASR)?

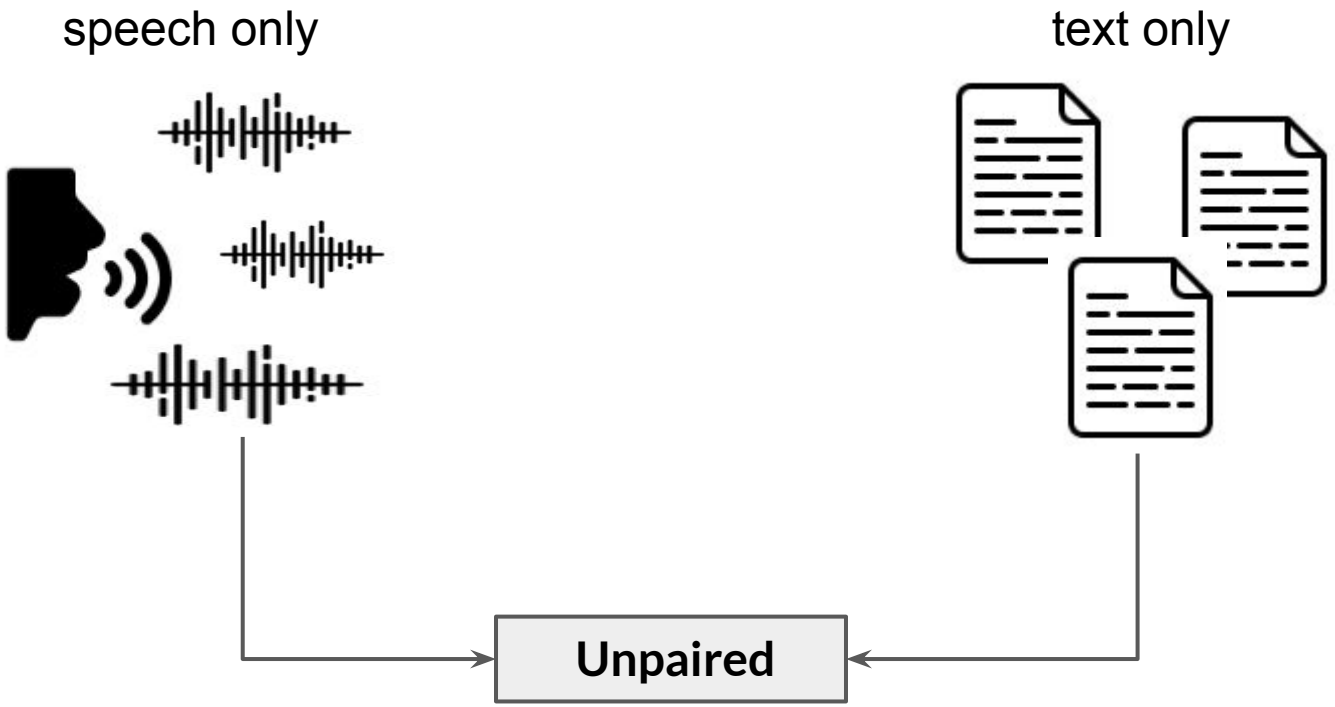
Learning a speech recognition system without any labeled data.

speech only

text only

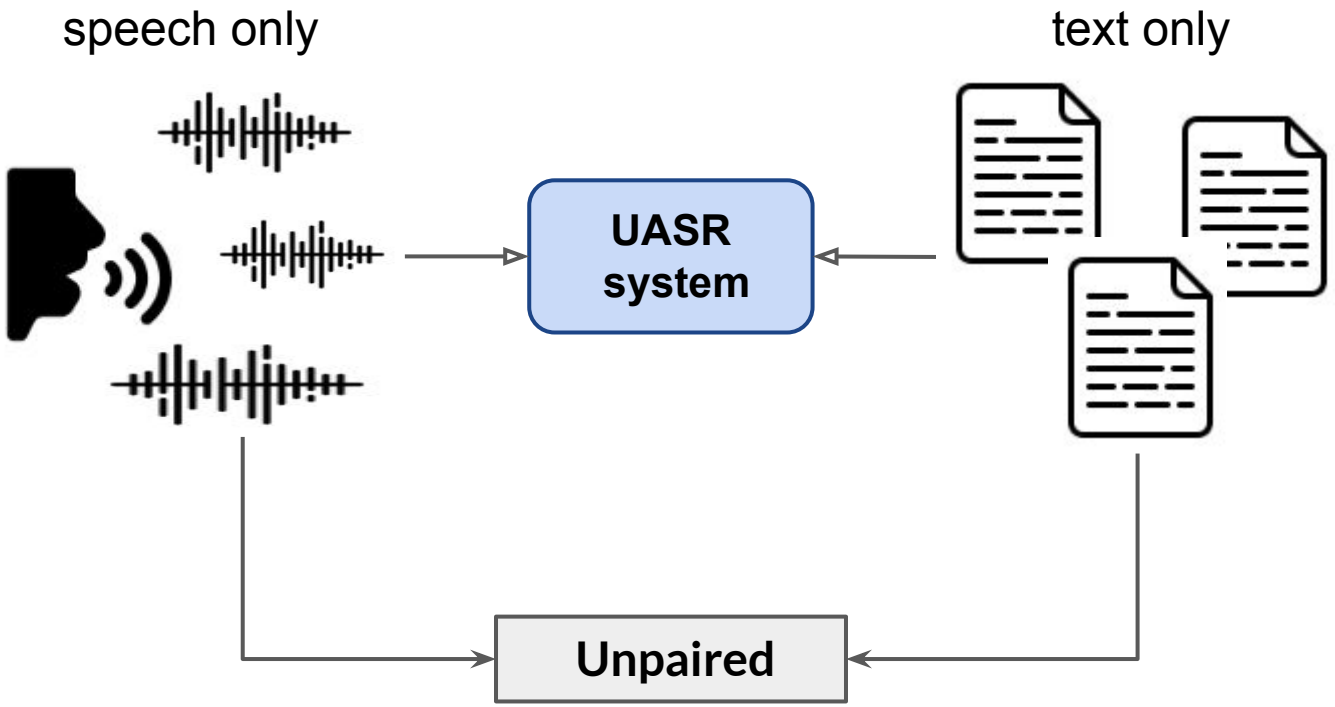
# What is Unsupervised ASR (UASR)?

Learning a speech recognition system without any labeled data.



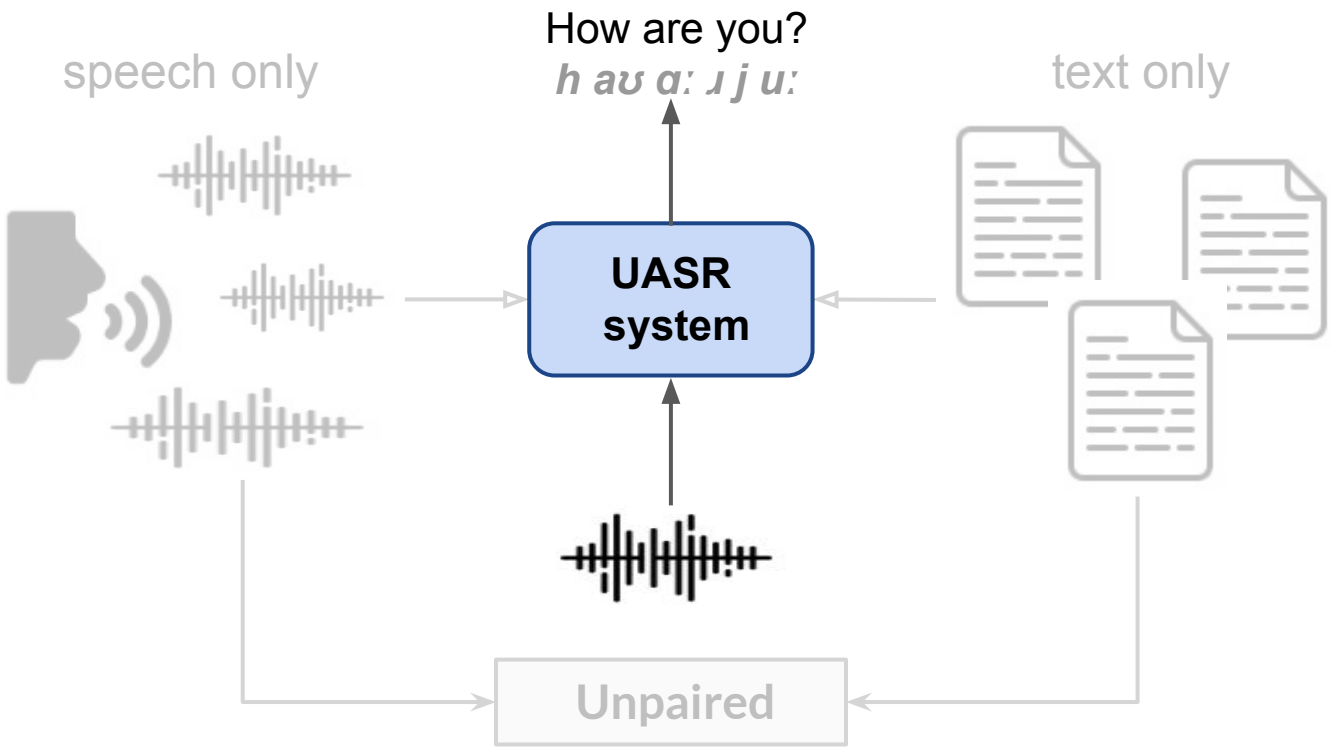
# What is Unsupervised ASR (UASR)?

Learning a speech recognition system without any labeled data.



# What is Unsupervised ASR (UASR)?

Learning a speech recognition system without any labeled data.



# UASR is hard because…

Speech and text are different!

# UASR is hard because…

Speech and text are different!

*continuous*

d i s c r e t e

# UASR is hard because…

Speech and text are different!

*continuous*

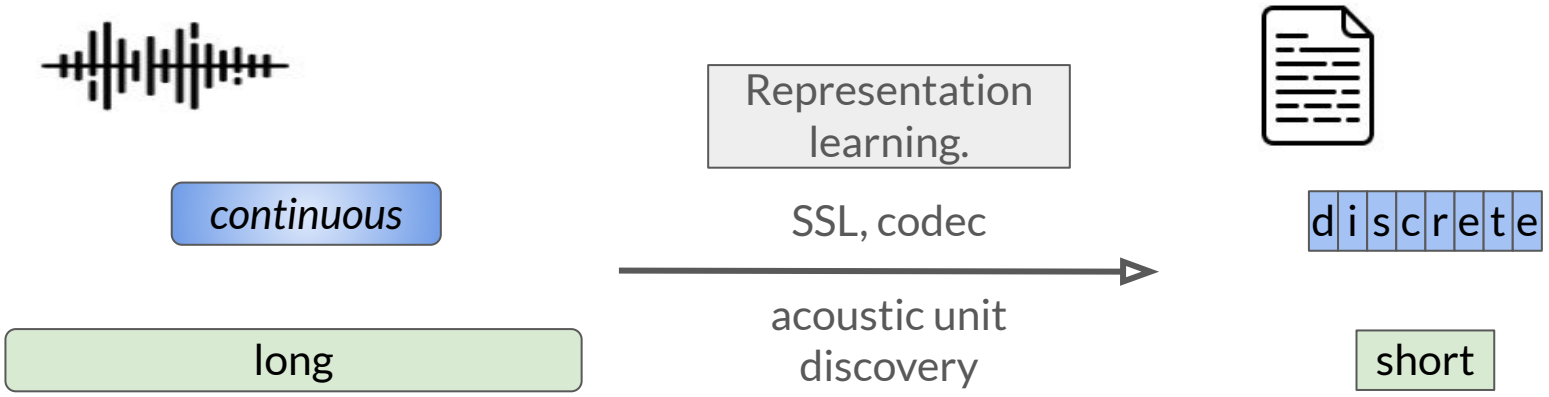
long

d i s c r e t e

short

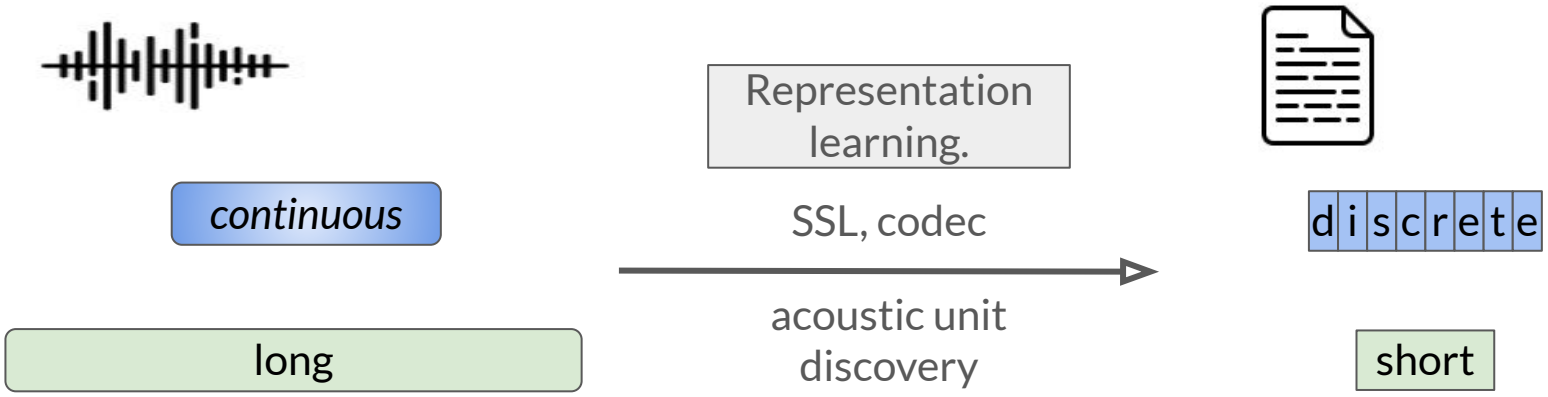
# UASR is hard because…

Speech and text are different!



