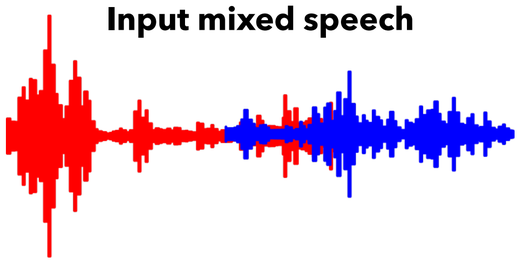
Anchored Speech Recognition With Neural Transducers

Desh Raj*, Junteng Jia, Jay Mahadeokar, Chunyang Wu, Niko Moritz, Xiaohui Zhang, Ozlem Kalinli

Background speech suppression

The problem



Speech Recognition

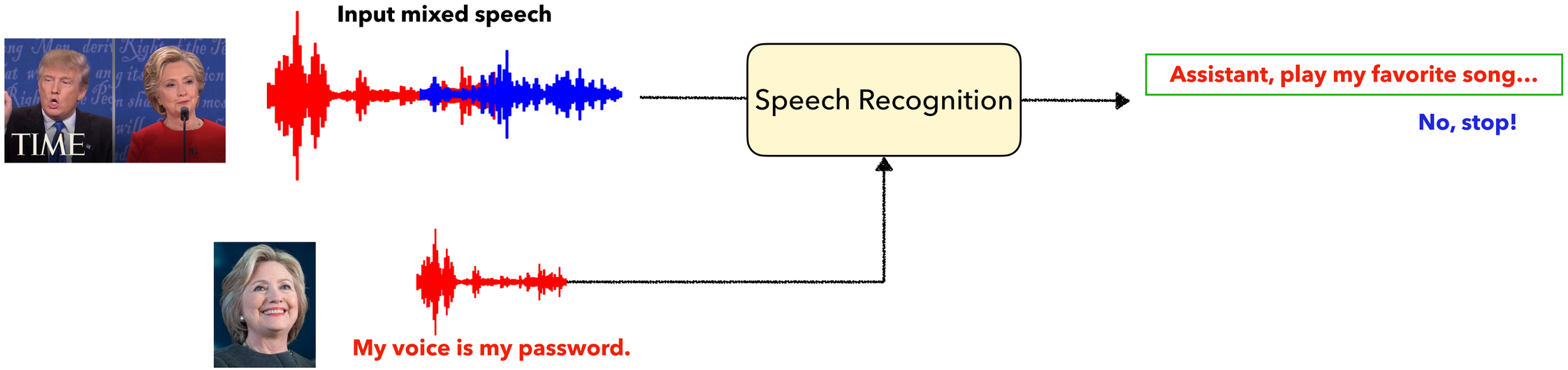
Device-directed speech?

Assistant, play my favorite song…

No, stop!

