FragmentVC: Any-to-Any Voice Conversion by End-to-End Extracting and Fusing Fine-Grained Voice Fragments With Attention

Yist Y. Lin*, Chung-Ming Chien*, Jheng-Hao Lin, Hung-yi Lee, Lin-shan Lee

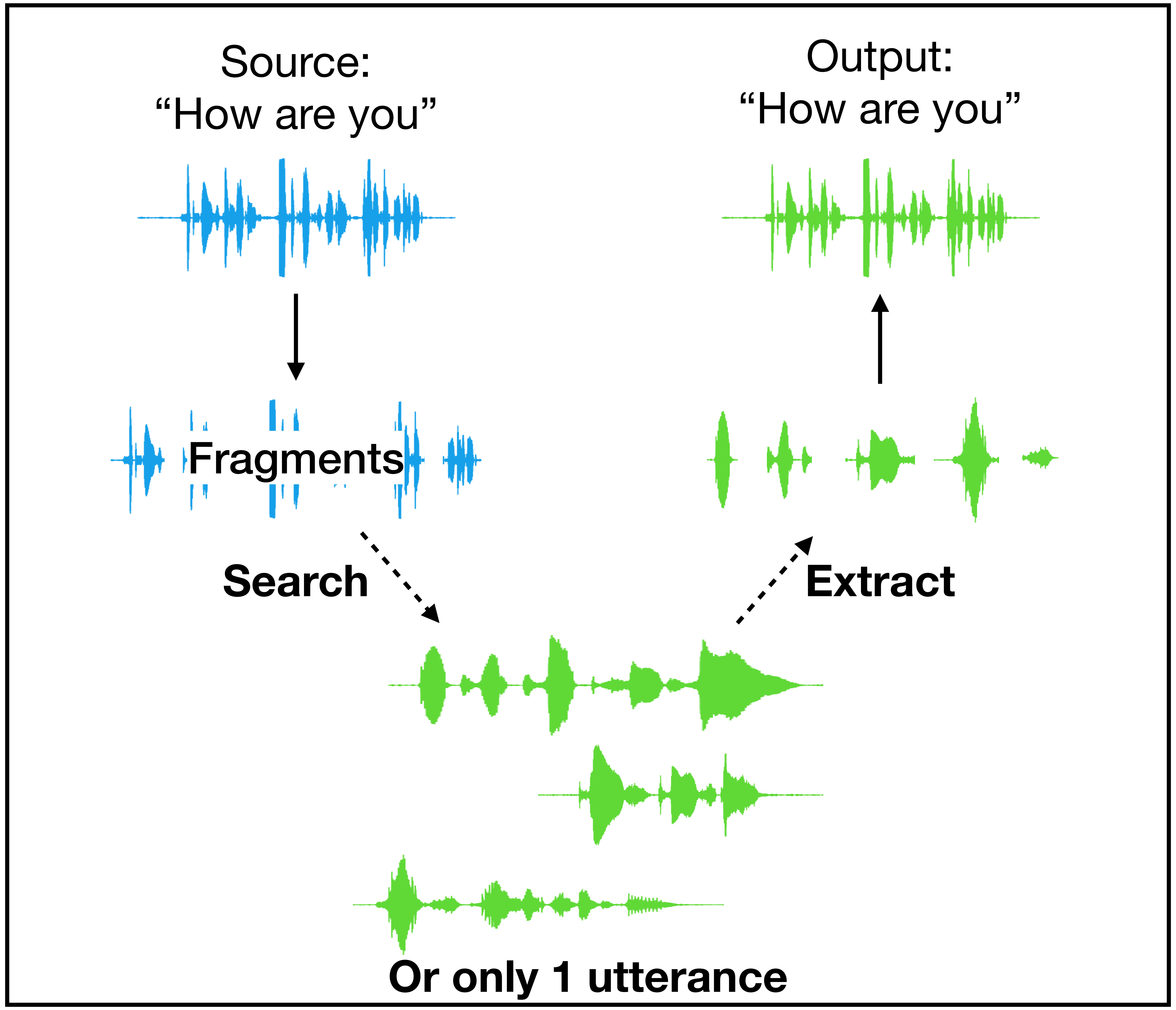
* These authors contributed equally.