# UASR is hard because…

Speech and text are different!




UASR aims to ***match the distribution*** from speech to text with **unpaired** data only!

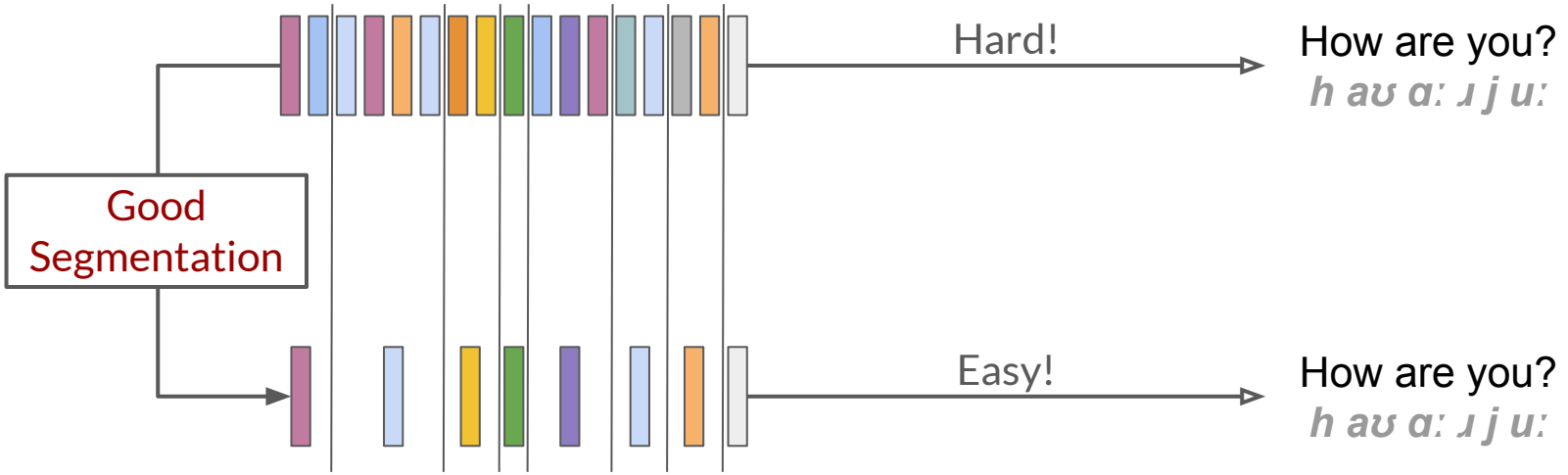
# Segmentation is what we need

Directly mapping from speech features to text (phoneme) is hard.

# Segmentation is what we need

Directly mapping from speech features to text (phoneme) is hard.

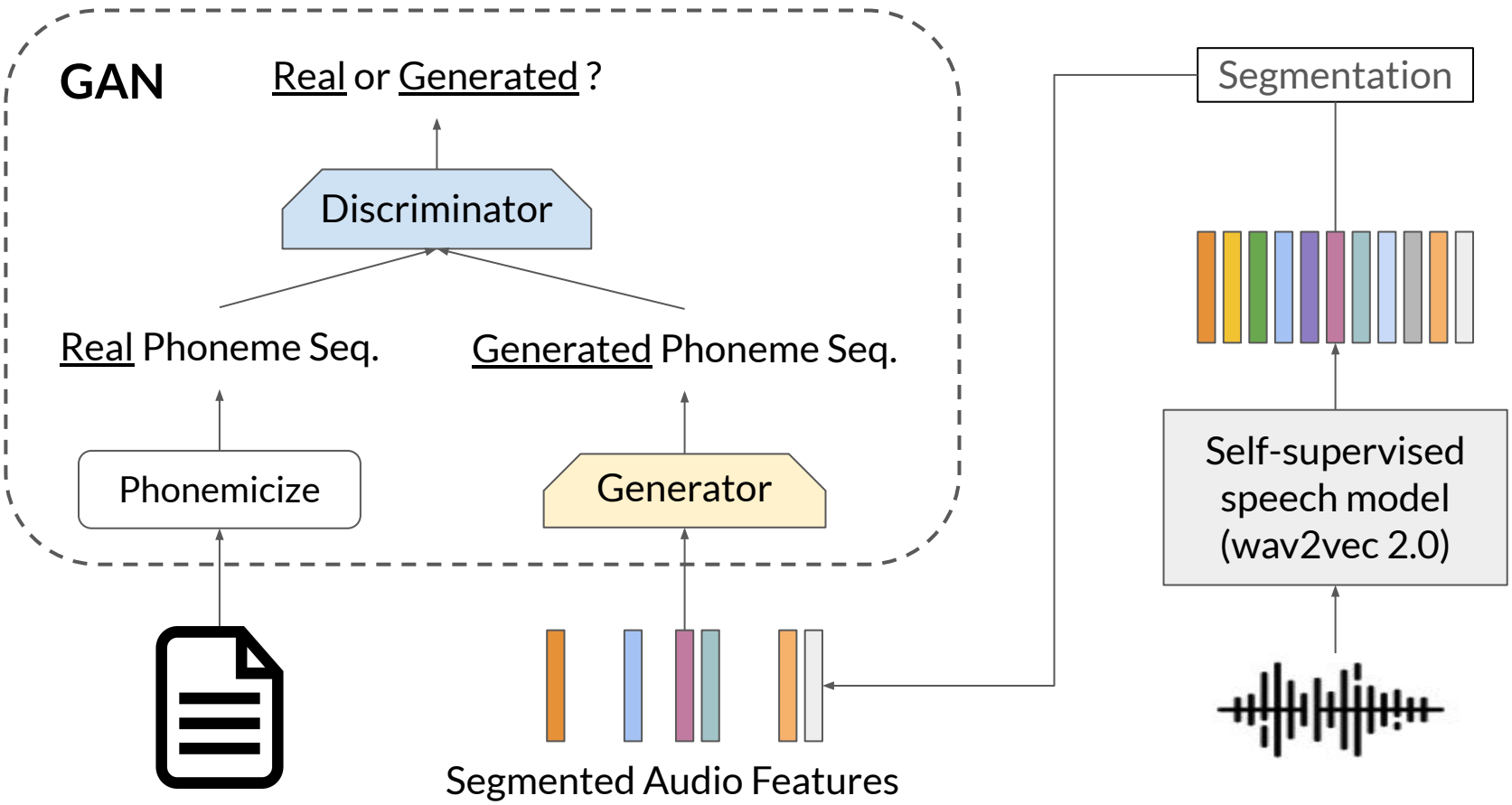
⇒ **Segmentation!**




The distribution matching problem would be easier given a ***good segmentation***.

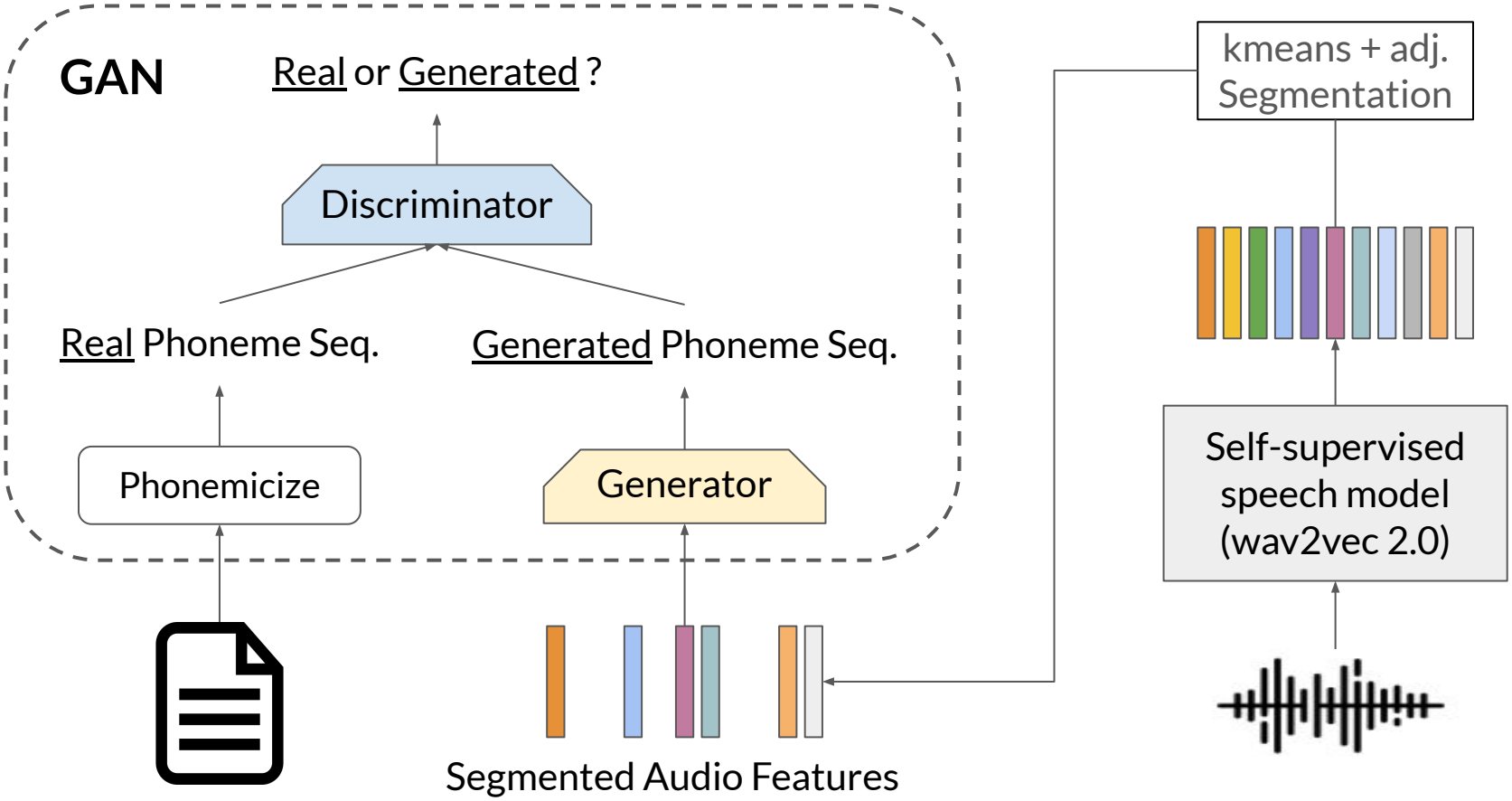
# *k*-means-based segmentation: wav2vec-U

