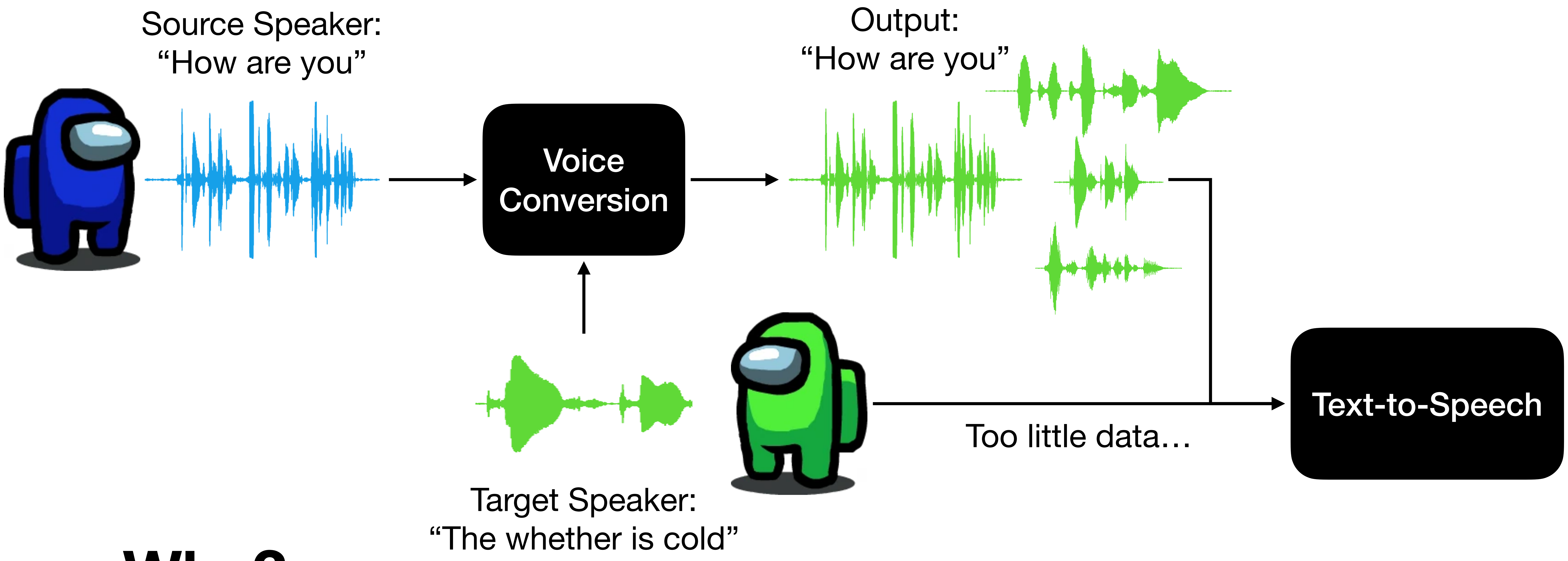
# Self-Supervised Pre-Trained Voice Conversion

Chung-Ming Chien

- * Work done at National Taiwan University
- * Collaborated with Yist Y. Lin, Jheng-Hao Lin, Hung-yi Lee and Lin-shan Lee
- * Published at IEEE ICASSP 2021 & InterSpeech 2021

# Background

# Voice Conversion



Source Speaker:
“How are you”

Voice Conversion

Output:
“How are you”

Target Speaker:
“The whether is cold”