Background speech suppression

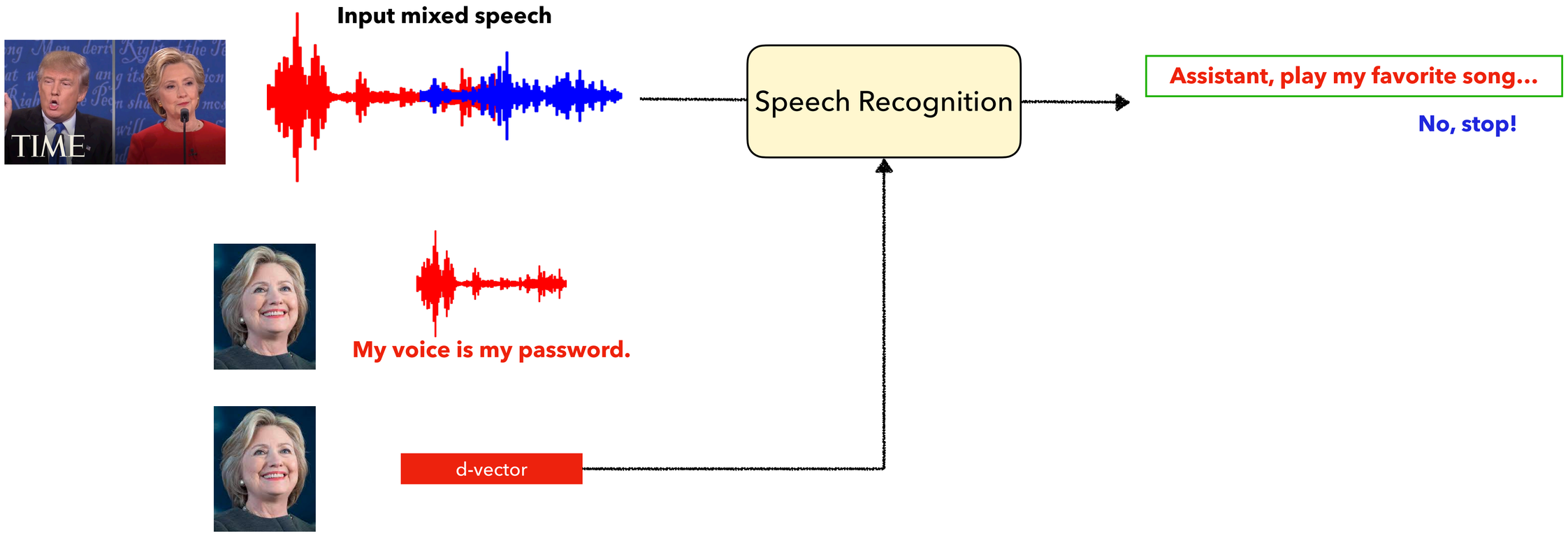
Enrollment utterance



K. Žmolíková *et al.*, "**SpeakerBeam**: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4.

Background speech suppression

Speaker embedding

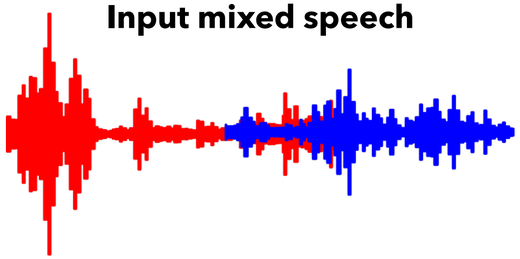


Wang, Quan et al. "**VoiceFilter-Lite**: Streaming Targeted Voice Separation for On-Device Speech Recognition." *InterSpeech* (2020).

Background speech suppression

Both of these require an *enrollment* stage

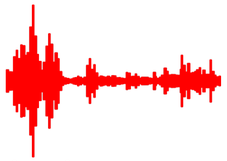
Input mixed speech



Speech Recognition

Assistant, play my favorite song…

No, stop!



My voice is my password.

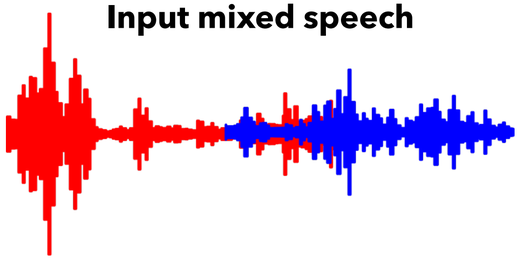
Speaker needs to be “enrolled” with the device

d-vector

Anchored Speech Recognition

Wake-word as the anchor

Input mixed speech



Speech Recognition

Assistant, play my favorite song…

No, stop!

"Assistant"