Feature extrator: wav2vec 2.0

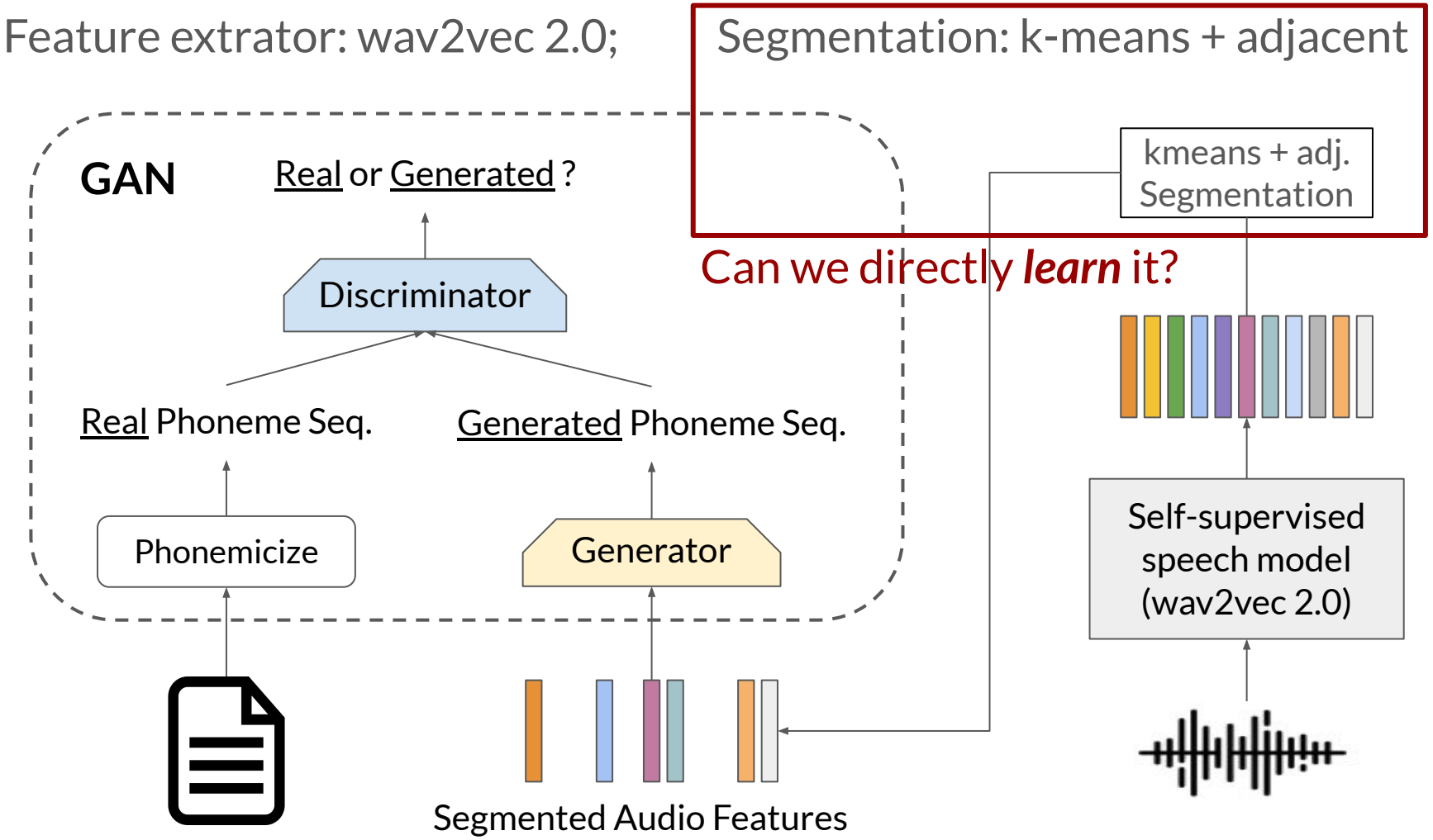


# *k*-means-based segmentation: wav2vec-U

Feature extrator: wav2vec 2.0; Segmentation: k-means + adjacent

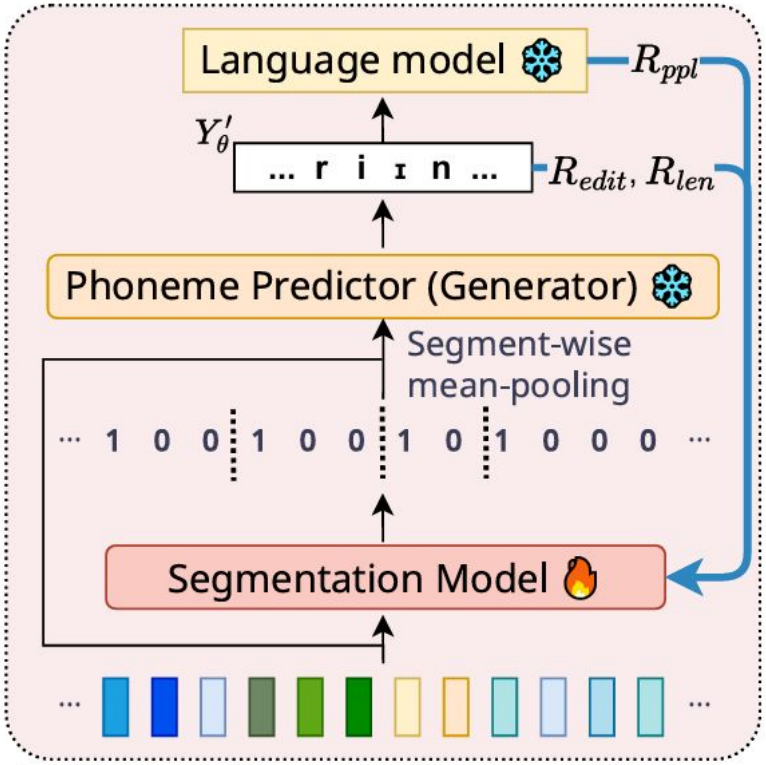


# *k*-means-based segmentation: wav2vec-U

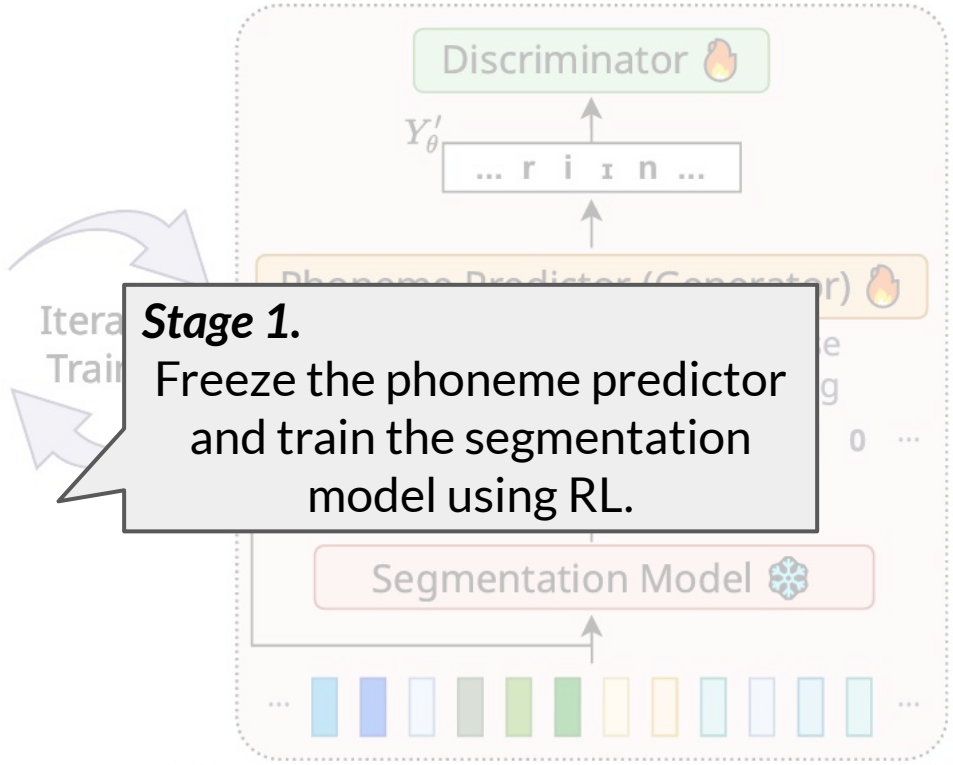


# REBORN

**Re**inforcement-Learned **Bo**undary Segmentation with Iterative Training for Unsupervised ASR



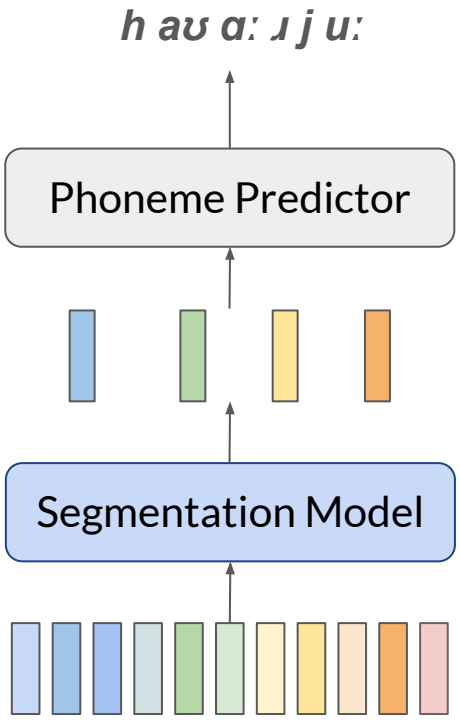

(a) Stage 1: Segmentation model training




(b) Stage 2: Phoneme prediction model training

# Using RL to learn the segmentation model

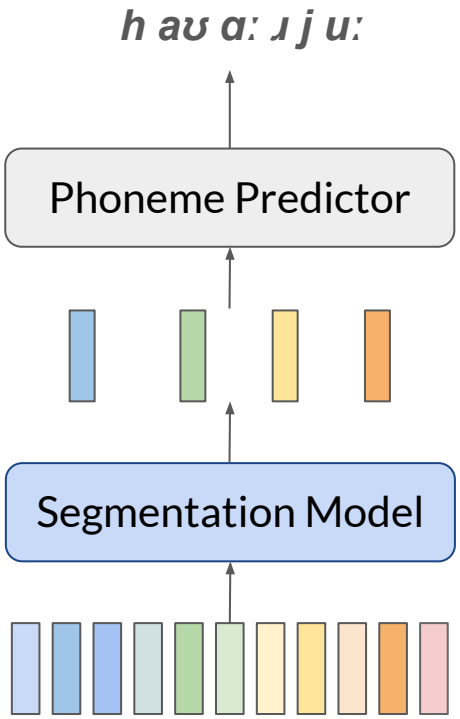
What we want is a segmentation that can yield a ***better transcription***.



# Using RL to learn the segmentation model

What we want is a segmentation that can yield a ***better transcription***.

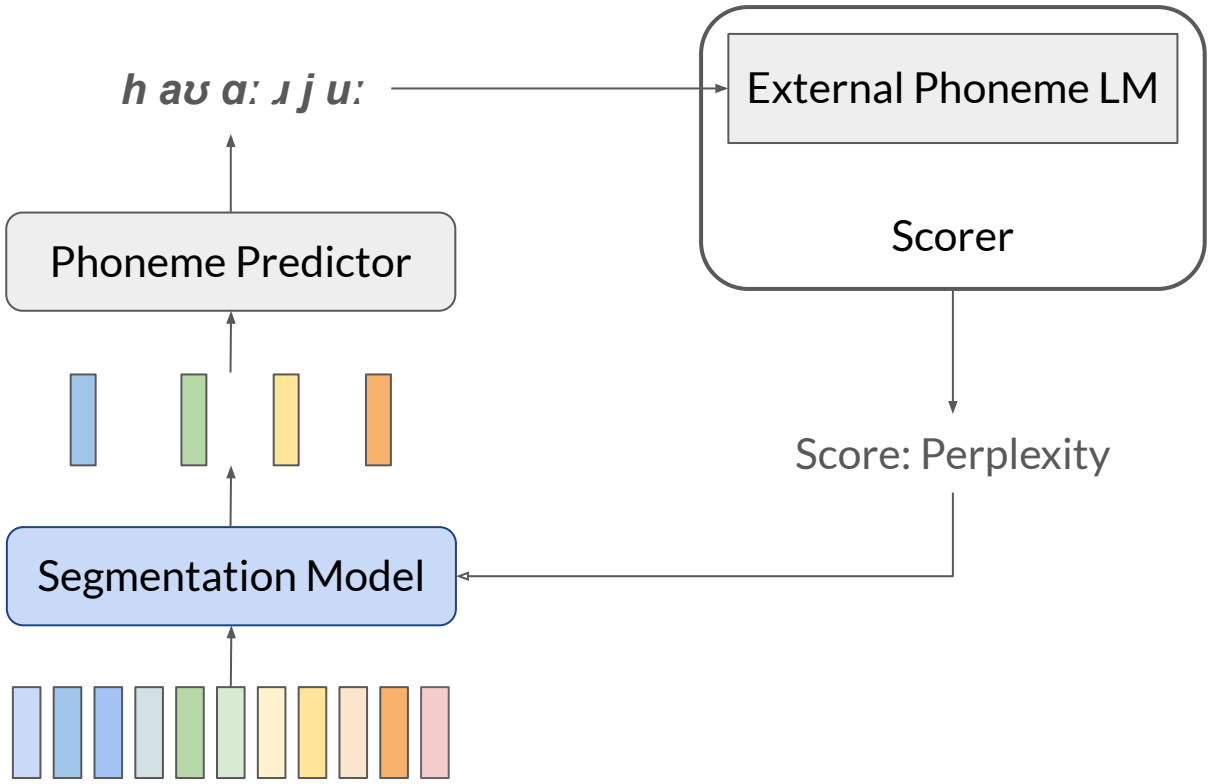
How to determine that a transcription is better?