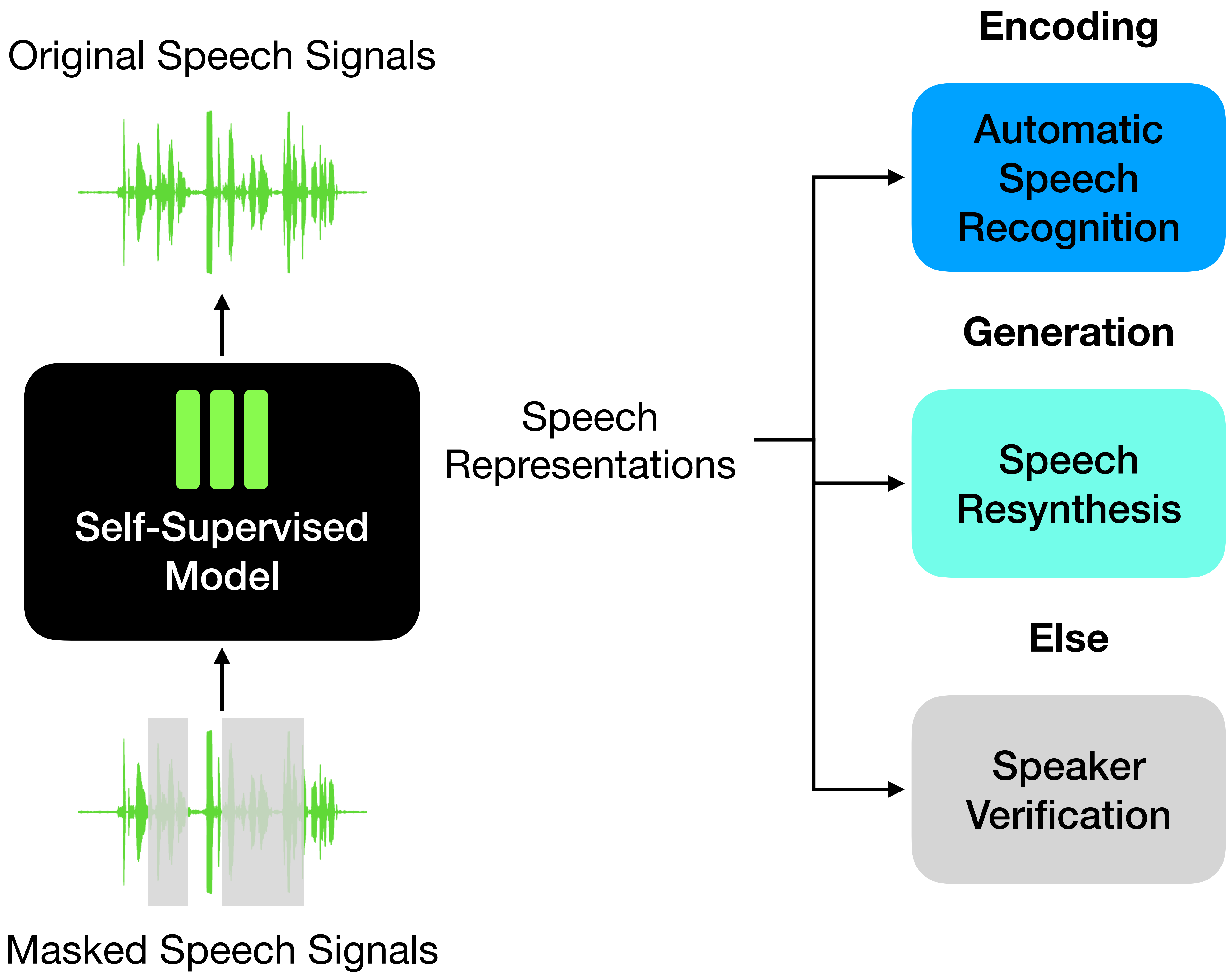
Too little data…

Text-to-Speech

**Why?**

And more applications…

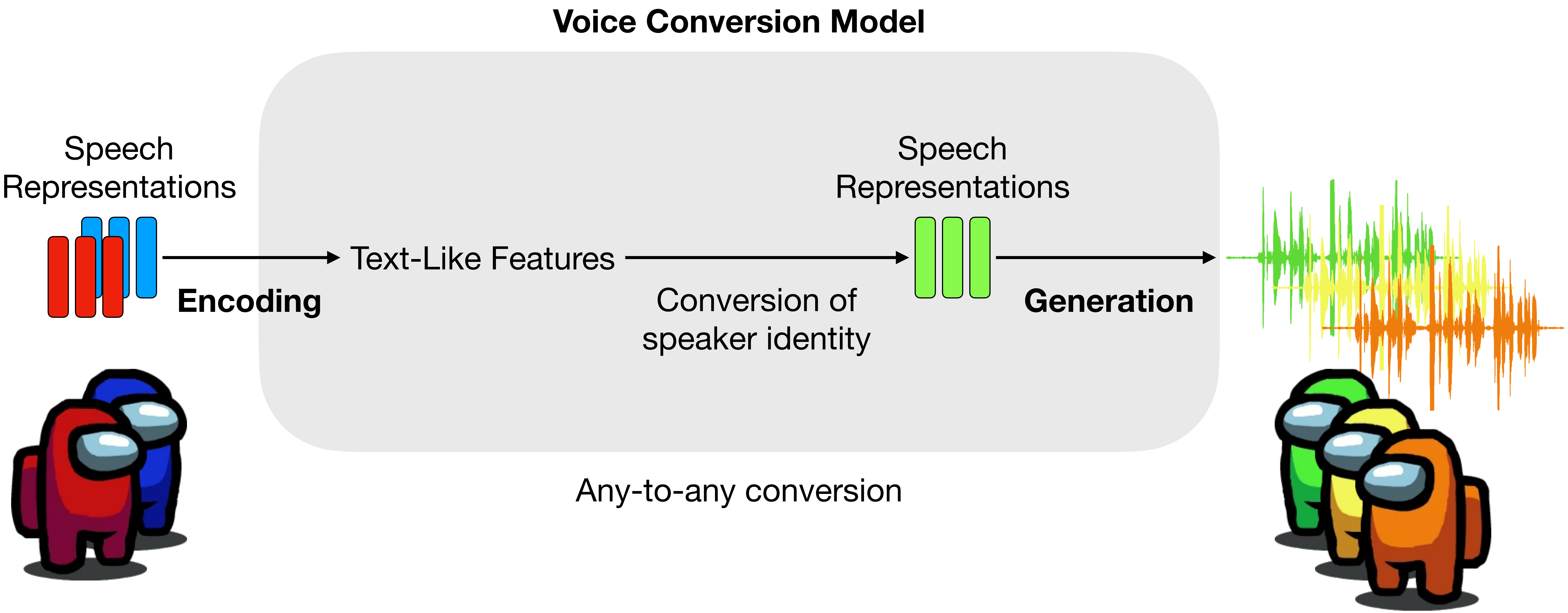
# Self-Supervised Learning (SSL) Representations