IEEE ICASSP 2021

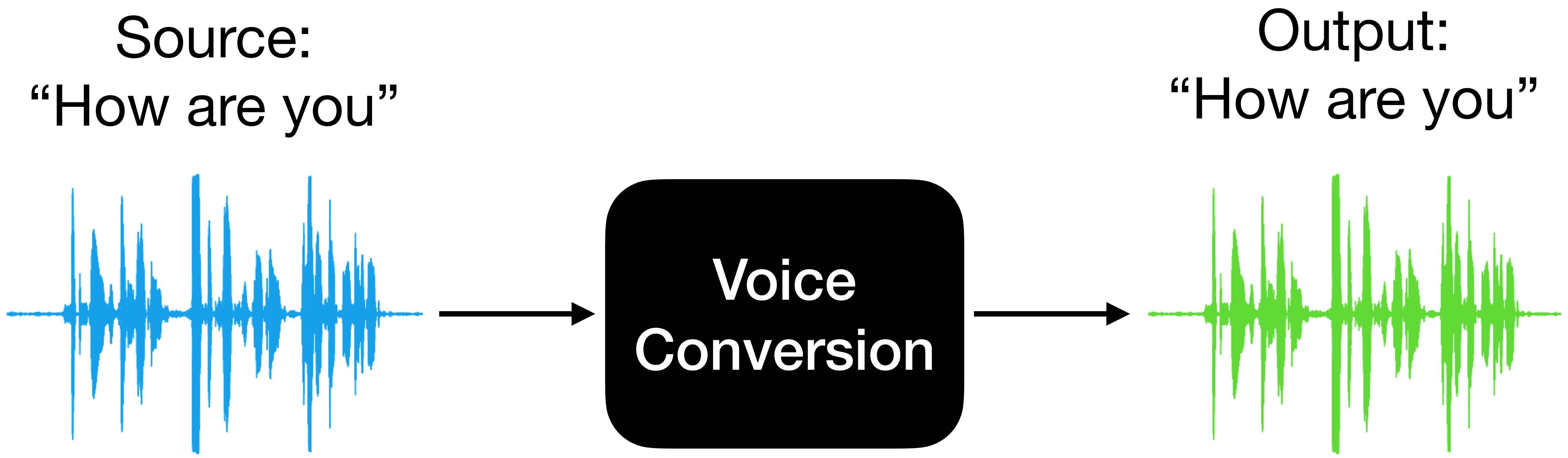
Highlights

- Any-to-any voice conversion
  - One-shot (zero-shot)
  - Parallel-data-free
  - SOTA performance

FragmentVC ⇒ attention + end-to-end



Voice Conversion

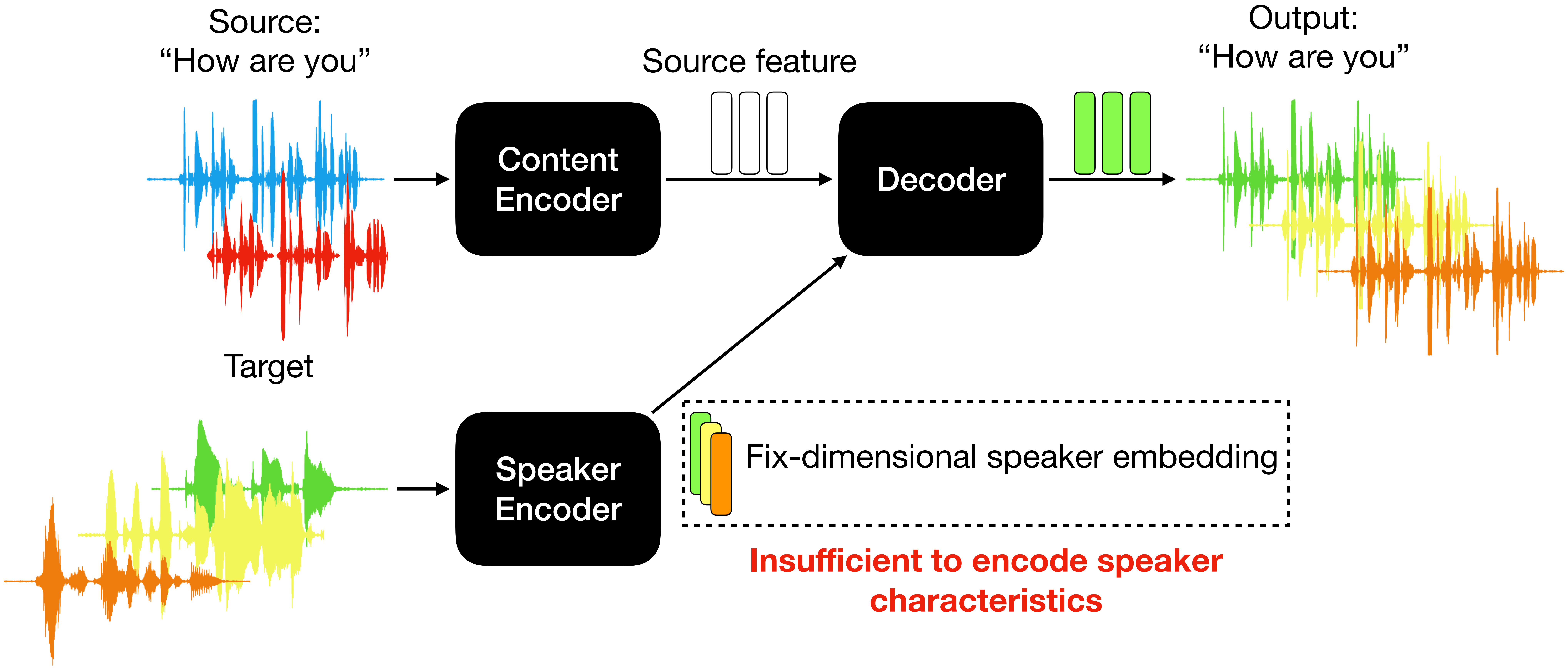


Source:
“How are you”