# Using RL to learn the segmentation model

What we want is a segmentation that can yield a ***better transcription***.

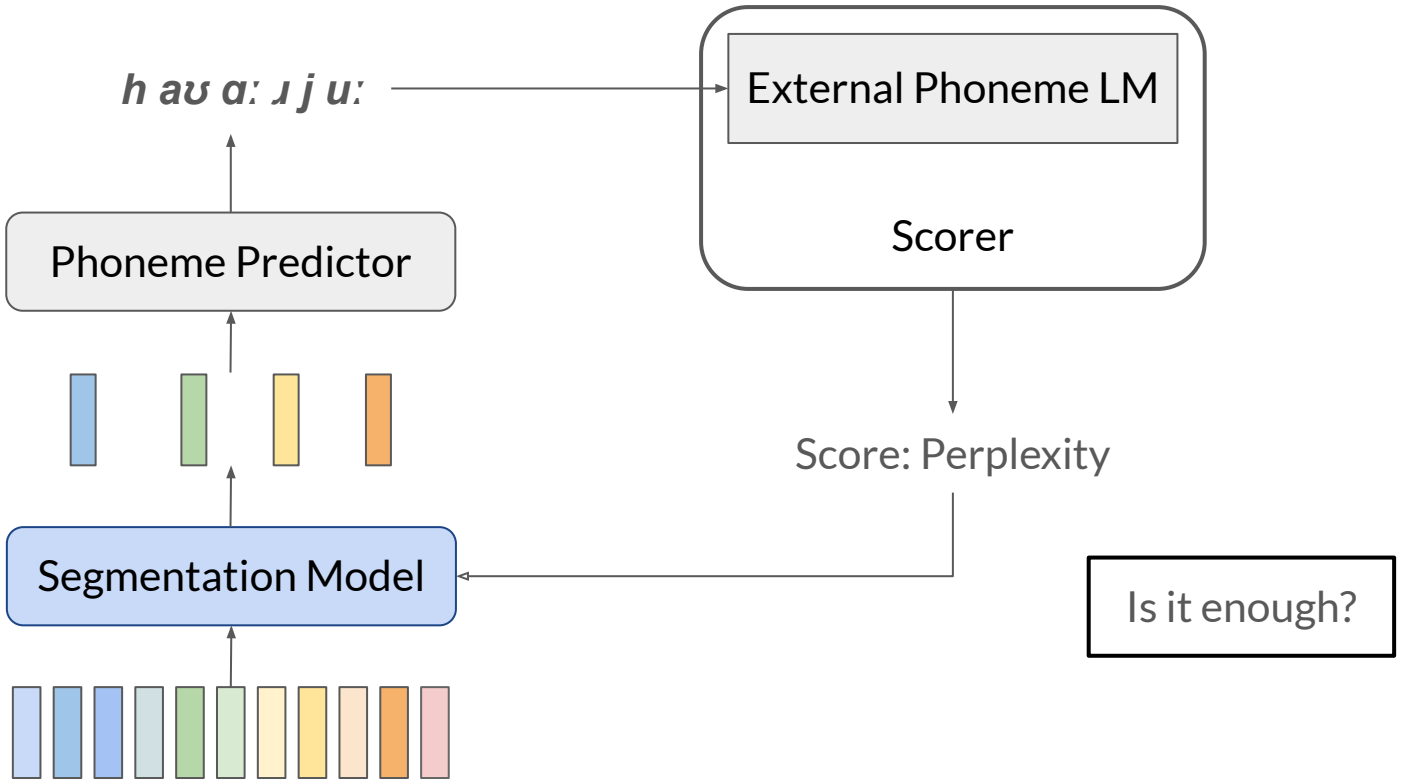
How to determine that a transcription is better? → Use the external LM.



# Using RL to learn the segmentation model

What we want is a segmentation that can yield a ***better transcription***.

How to determine that a transcription is better? → Use the external LM.



# Utilizing relativeness for reward design

Our reward is calculated by ***comparing with the transcription from the previous iteration***.

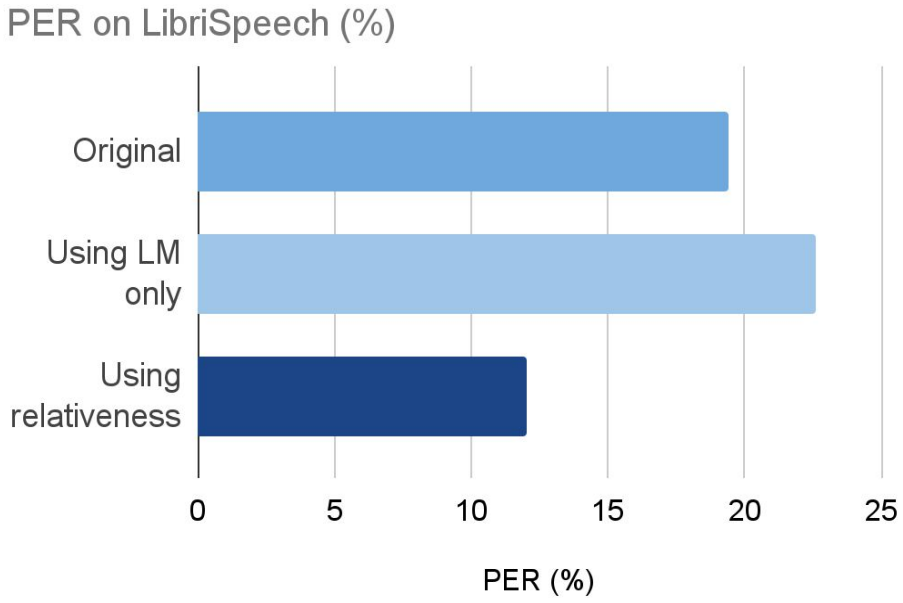
# Utilizing relativeness for reward design

Our reward is calculated by ***comparing with the transcription from the previous iteration***.

We introduce the three different rewards:

- PPL Difference Reward
- Edit Distance Reward
- Length Difference Reward

Using perplexity **difference** between iterations yields more stable and better results



# Utilizing relativeness for reward design

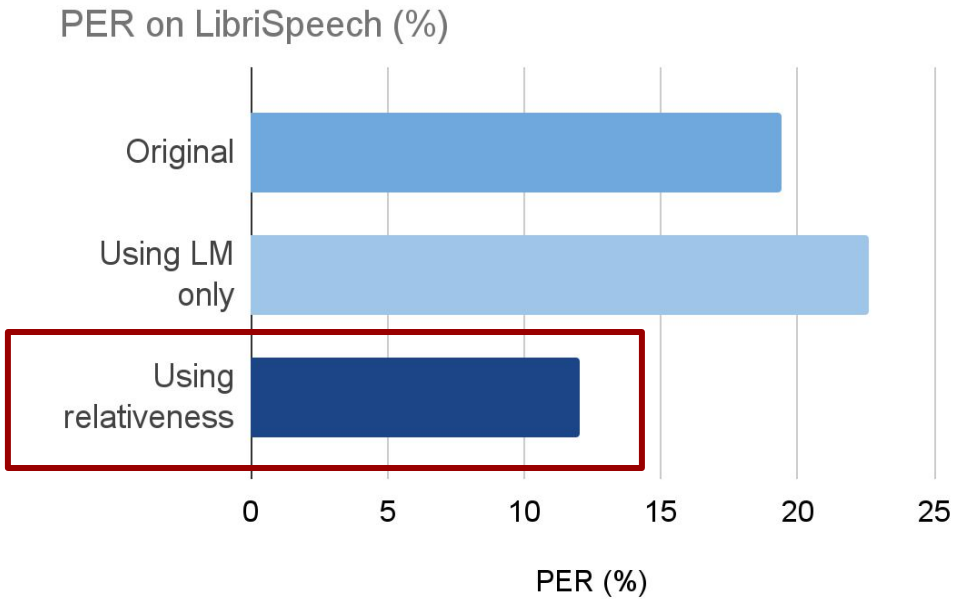
Our reward is calculated by ***comparing with the transcription from the previous iteration***.

We introduce the three different rewards:

- PPL Difference Reward
- Edit Distance Reward
- Length Difference Reward

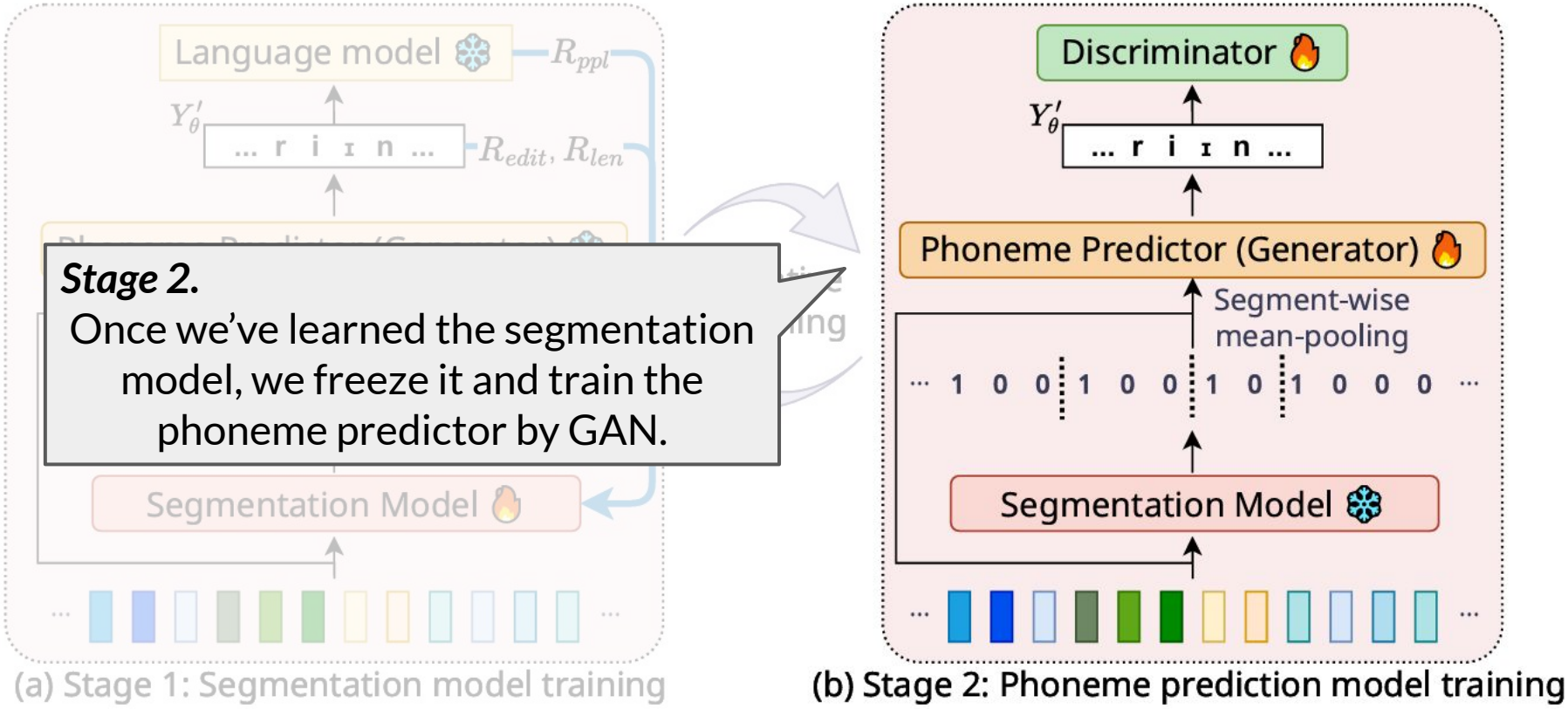
Using perplexity **difference** between iterations yields more stable and better results

We discover that the concept of using relativeness is essential for our RL rewards.