SOTA performance with 1 linear layer

➔ text-like

# Proposed: Encoding & Generation in One Model



**Voice Conversion Model**

Speech Representations

Text-Like Features

Speech Representations

**Encoding**

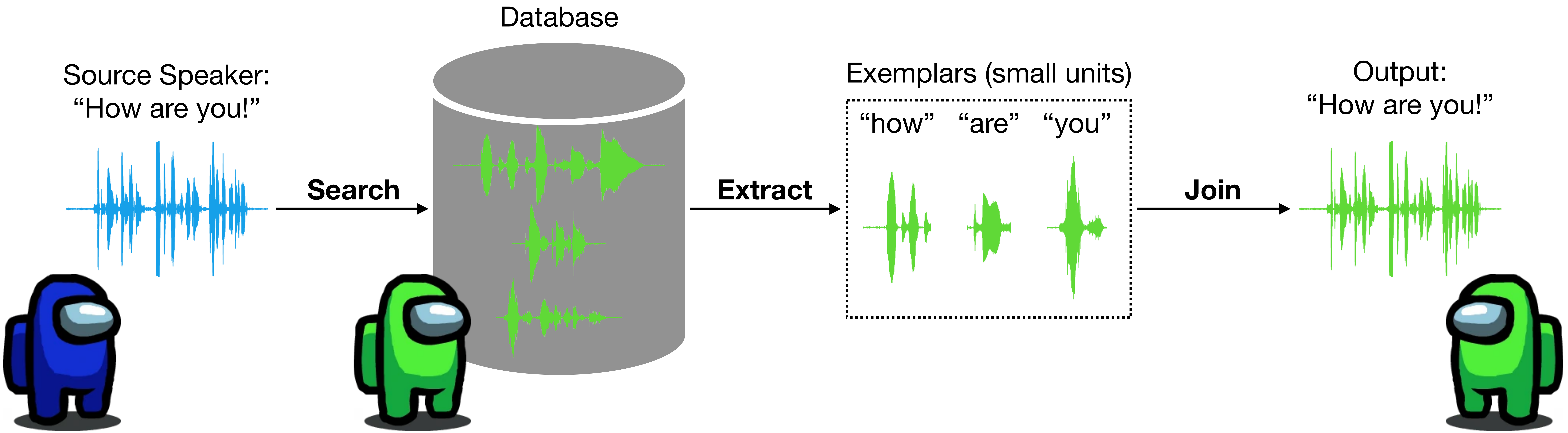
Conversion of speaker identity

**Generation**

Any-to-any conversion

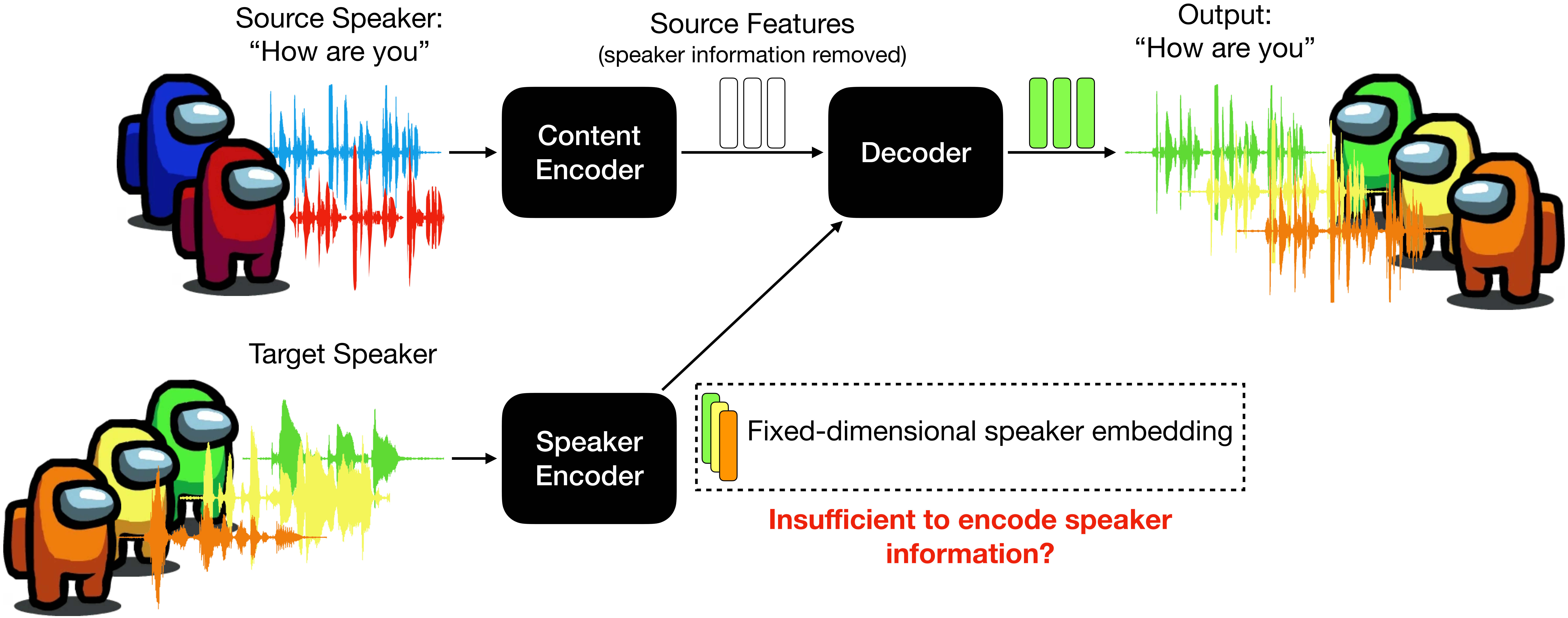
# Prior Arts

# Prior Art 1: Exemplar-Based Voice Conversion



Heavily handcrafted ➔ end-to-end + self-supervised representations

# Prior Art 2: Any-to-Any Voice Conversion



Source Speaker:
"How are you"

Source Features
(speaker information removed)

Output:
"How are you"

Content Encoder

Decoder

Target Speaker

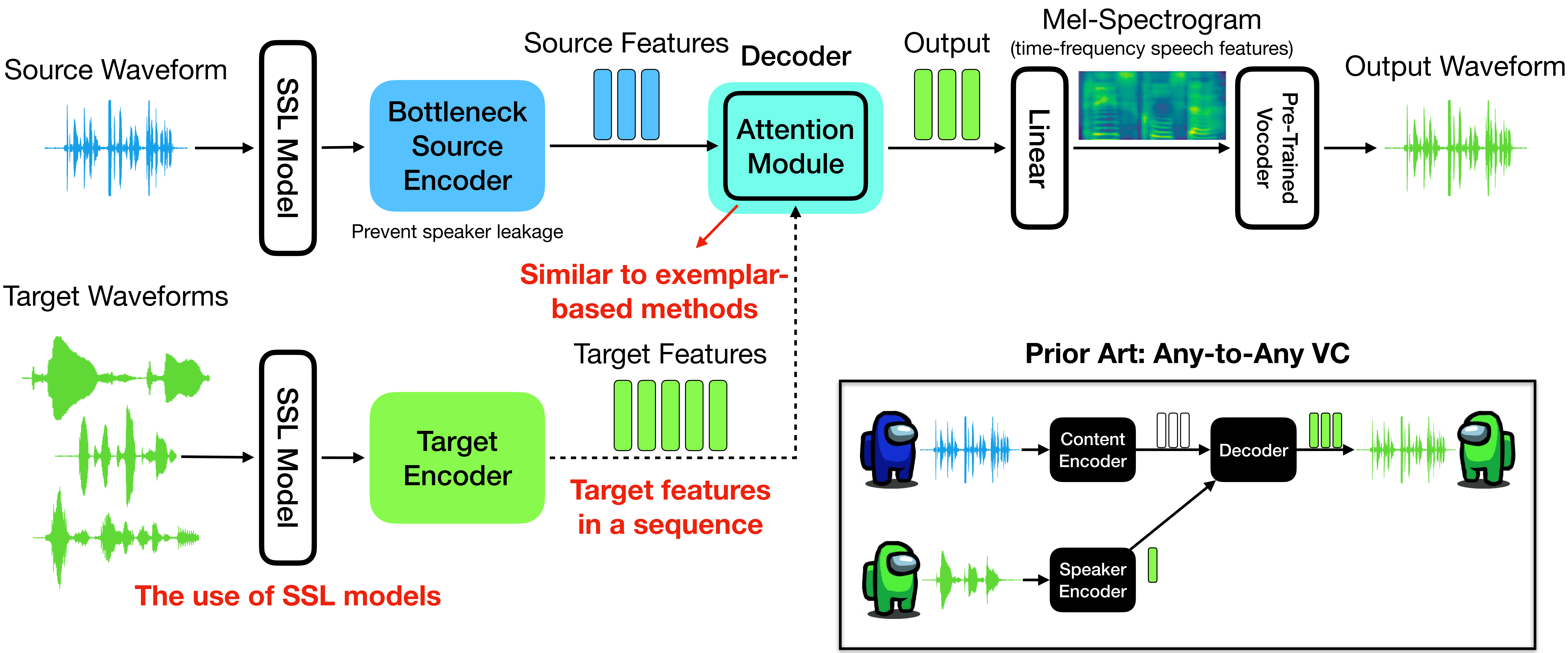
Speaker Encoder

Fixed-dimensional speaker embedding

**Insufficient to encode speaker information?**

# Proposed Methods

# Model Architecture



Source Waveform

SSL Model

Bottleneck Source Encoder

Prevent speaker leakage

Source Features

Decoder

Attention Module

Output

Linear

Mel-Spectrogram

(time-frequency speech features)