Voice Conversion

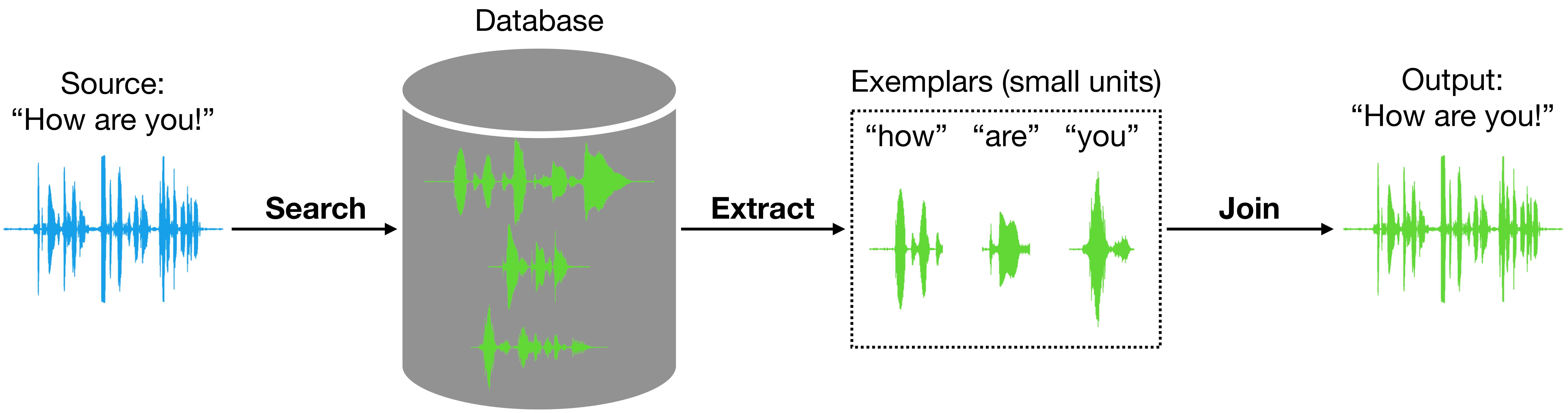
Output:
“How are you”

Prior Art 1: Any-to-Any Voice Conversion



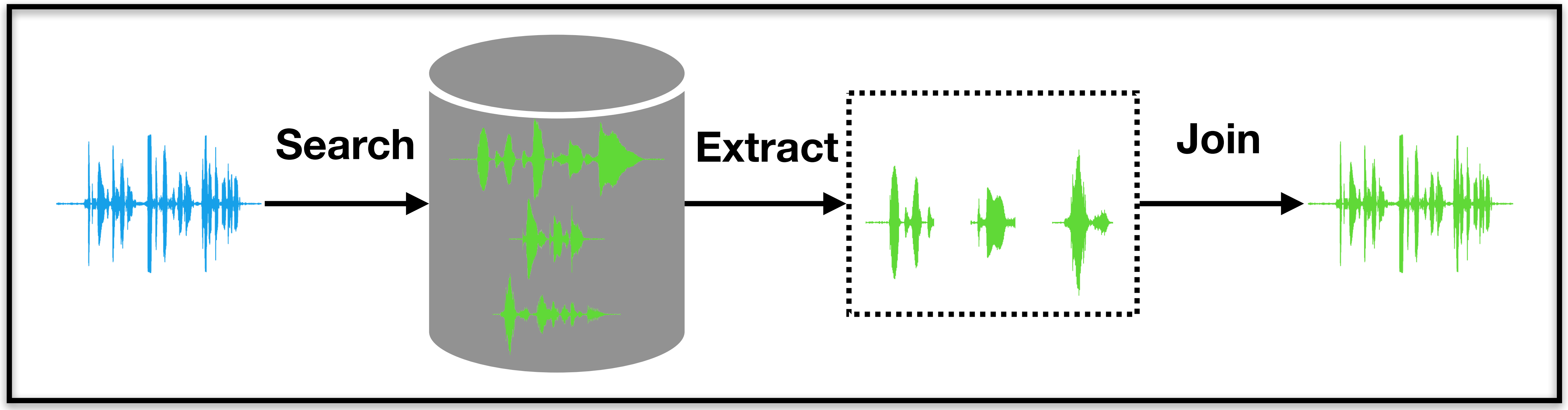
Insufficient to encode speaker characteristics