# REBORN

**Re**inforcement-Learned **Bo**undary Segmentation with Ite**r**ative Trai**n**ing for Unsupervised ASR

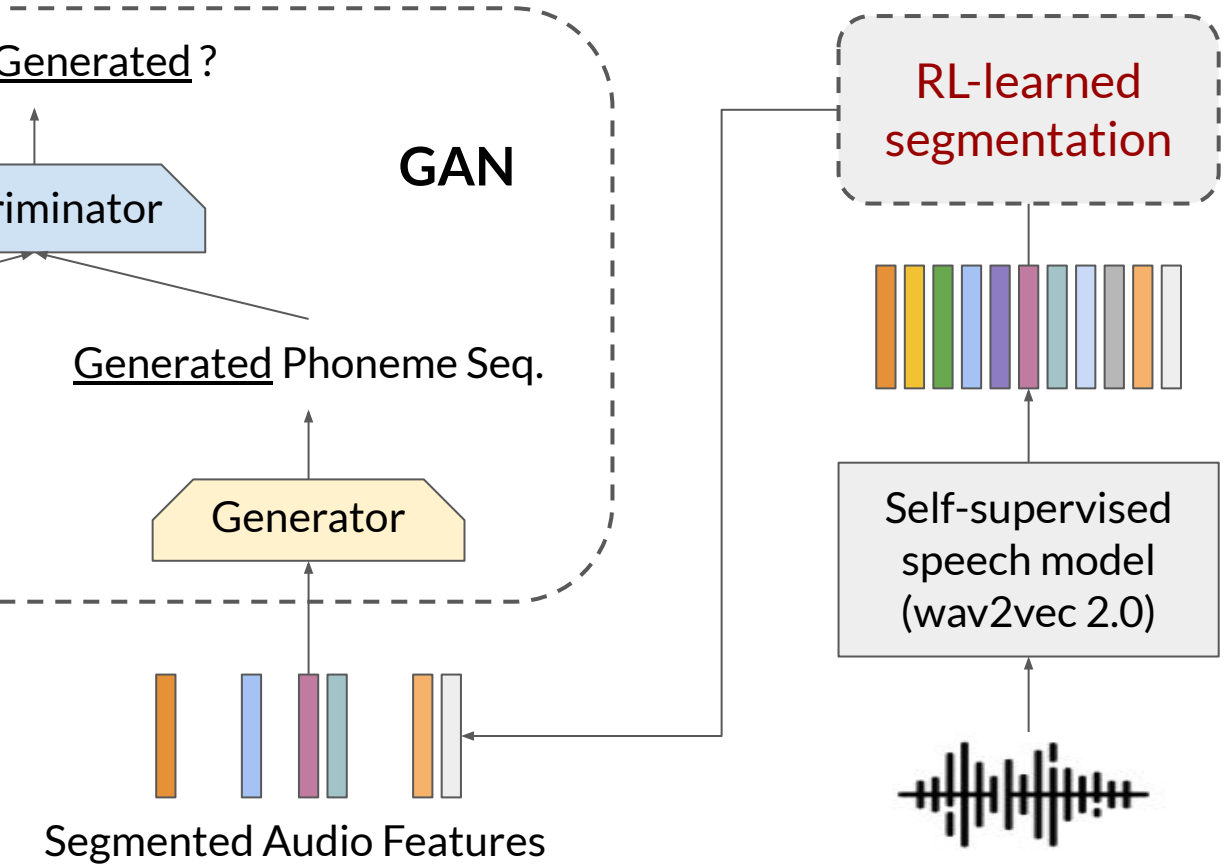



(a) Stage 1: Segmentation model training

(b) Stage 2: Phoneme prediction model training

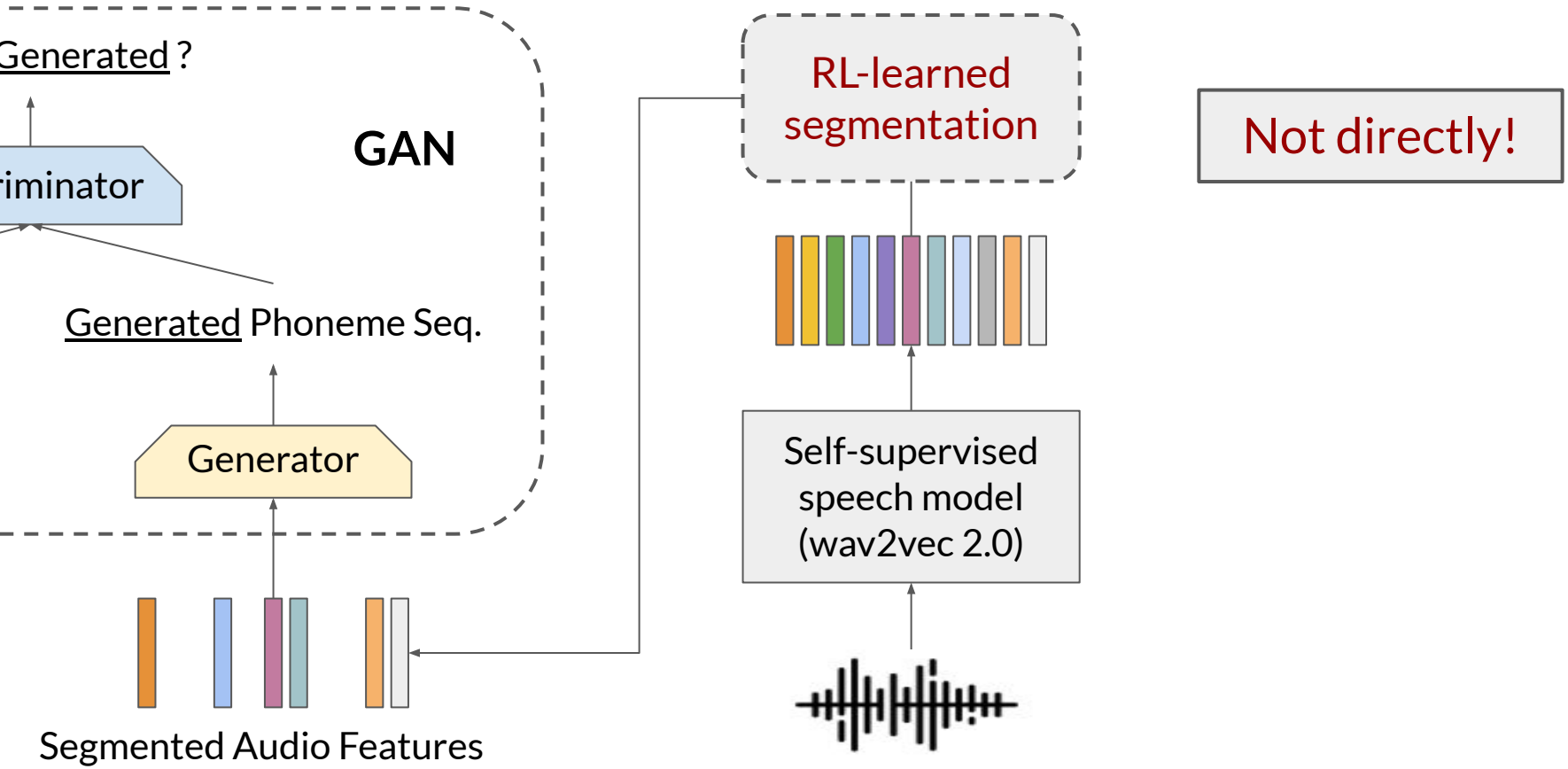
# Improving the phoneme predictor further

Can we directly use the new RL-learned segmentation for the stage 2 GAN training?



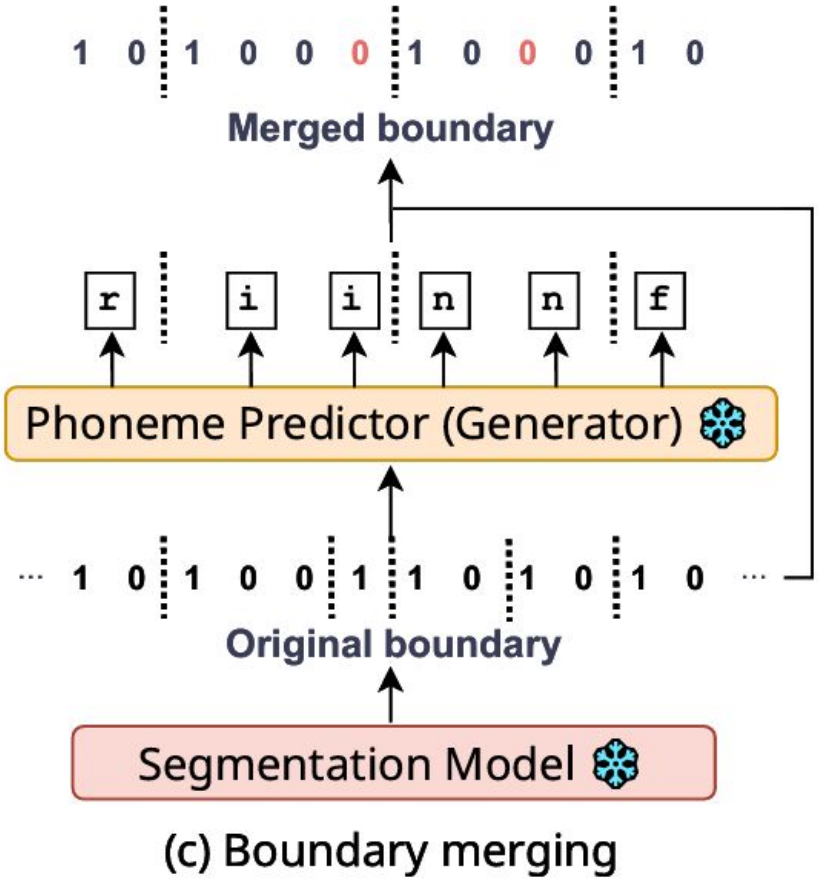
# Improving the phoneme predictor further

Can we directly use the new RL-learned segmentation for the stage 2 GAN training?



# Introducing boundary merging

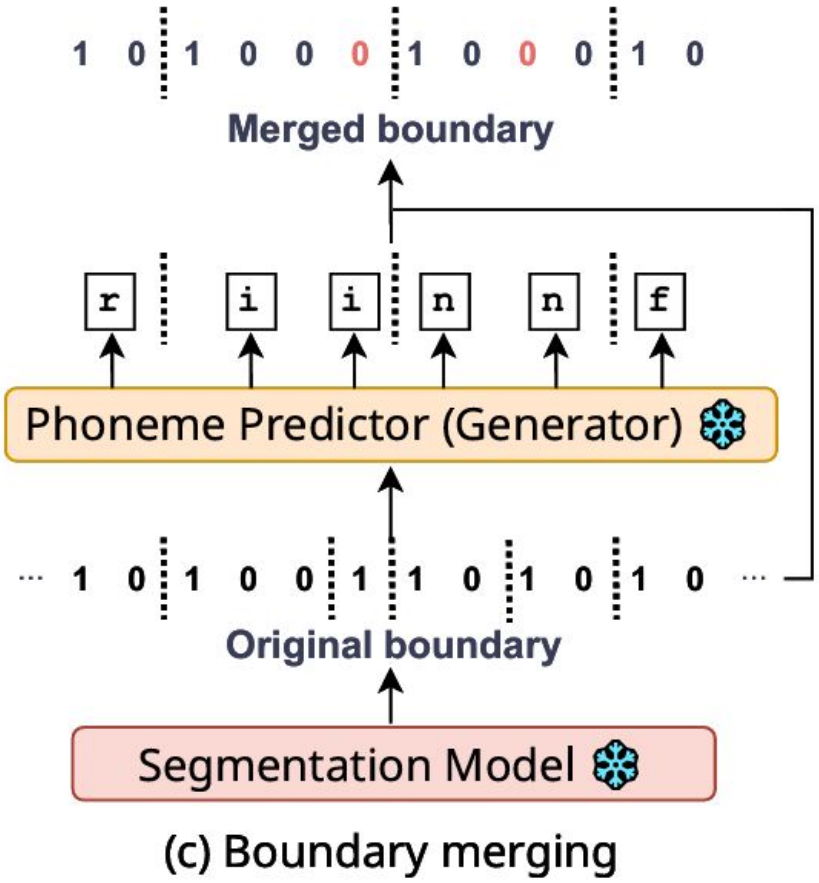
We merge the segments yielding consecutive phoneme predictions.