Pre-Trained Vocoder

Output Waveform

**Similar to exemplar-based methods**

Target Waveforms

SSL Model

Target Encoder

Target Features

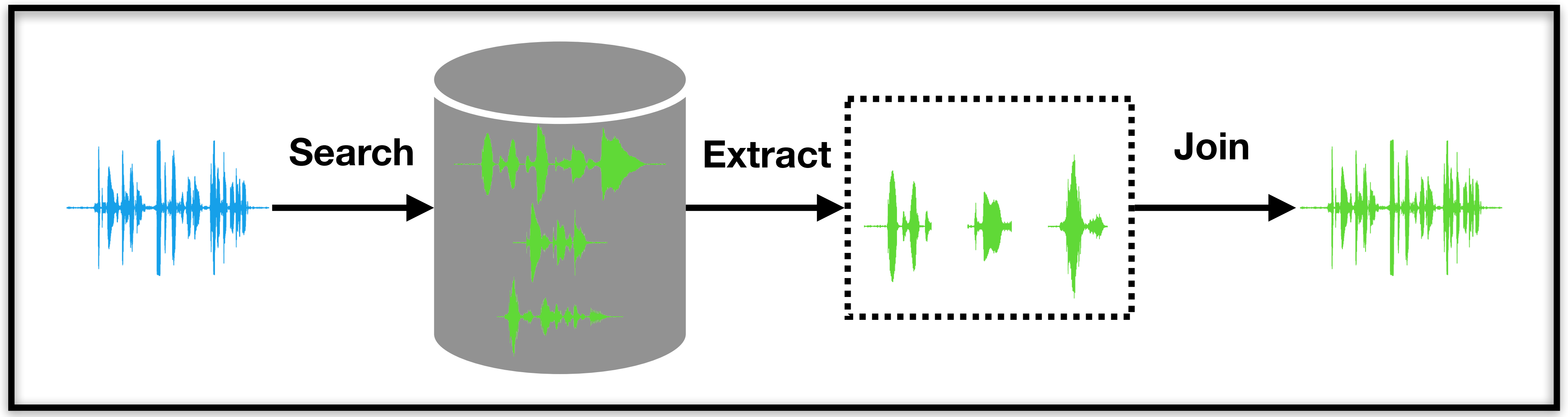
**Target features in a sequence**

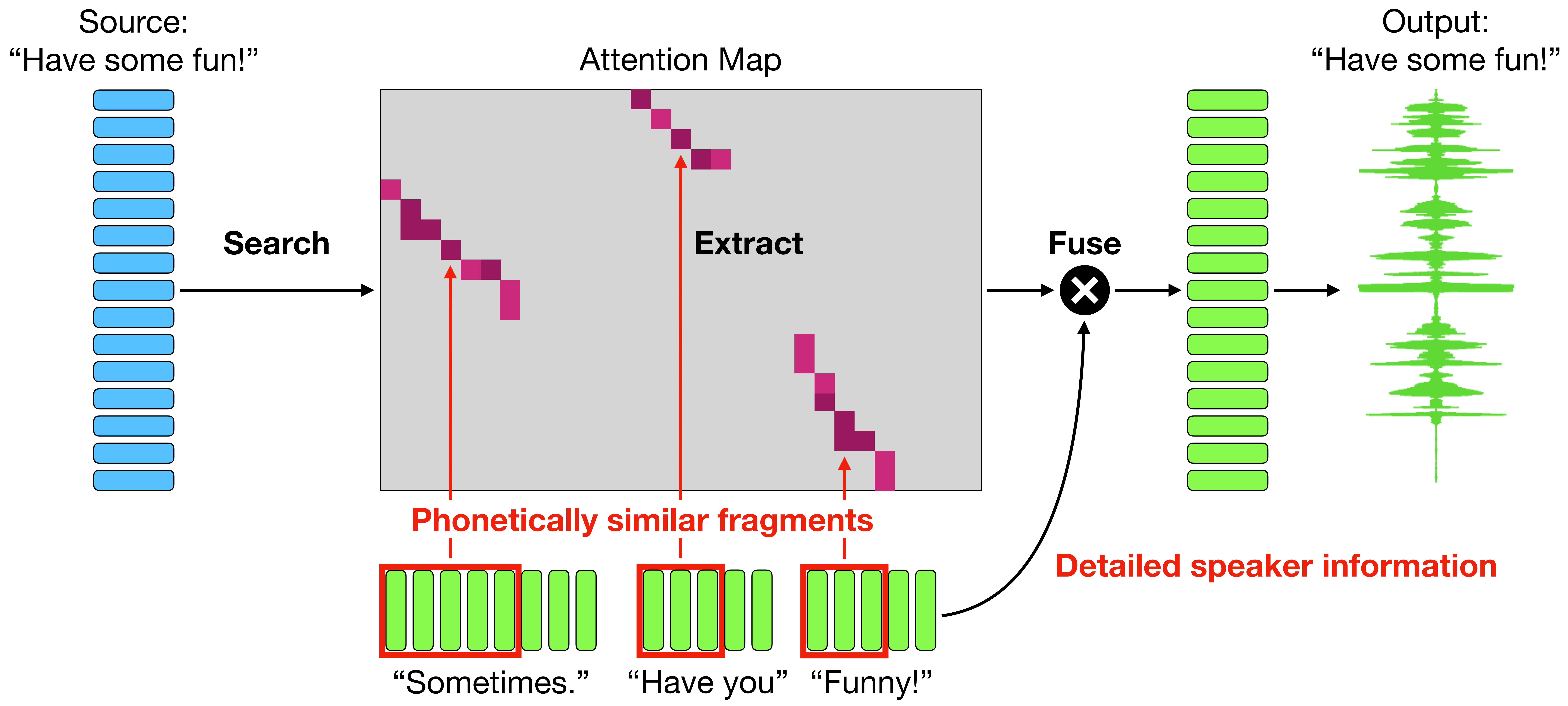
**The use of SSL models**

**Prior Art: Any-to-Any VC**

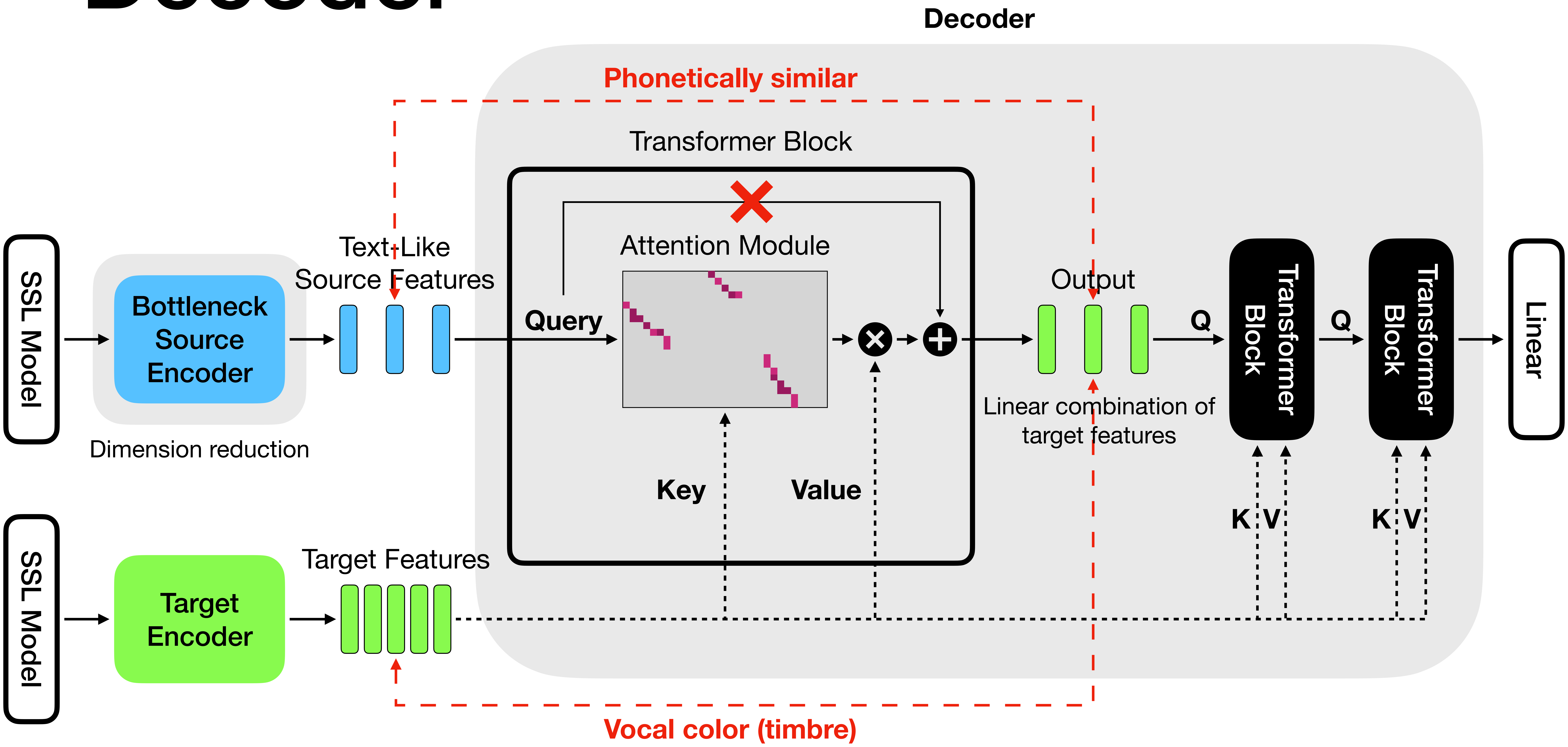
# Attention Module

**Exemplar-based Voice Conversion**

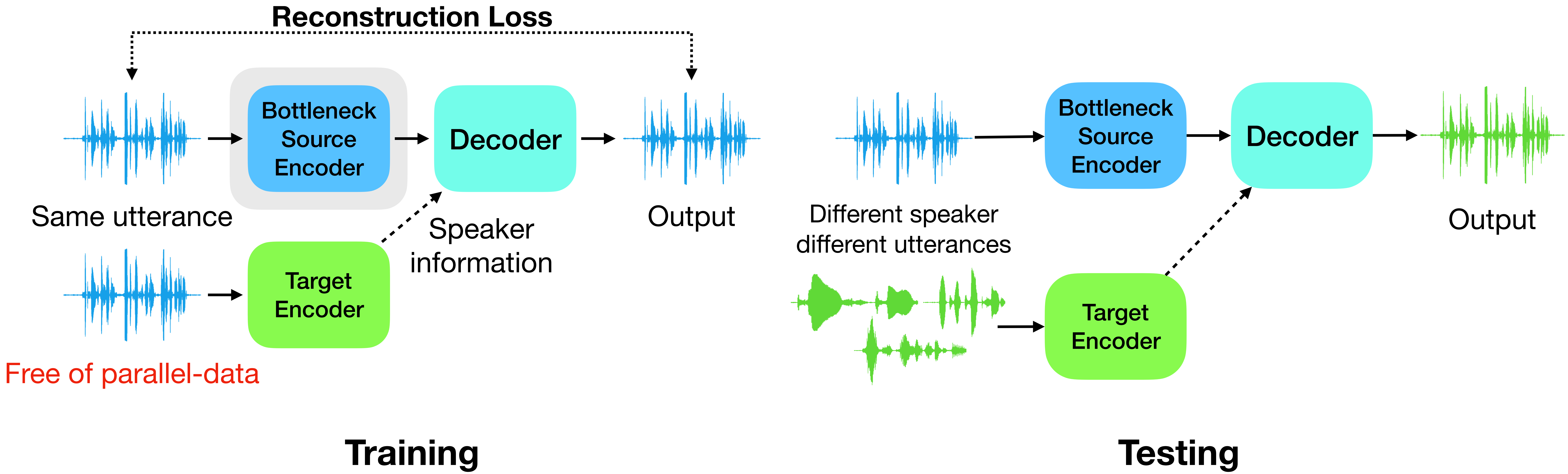




# Decoder


Decoder
Phonetically similar
Transformer Block
Attention Module
SSL Model
Bottleneck Source Encoder
Dimension reduction
Text-Like Source Features
Query
Key
Value
Output
Linear combination of target features
Q
Transformer Block
Q
Transformer Block
Linear
K
V
K
V
SSL Model
Target Encoder
Target Features
Vocal color (timbre)

# Training


Reconstruction Loss
Bottleneck Source Encoder
Decoder
Same utterance
Output
Speaker information
Target Encoder
Free of parallel-data
Training
Bottleneck Source Encoder
Decoder
Different speaker different utterances
Target Encoder
Output
Testing