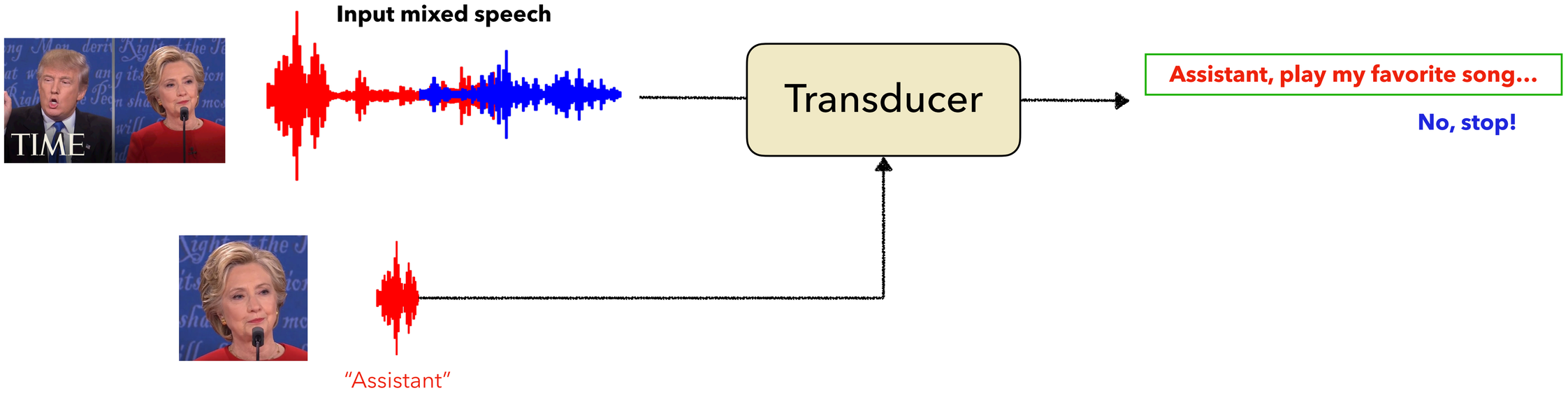
✓ On-the-fly, i.e., no "enrollment"

✓ Matched noise conditions

Wang, Yiming et al. "End-to-end Anchored Speech Recognition." IEEE ICASSP, 2019.

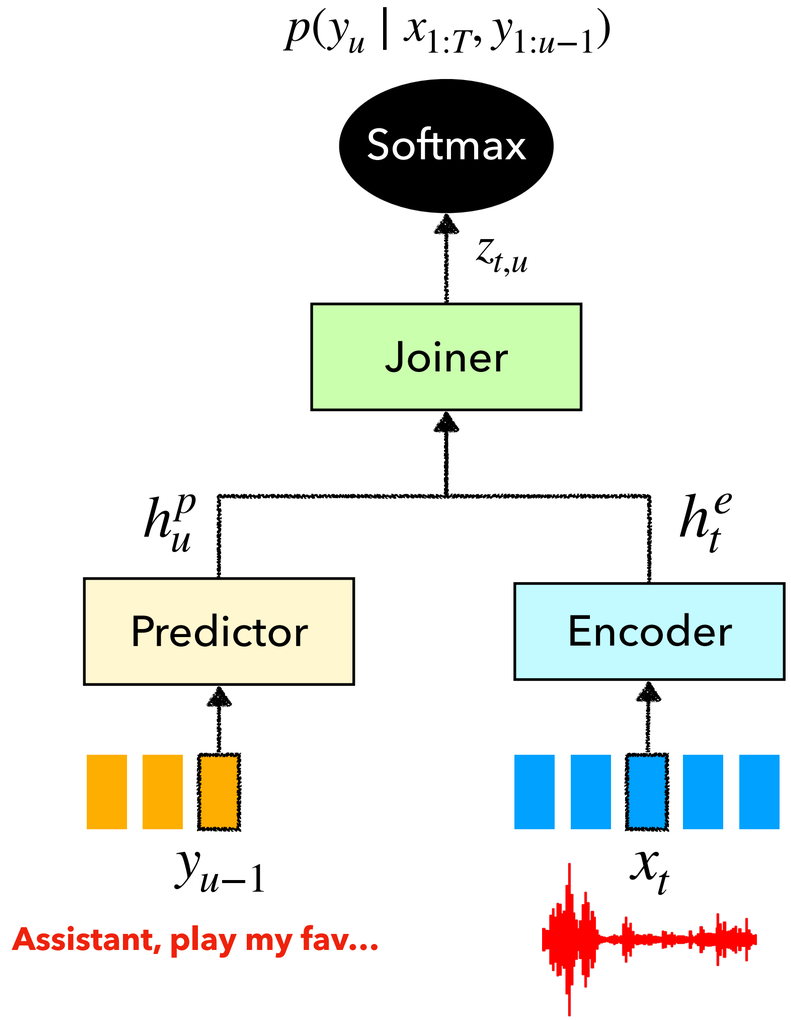
Our work

Anchored Speech Recognition + Transducer-based ASR



What is a transducer?