(c) Boundary merging

# Introducing boundary merging

We merge the segments yielding consecutive phoneme predictions.

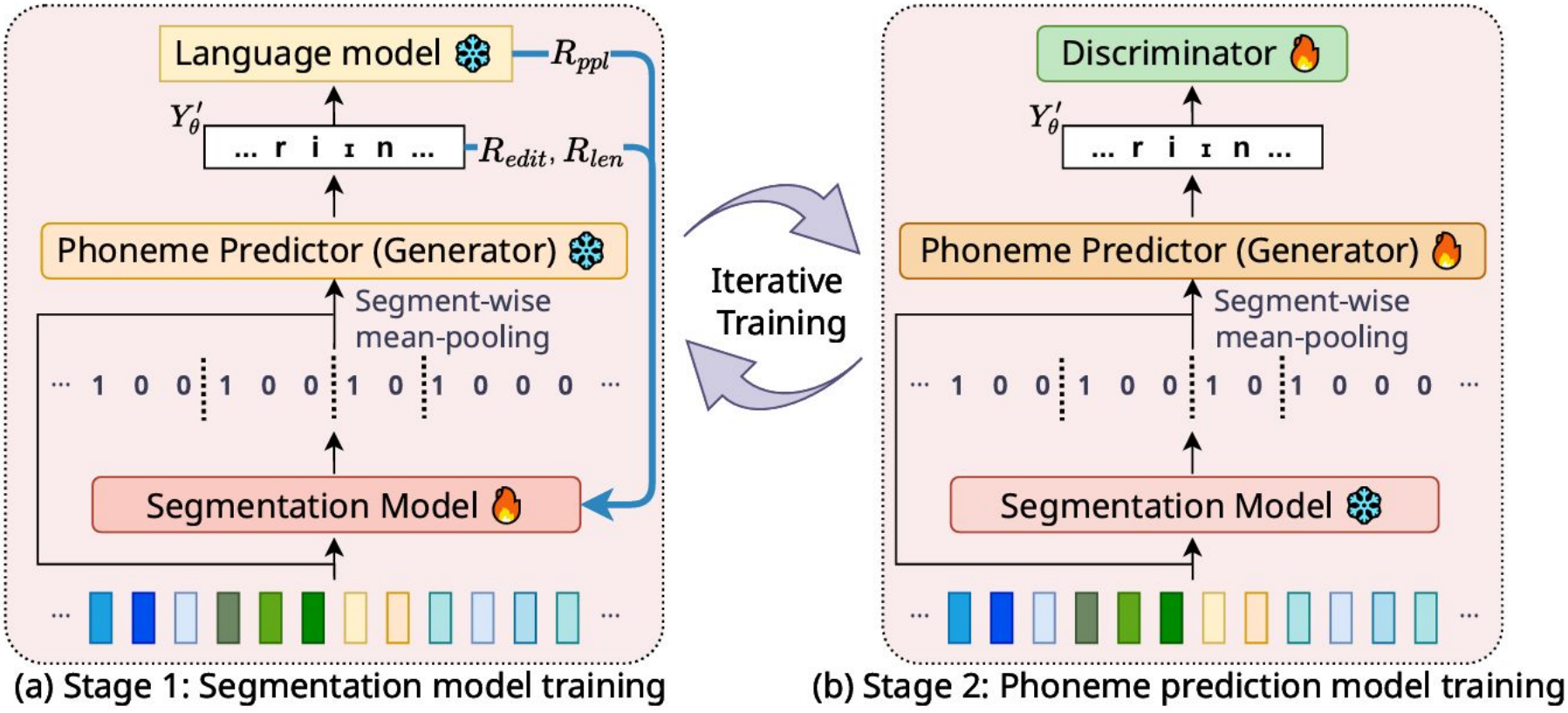



(c) Boundary merging

The amount of segments decreases, the GAN training is more stable and can yield better performance.