# Experiments

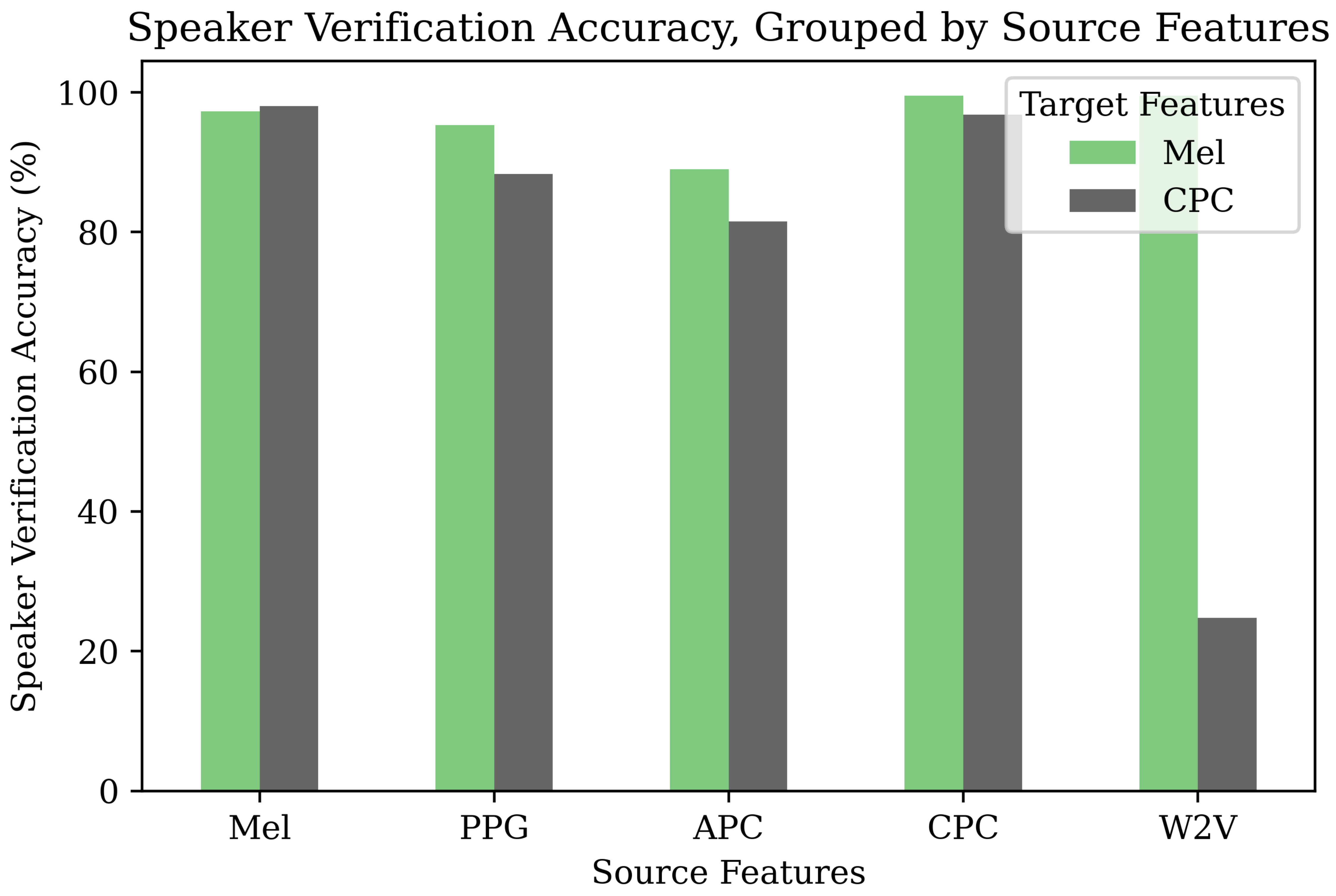
# Experimental Setup

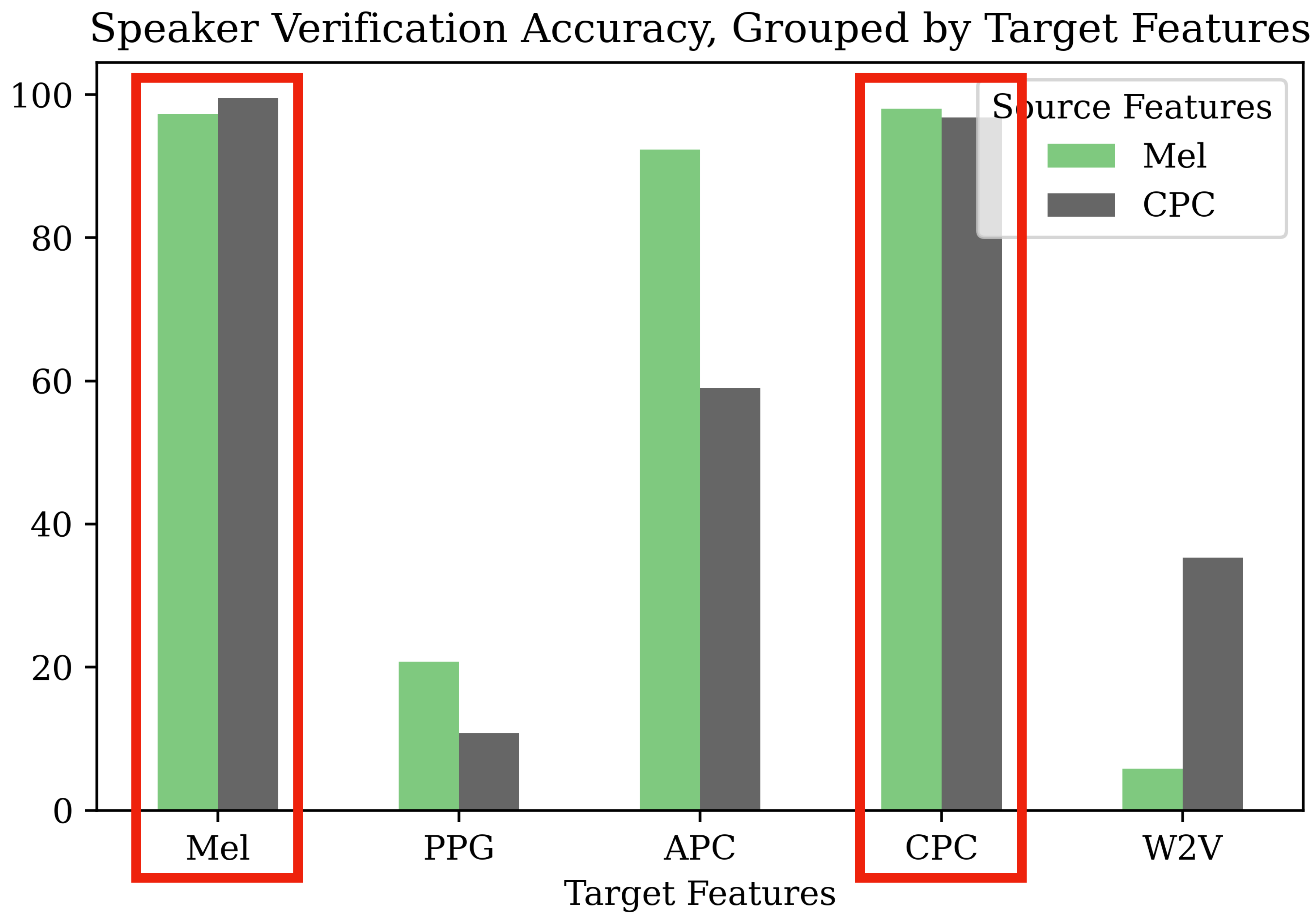
- Training
  - VCTK corpus (109 speakers)
- Testing
  - seen speaker (VCTK)
  - unseen speakers (CMU)
    - one-shot conversion