Prior Art 2: Exemplar-based Voice Conversion

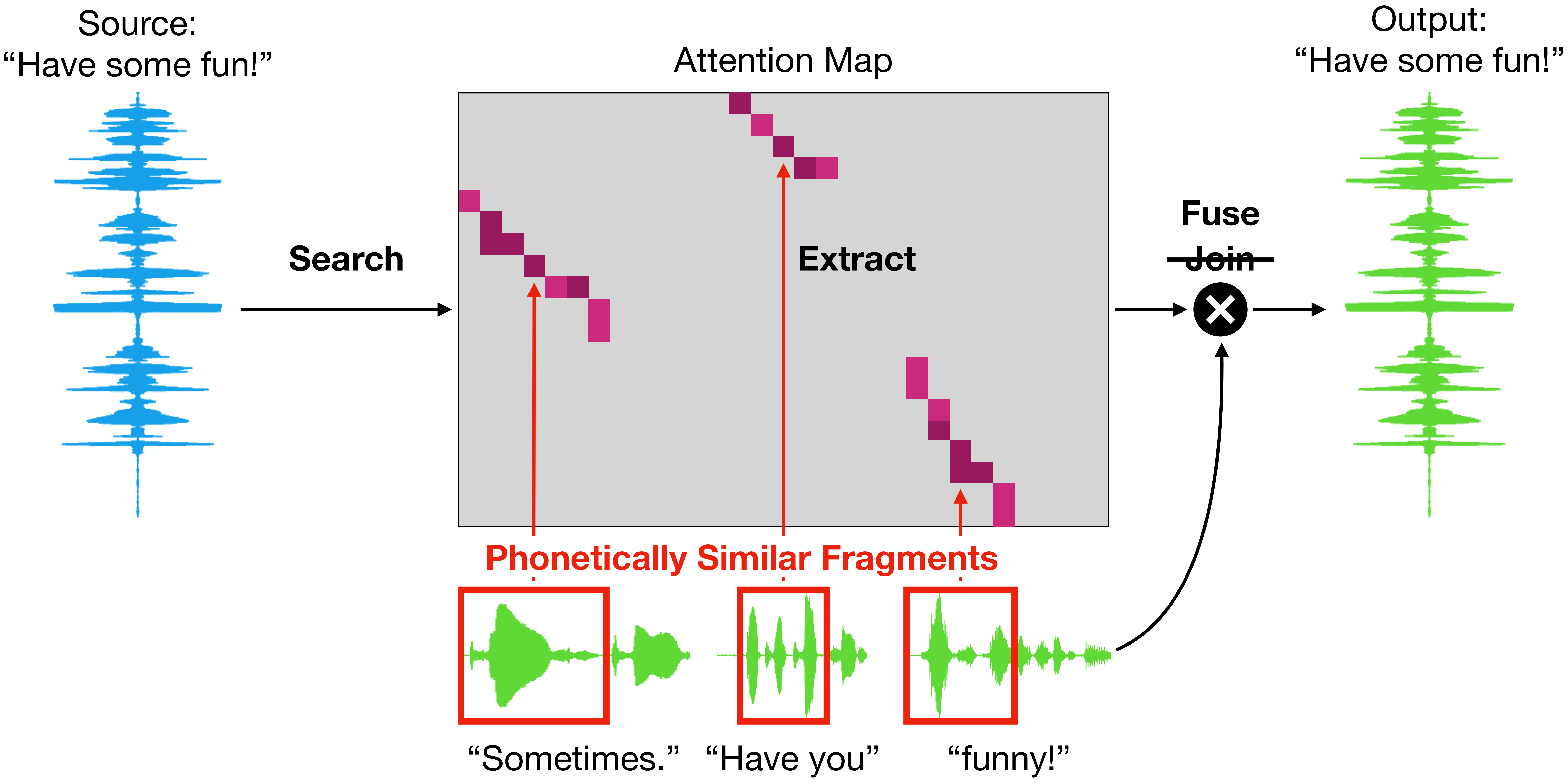


Heavily handcrafted ⇒ DNN (attention) + end-to-end

Illustration



Exemplar-based Voice Conversion


Source:
“Have some fun!”

Attention Map

Output:
“Have some fun!”