# REBORN

**Re**inforcement-Learned **Bo**undary Segmentation with Ite**r**ative Trai**n**ing for Unsupervised ASR

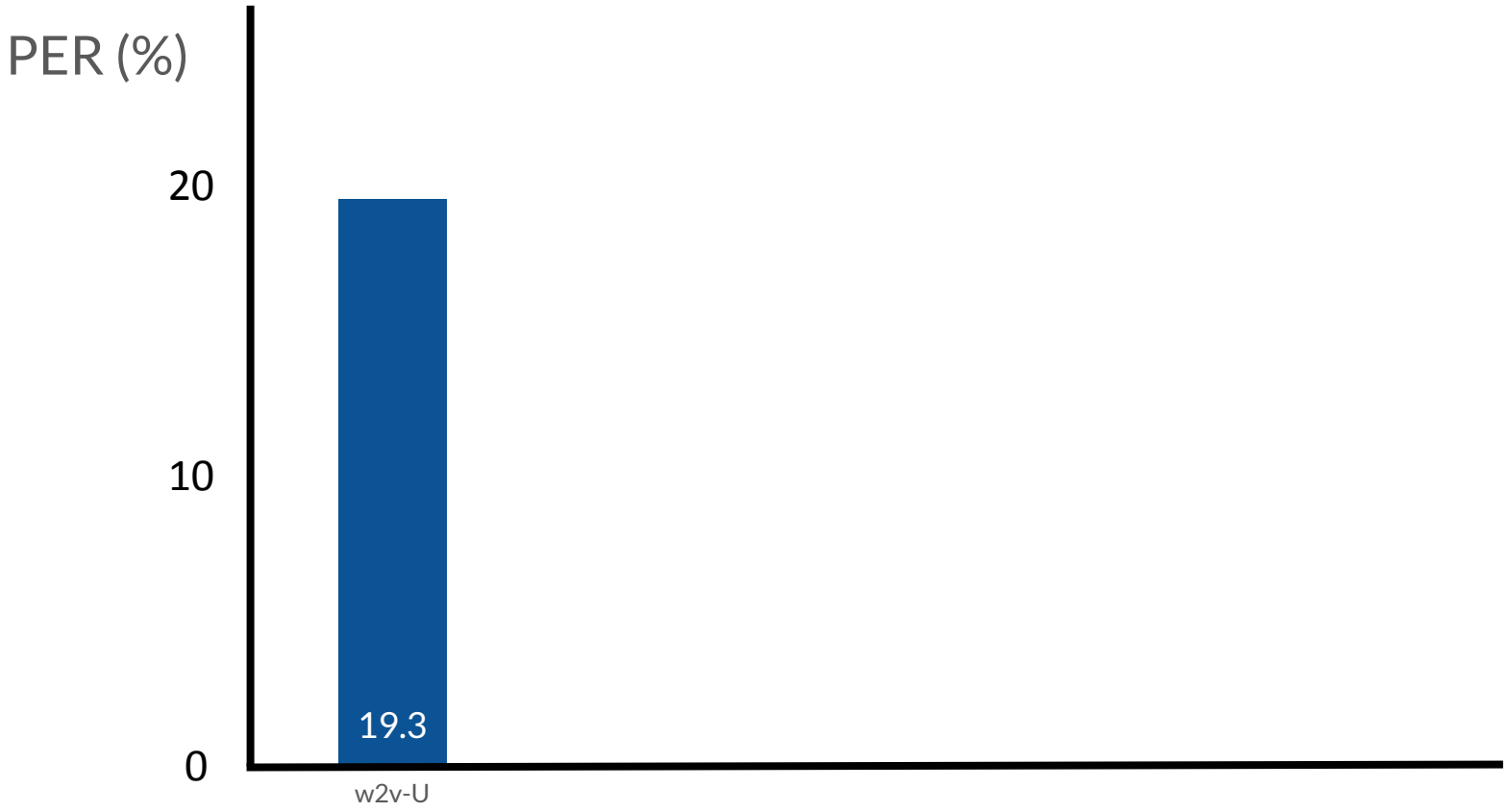



(a) Stage 1: Segmentation model training

(b) Stage 2: Phoneme prediction model training

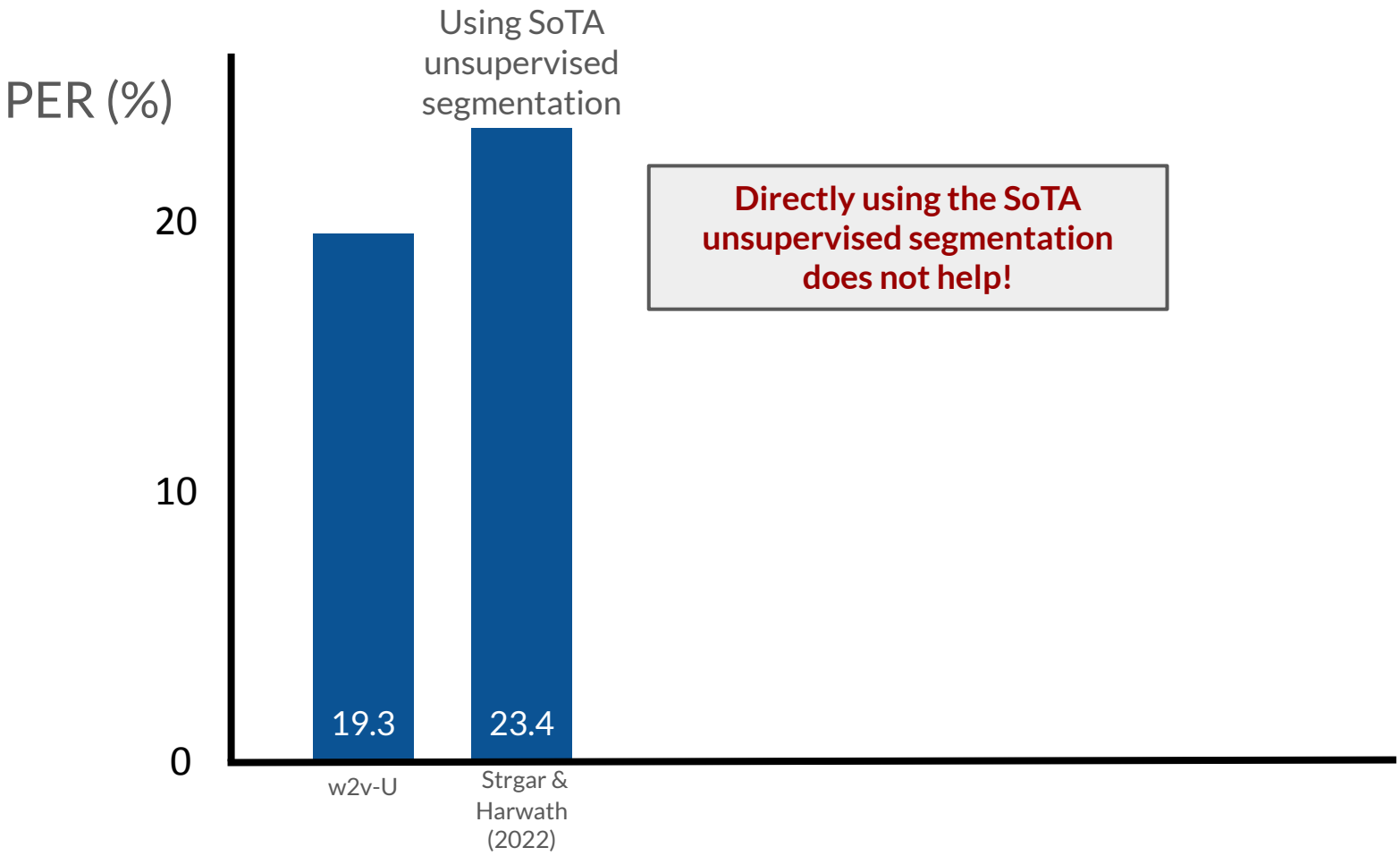
# Result on LibriSpeech

Evaluation metric: PER (%), the lower the better. Training data: 100 hours



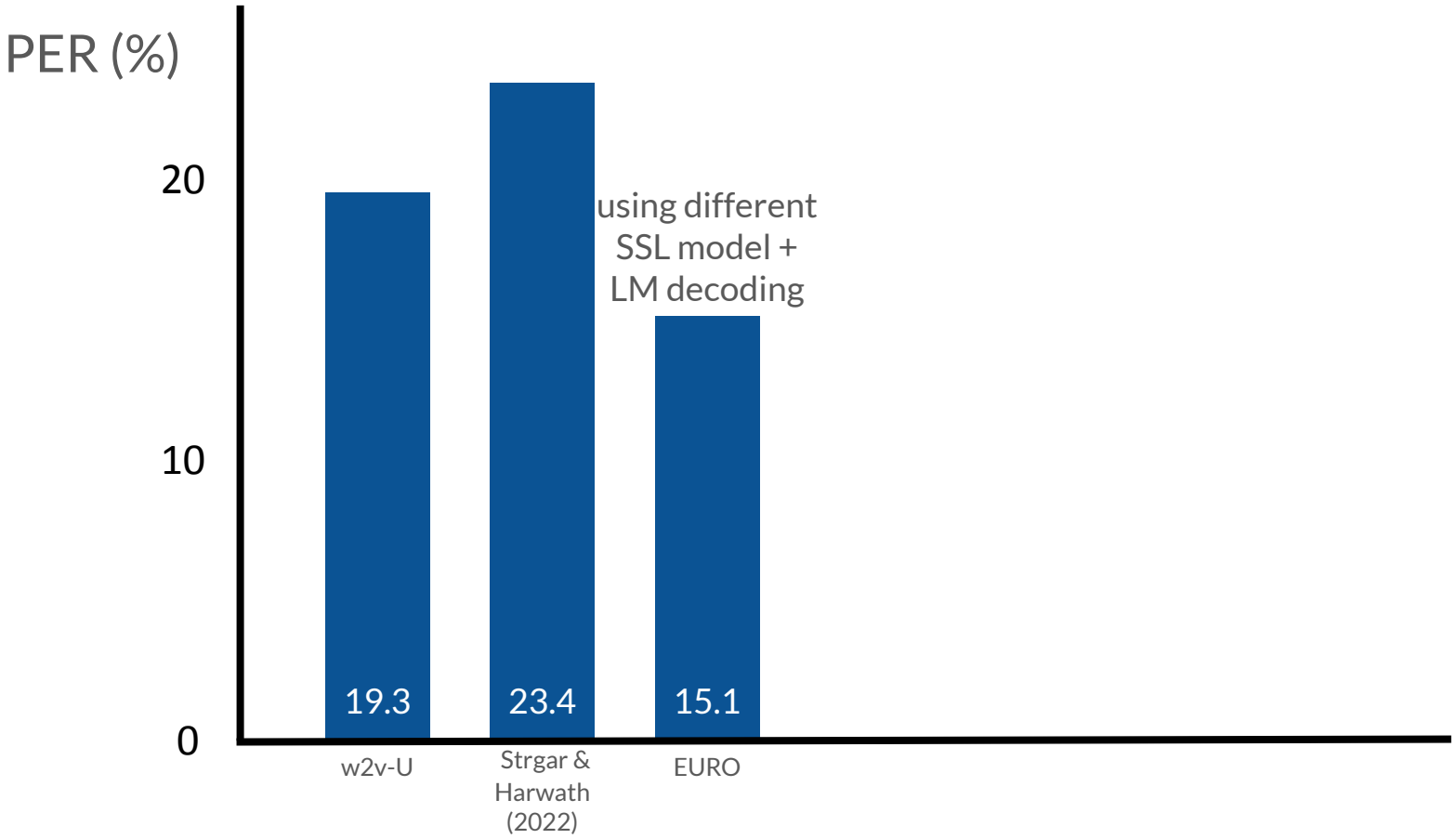
# Result on LibriSpeech

Evaluation metric: PER (%), the lower the better. Training data: 100 hours



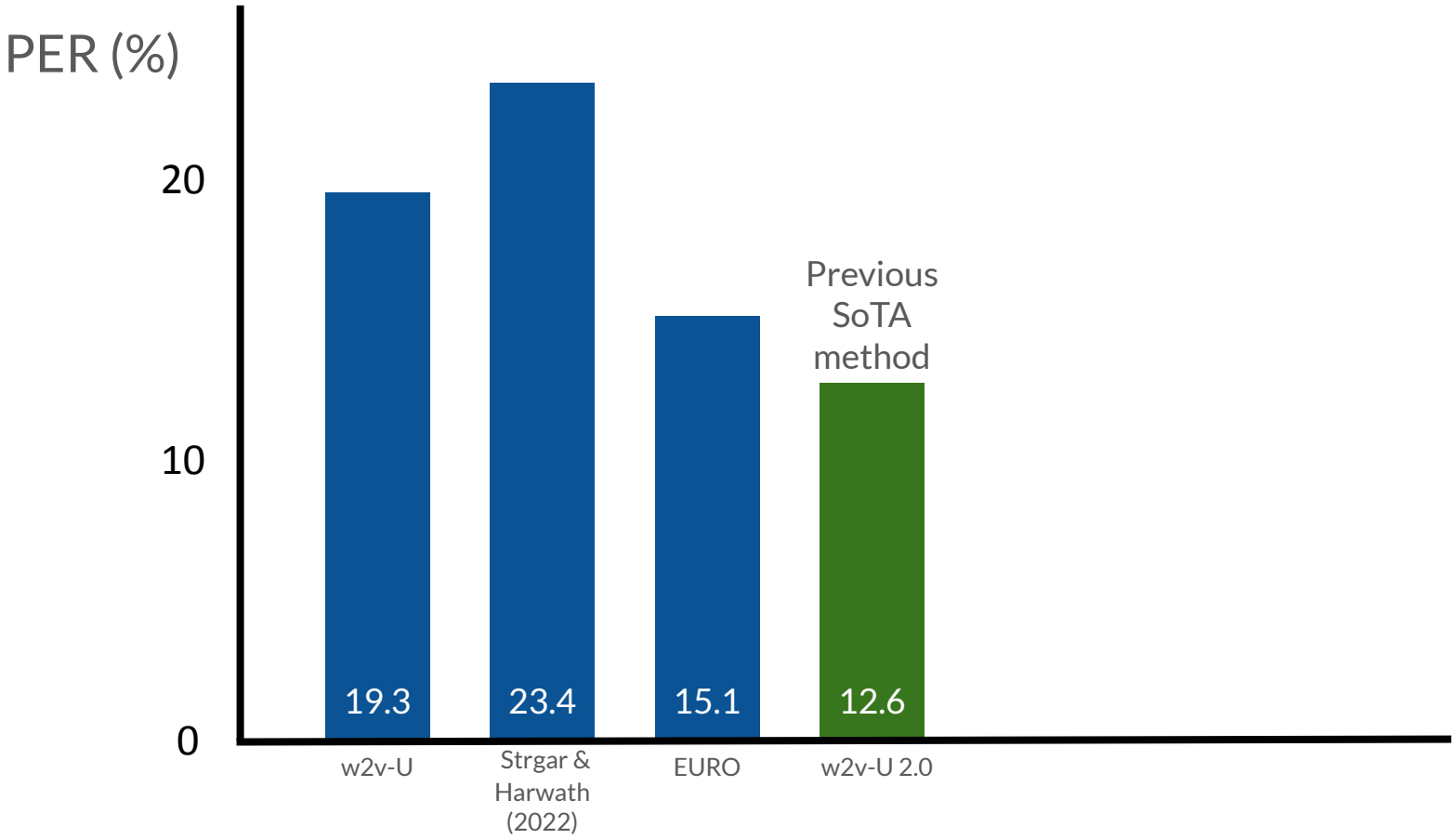
# Result on LibriSpeech

Evaluation metric: PER (%), the lower the better. Training data: 100 hours



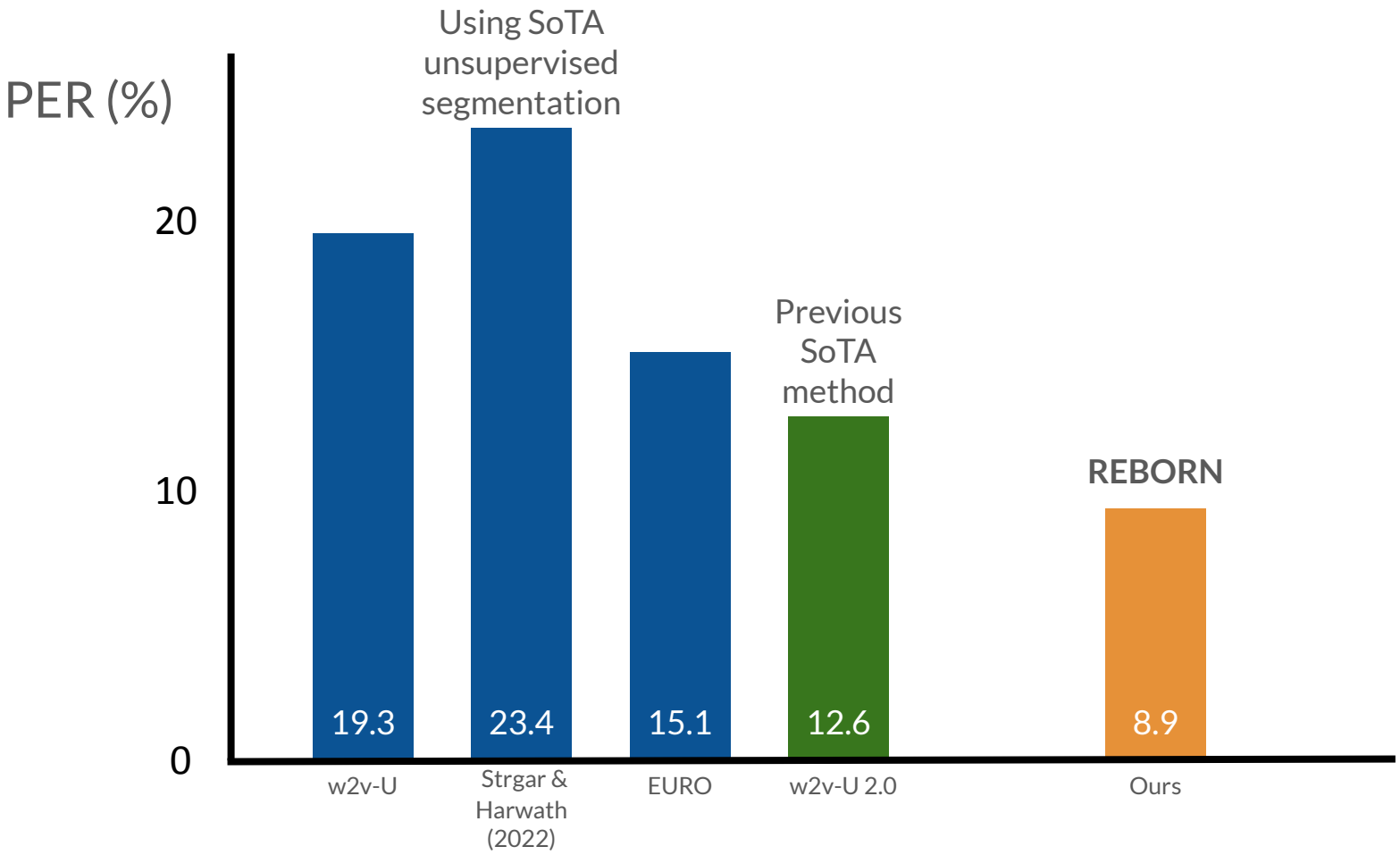
# Result on LibriSpeech

Evaluation metric: PER (%), the lower the better. Training data: 100 hours



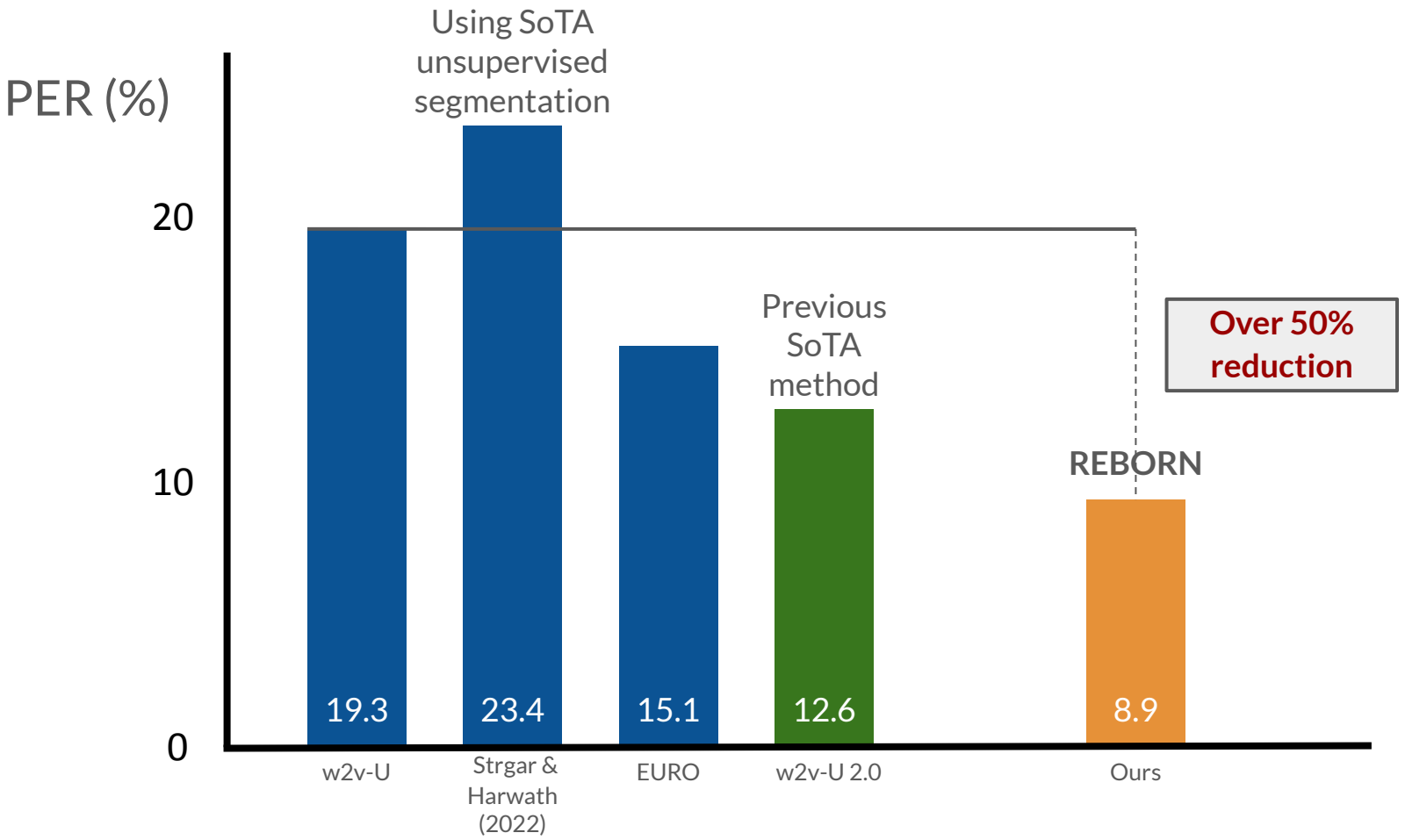
# Result on LibriSpeech

Evaluation metric: PER (%), the lower the better. Training data: 100 hours



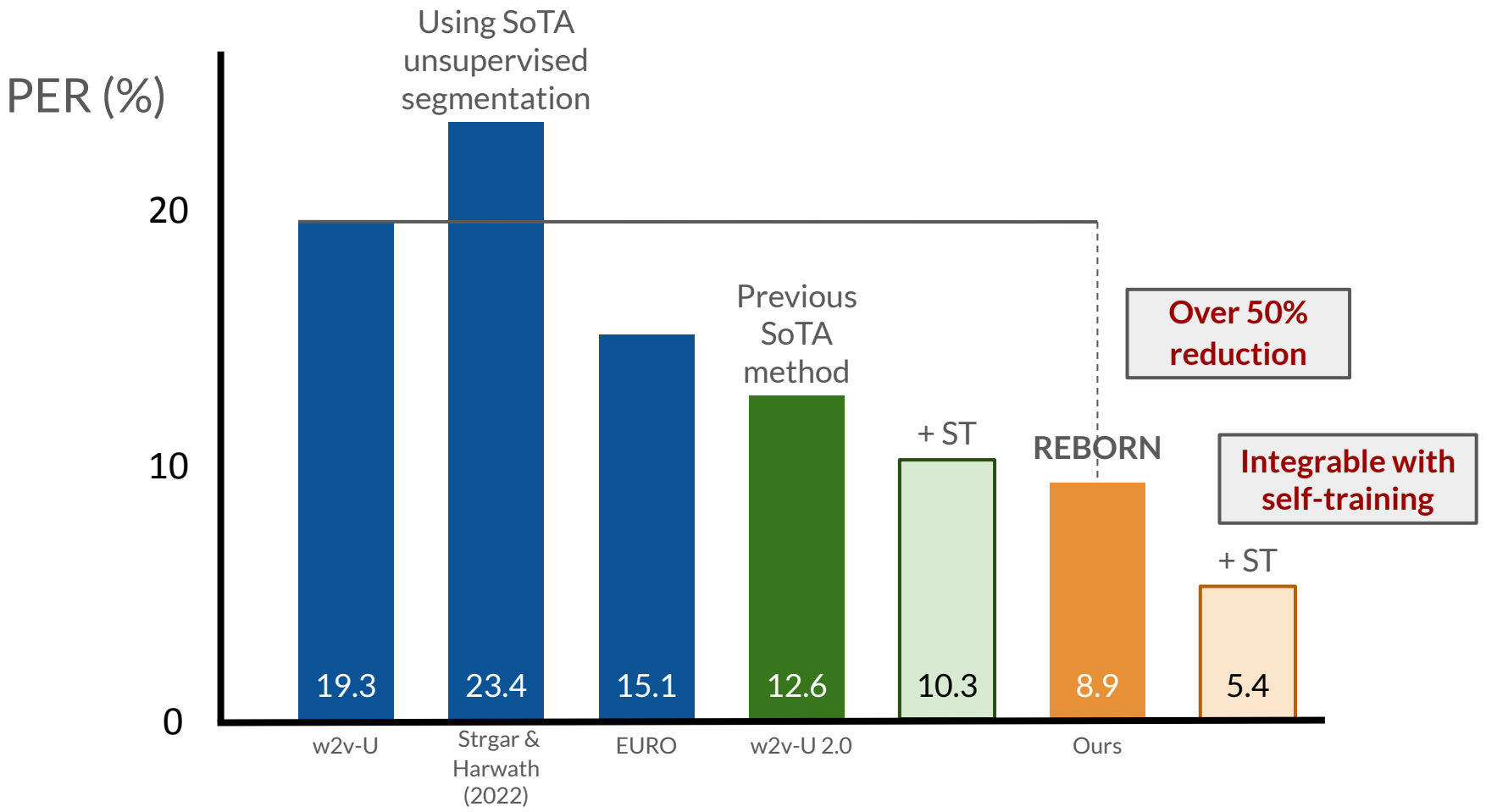
# Result on LibriSpeech

Evaluation metric: PER (%), the lower the better. Training data: 100 hours



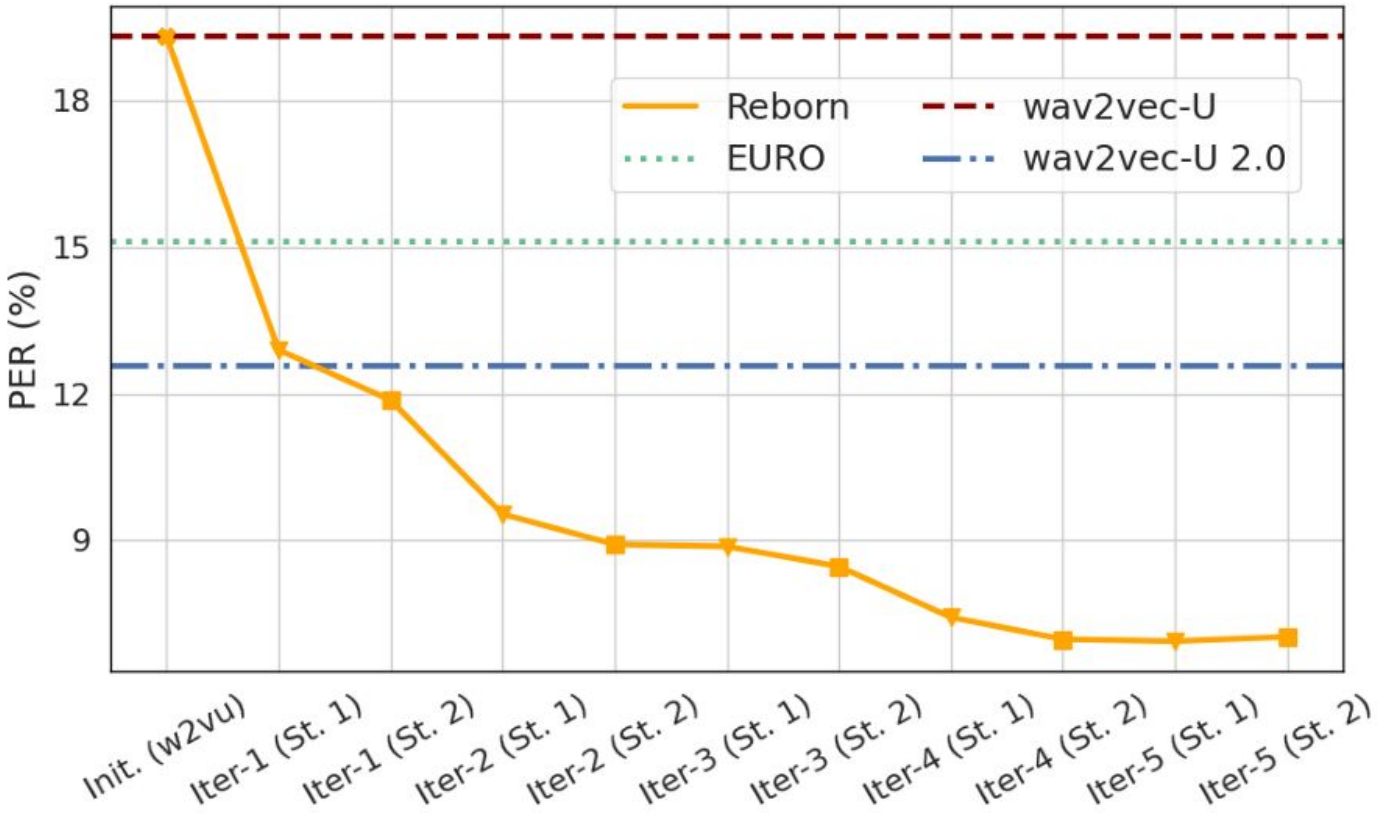
# Result on LibriSpeech

Evaluation metric: PER (%), the lower the better. Training data: 100 hours



# Result of iterative training

REBORN has large performance improvements in the early iterations, and can still gradually improve afterward.



# REBORN generalizes well to other languages

We conduct experiment on Multi-lingual LibriSpeech across 6 languages.