- Compared SSL Features
  - CPC (contrastive predictive coding)
  - APC (autoregressive predictive coding)
  - Wav2Vec 2.0
- Non SSL Features
  - Mel spectrograms
  - PPG (phoneme posteriorgram trained with text annotations)

# Automatic Speaker Similarity Evaluation

- Off-the-shelf speaker verification system
  - the percentage of outputs passing the system (the higher the better)

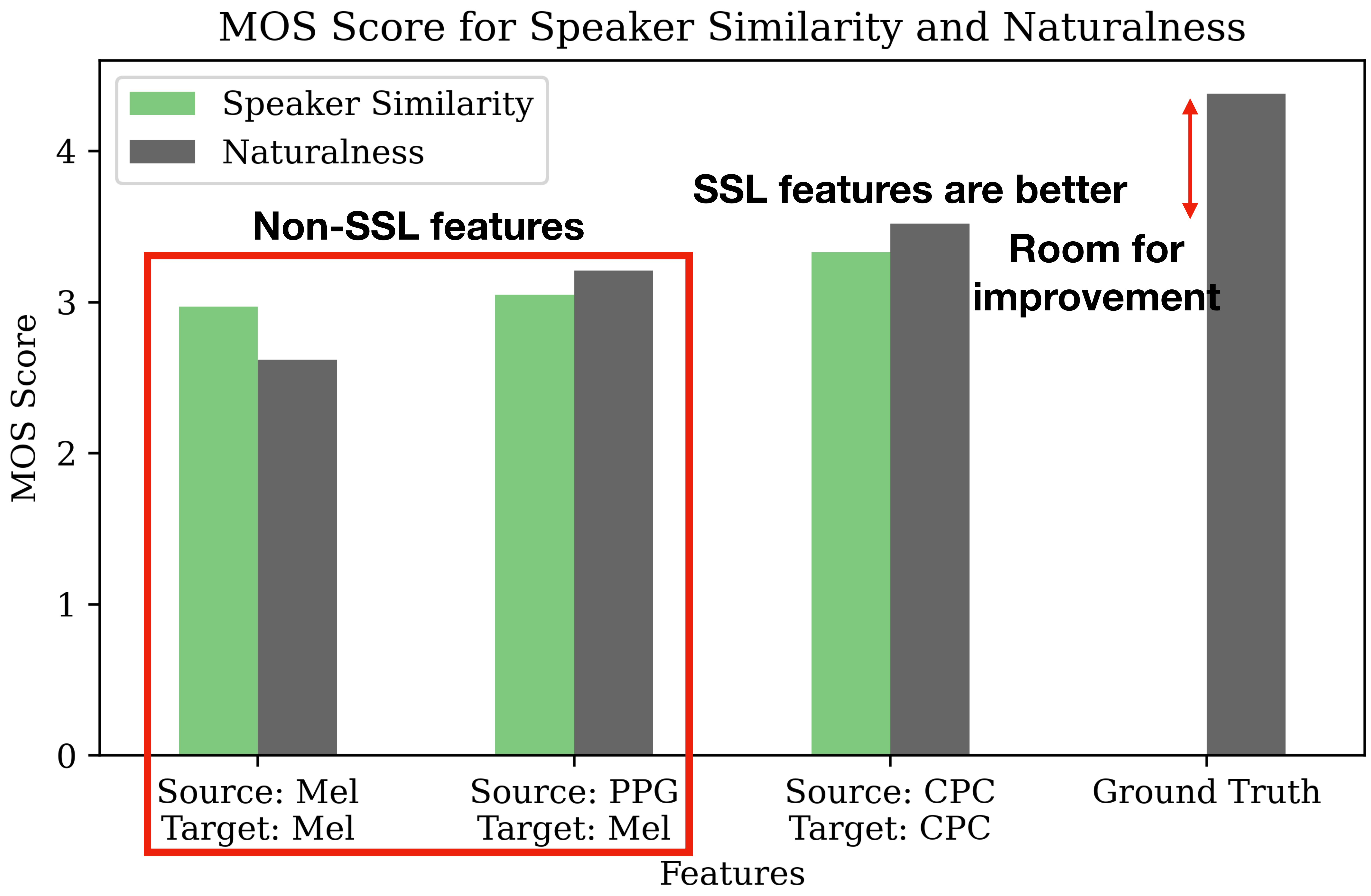




**Target features affect speaker similarity more**

# Subjective Evaluation

- 5-scale Mean Opinion Score (MOS) of synthetic utterances
    - Speaker similarity
    - Naturalness