Search

Extract

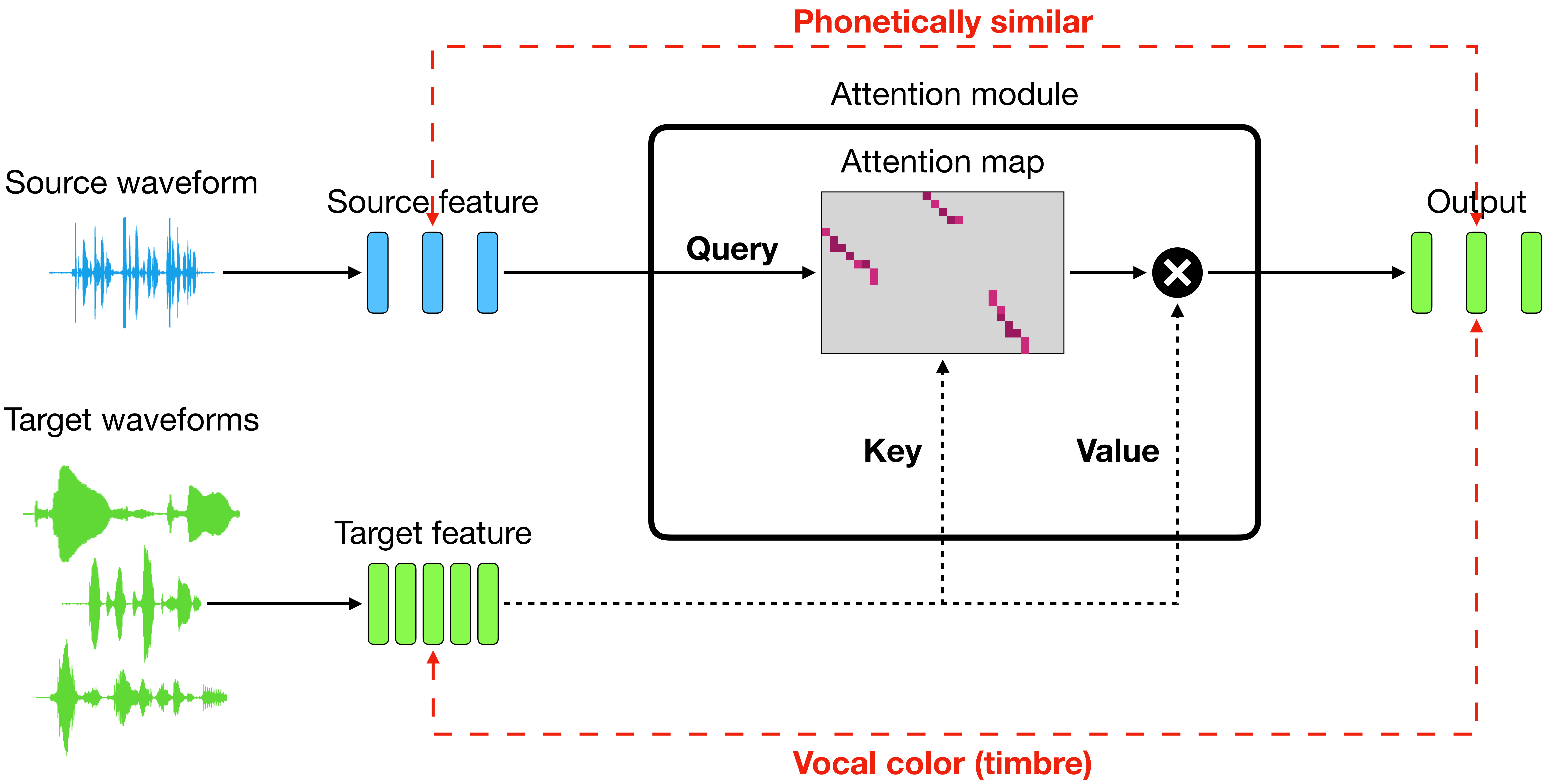
Fuse

~~**Join**~~