REBORN achieves the best performance on 5 of them.

| Approach | WER (%) ↓ | | | | | | |
|---|---|---|---|---|---|---|---|
| | de | nl | fr | es | it | pt | Avg. |
| wav2vec-U† | 32.5 | 40.2 | 39.8 | 33.3 | 58.1 | 59.8 | 44.0 |
| wav2vec-U* | 33.9 | 38.1 | 37.7 | 33.1 | 51.8 | 59.4 | 42.3 |
| wav2vec-U 2.0‡ | 23.5 | 35.1 | 35.7 | 25.8 | 46.9 | **48.5** | 35.9 |
| REBORN | **20.9** | **26.9** | **28.2** | **24.7** | **39.9** | 51.5 | **32.0** |

# Conclusion

- We propose a method based on RL to learn the segmentation model **tailored for the phoneme predictor**.

# Conclusion

- We propose a method based on RL to learn the segmentation model **tailored for the phoneme predictor**.
- We find out that **relativeness is the key** in our reward design.

# Conclusion

- We propose a method based on RL to learn the segmentation model **tailored for the phoneme predictor**.
- We find out that **relativeness is the key** in our reward design.
- We establish an iterative learning framework that can **polish the segmentation model and phoneme predictor in turn**.

# Conclusion

- We propose a method based on RL to learn the segmentation model **tailored for the phoneme predictor**.
- We find out that **relativeness is the key** in our reward design.
- We establish an iterative learning framework that can **polish the segmentation model and phoneme predictor in turn**.
- Our method generates **excellent or even SoTA performance** on multiple datasets across different languages.

# Thank you!