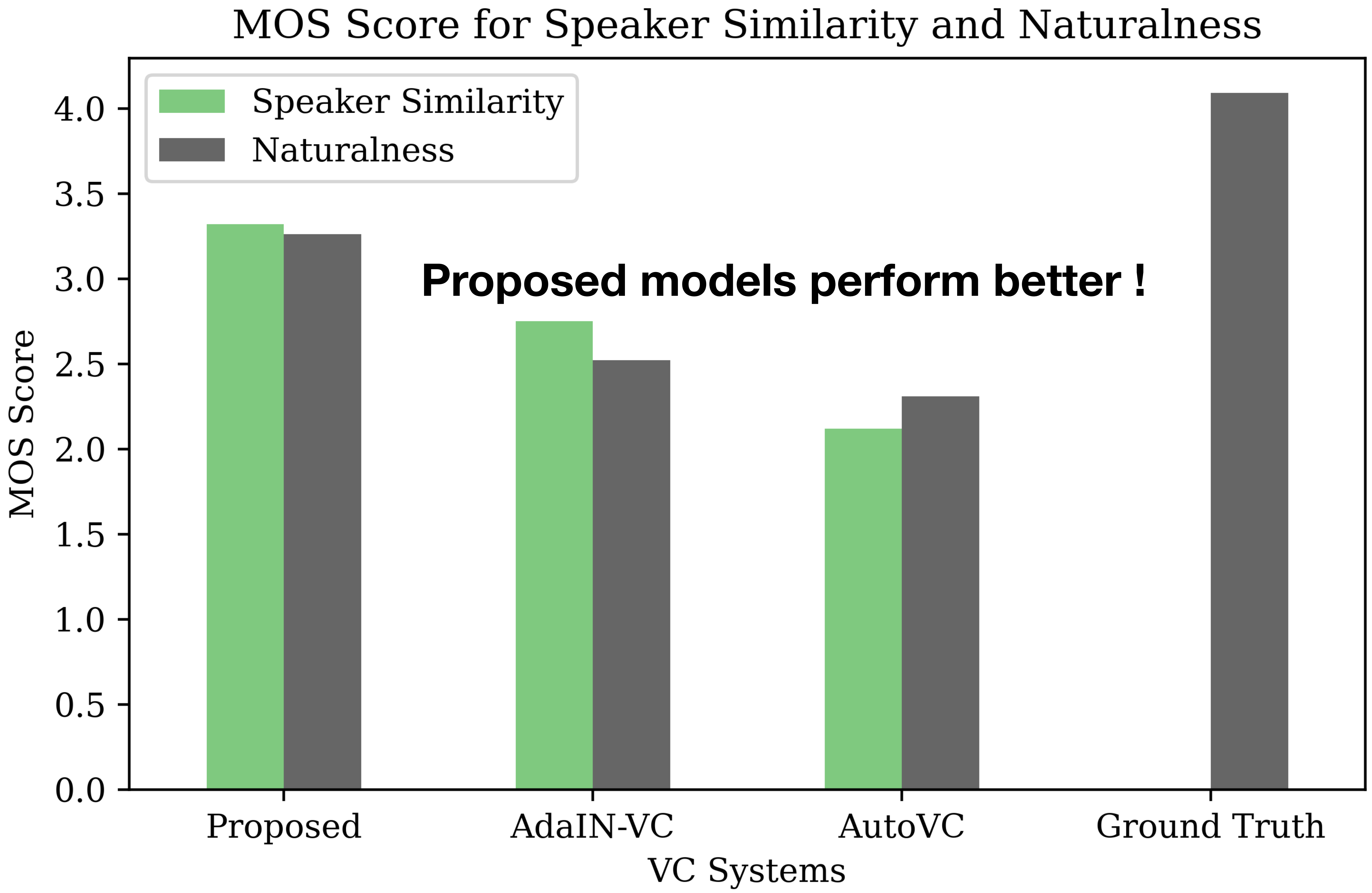
Non-SSL features

SSL features are better

Room for improvement

# Compared with Previous Works

- Compared with previous works that are also
  - One-shot
  - Any-to-any voice conversion
  - Parallel-data-free



**Proposed models perform better !**

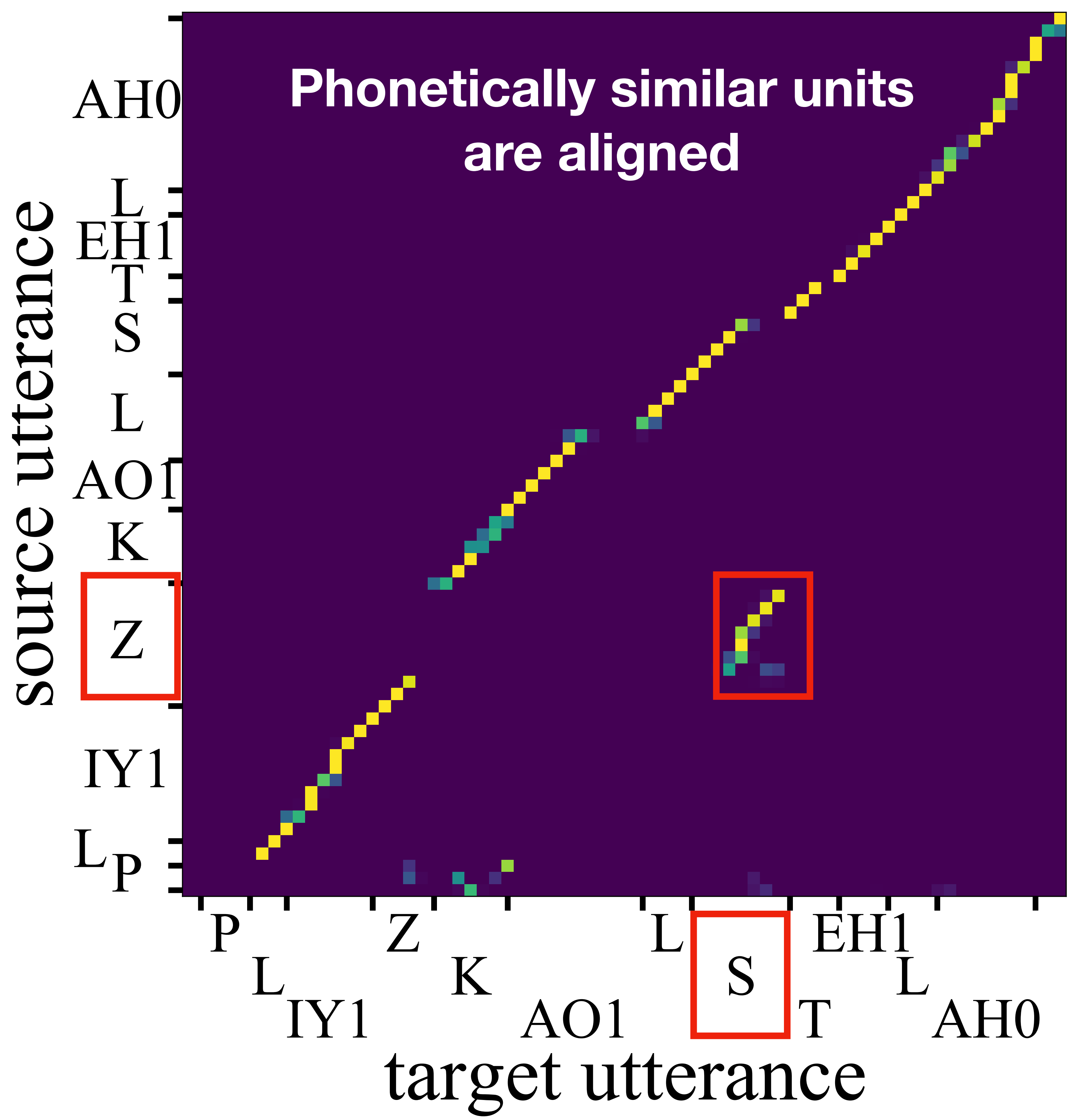
[1] Chou et al., One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization
[2] Qian et al., AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss

# Demo

# Attention Analysis

- Same sentence, different speakers

- Attention map alignment from the Transformer block


Phonetically similar units are aligned

**Source Speaker**
"Please call Stella."

**Target Speaker**
"Please call Stella."

**Converted**
"Please call Stella."

# Conclusion

# Conclusion

- A SOTA approach to **any-to-any** voice conversion
  - **One-shot** and **parallel-data-free**
  - Show the advantage of **sequence speaker features** over fixed-dimensional embeddings
- Combine SSL **encoding & generation** in a voice conversion task **without any annotation**
  - Compare different SSL features
  - SSL features are better than traditional features

# Future Work

- The bottleneck has to be carefully monitored to balance the content correctness and speaker information leakage
  - Better disentanglement of speaker and content information
  - Will discrete SSL features be more text-like?

# Questions?