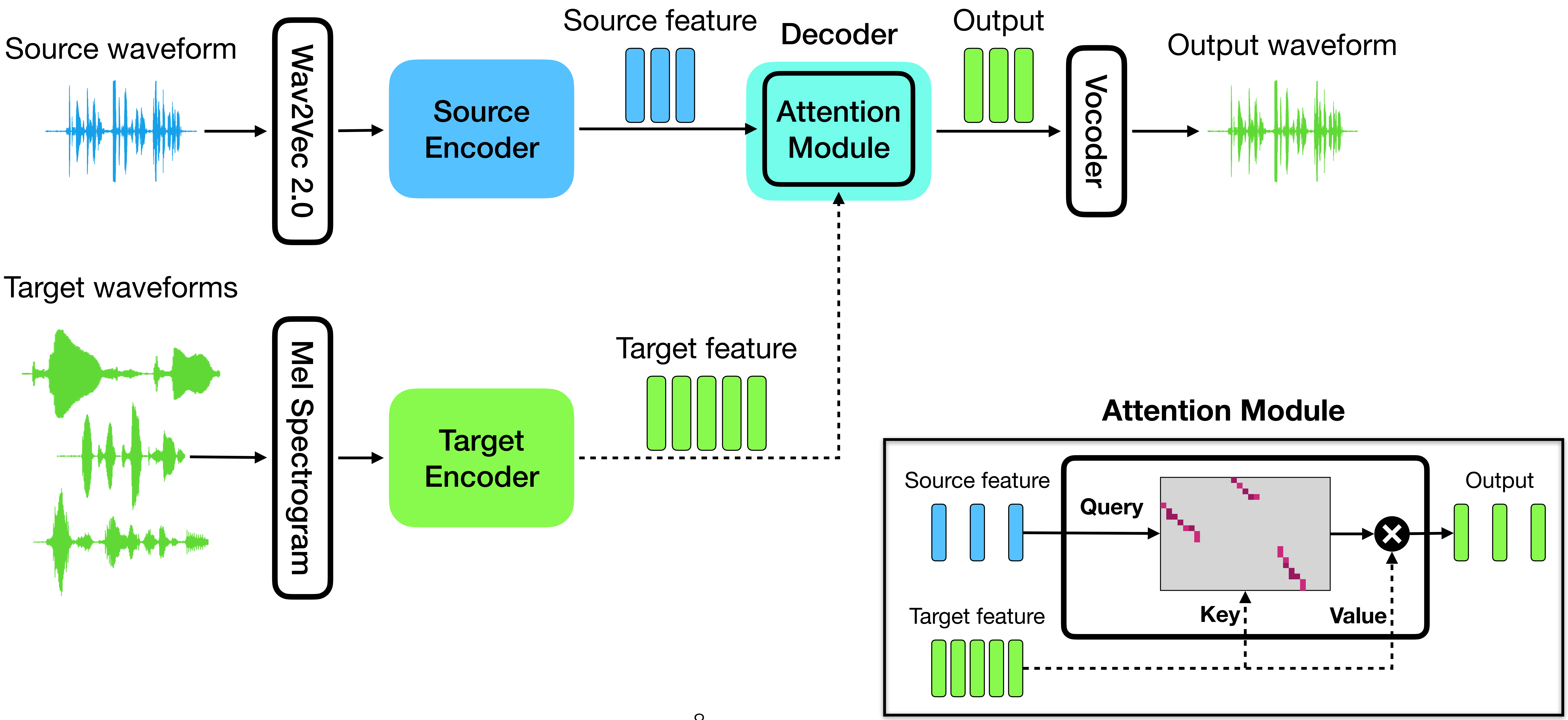
Phonetically Similar Fragments

“Sometimes.” “Have you” “funny!”

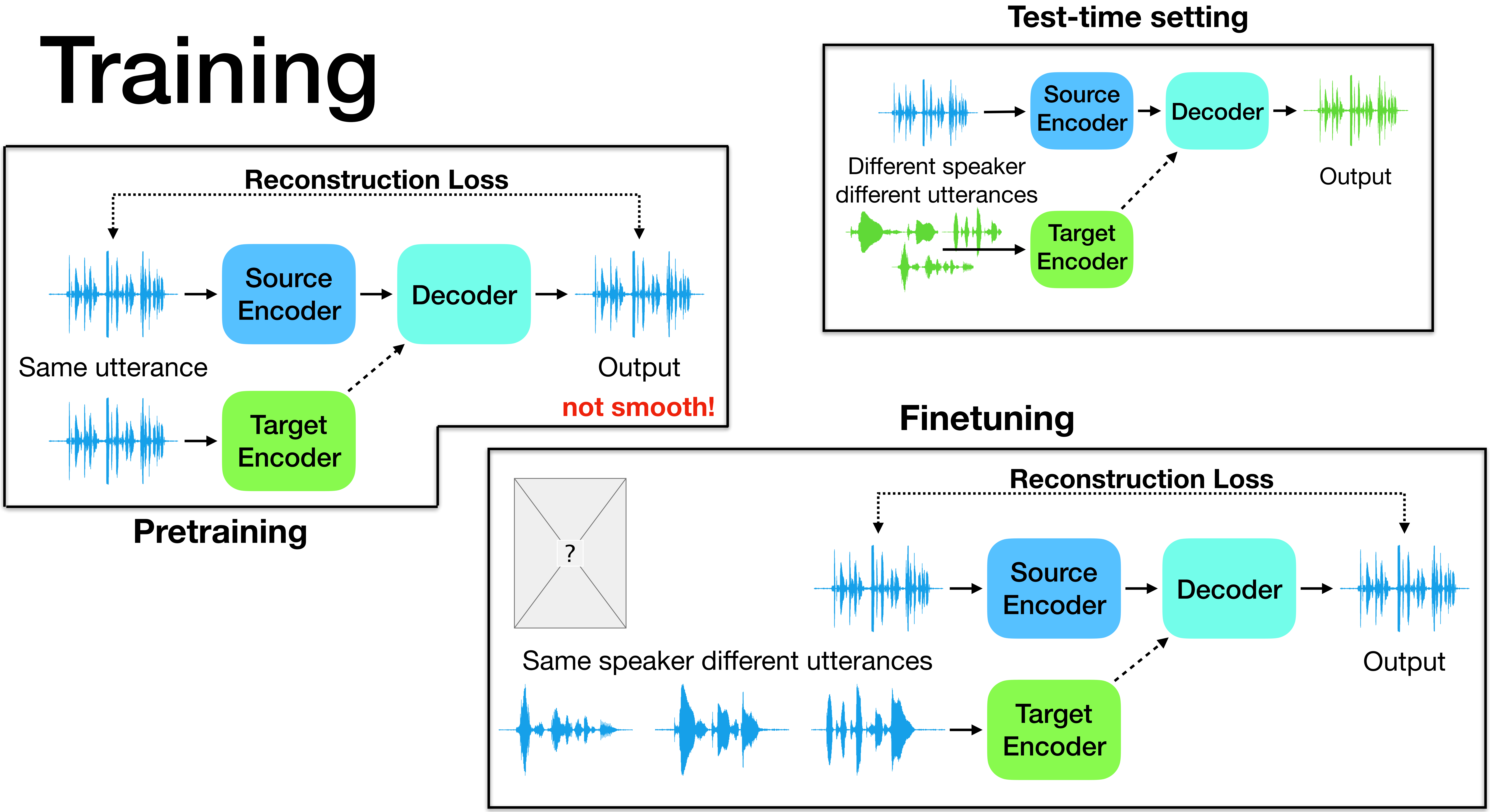
Inside the Attention Module



Model Architecture



Training


Test-time setting
Source Encoder
Decoder
Different speaker different utterances
Target Encoder
Output
Reconstruction Loss
Source Encoder
Decoder
Same utterance
Output
Target Encoder
not smooth!
Pretraining
Finetuning
Reconstruction Loss
?
Source Encoder
Decoder
Same speaker different utterances
Output
Target Encoder

Experimental Setup

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

Automatic Speaker Similarity Evaluation

- Speaker similarity: outputs ⇔ target speakers' utterances
- Off-the-shelf speaker verification system
  - The percentage of outputs passing the system (the higher the better)

	Proposed	Proposed w/o finetune	AdaIN-VC [1]	AUTOVC [2]
seen-to-seen	94.8	94.7	97.8	39.3
unseen-to-unseen	92.5	99.8	87.1	19.0

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

Subjective Evaluation

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

	Proposed	Proposed w/o finetune	AdaIN-VC	AutoVC	Auth.
Speaker similarity	3.32 ± 0.15	**3.81** ± 0.15	2.75 ± 0.15	2.12 ± 0.14	-
Naturalness	**3.26** ± 0.12	2.73 ± 0.11	2.52 ± 0.12	2.31 ± 0.12	4.09 ± 0.12

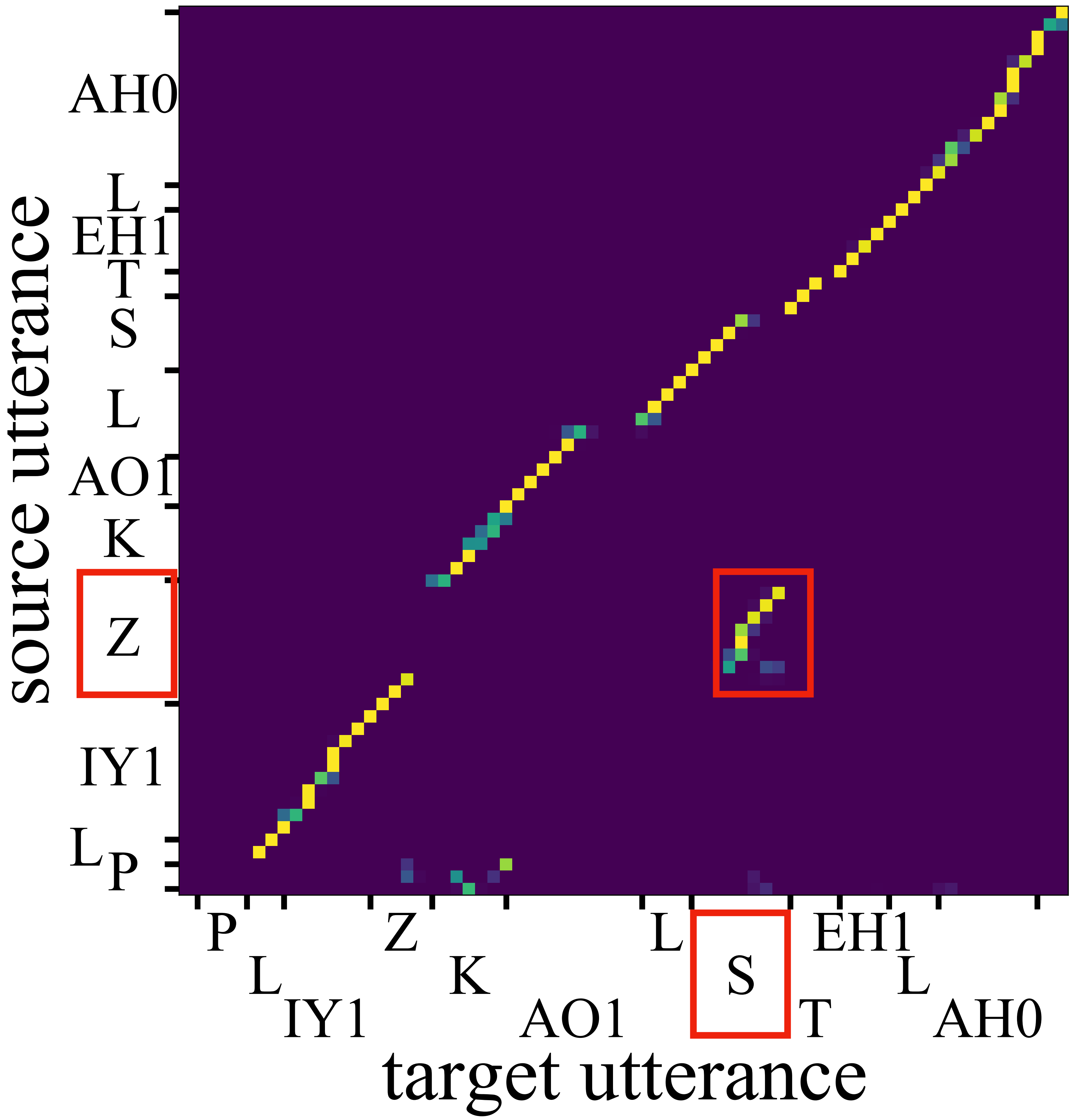
Trade similarity for naturalness

Proposed models perform better !

Attention Analysis

- Same sentence, different speakers

- Alignment of phonetically similar fragments



Source
“Please call Stella.”

Target
“Please call Stella.”

Converted
“Please call Stella.”

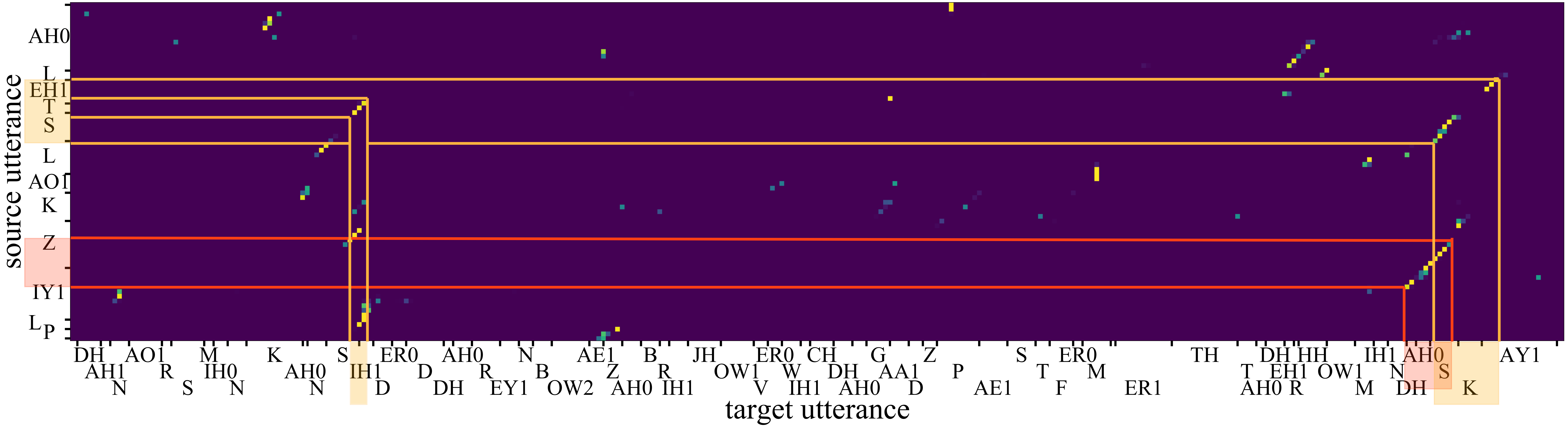
Attention Analysis

- Different sentence, different speakers

Source
“Please call Stella.”

Target
“The Norsemen considered the rainbow as a bridge over which the gods passed from earth to their home in the sky.”

Converted
“Please call Stella.”



Conclusion

- A SOTA approach to any-to-any voice conversion
- Utilize attention mechanism to end-to-end
  - **Extract** target fragments phonetically similar to source fragments
  - **Fuse** the extracted fragments to achieve voice conversion
- Source code & model: https://github.com/yistLin/FragmentVC