Conditional dependence + streaming

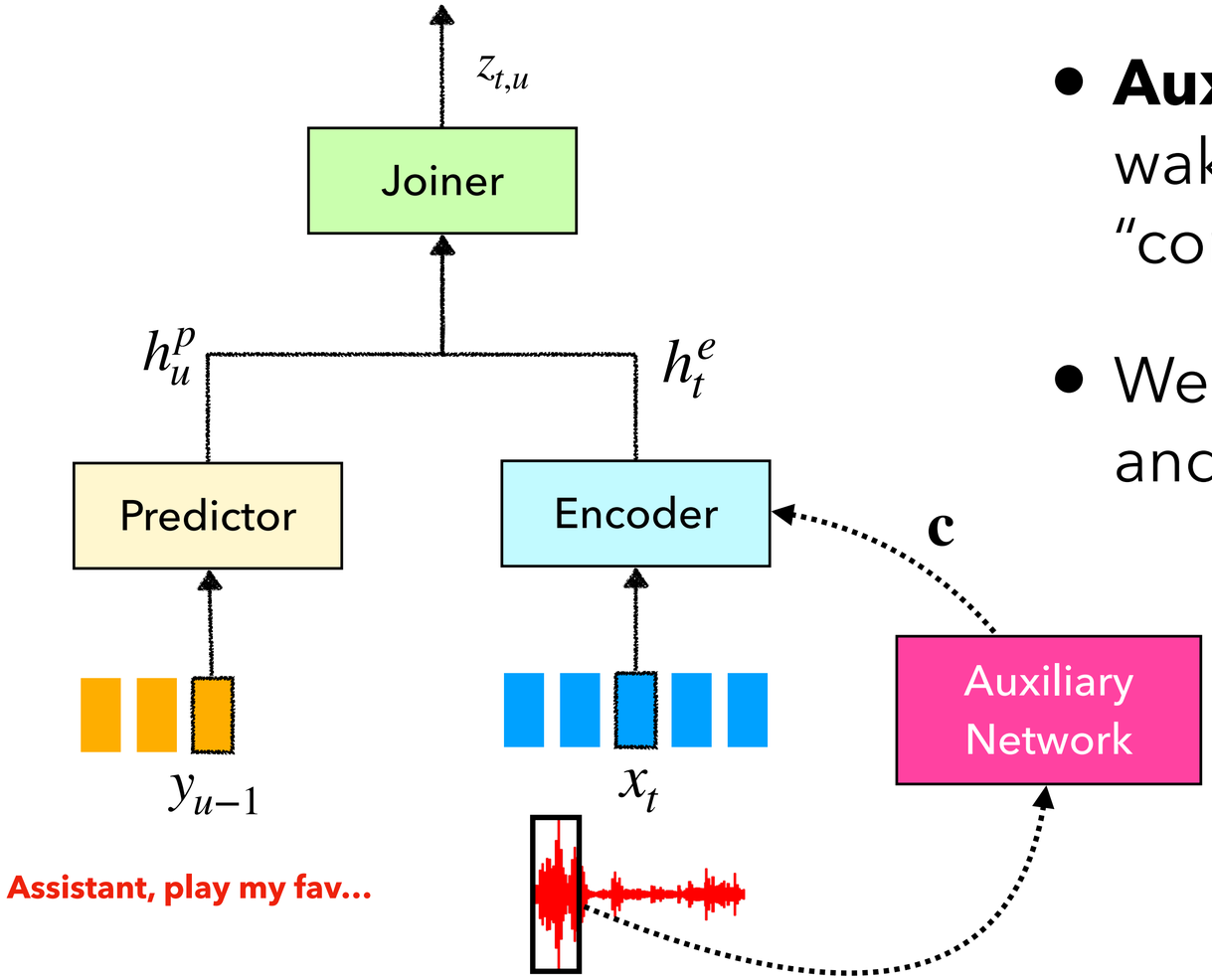


- **Encoder** converts input *audio* to high-dimensional representation
- **Predictor** is an autoregressive model that encodes input *text*
- **Joiner** combines audio and text representations to predict next token

Anchoring the transducer

1. Bias the encoder with context

Encoder can use context embedding to suppress background speech.

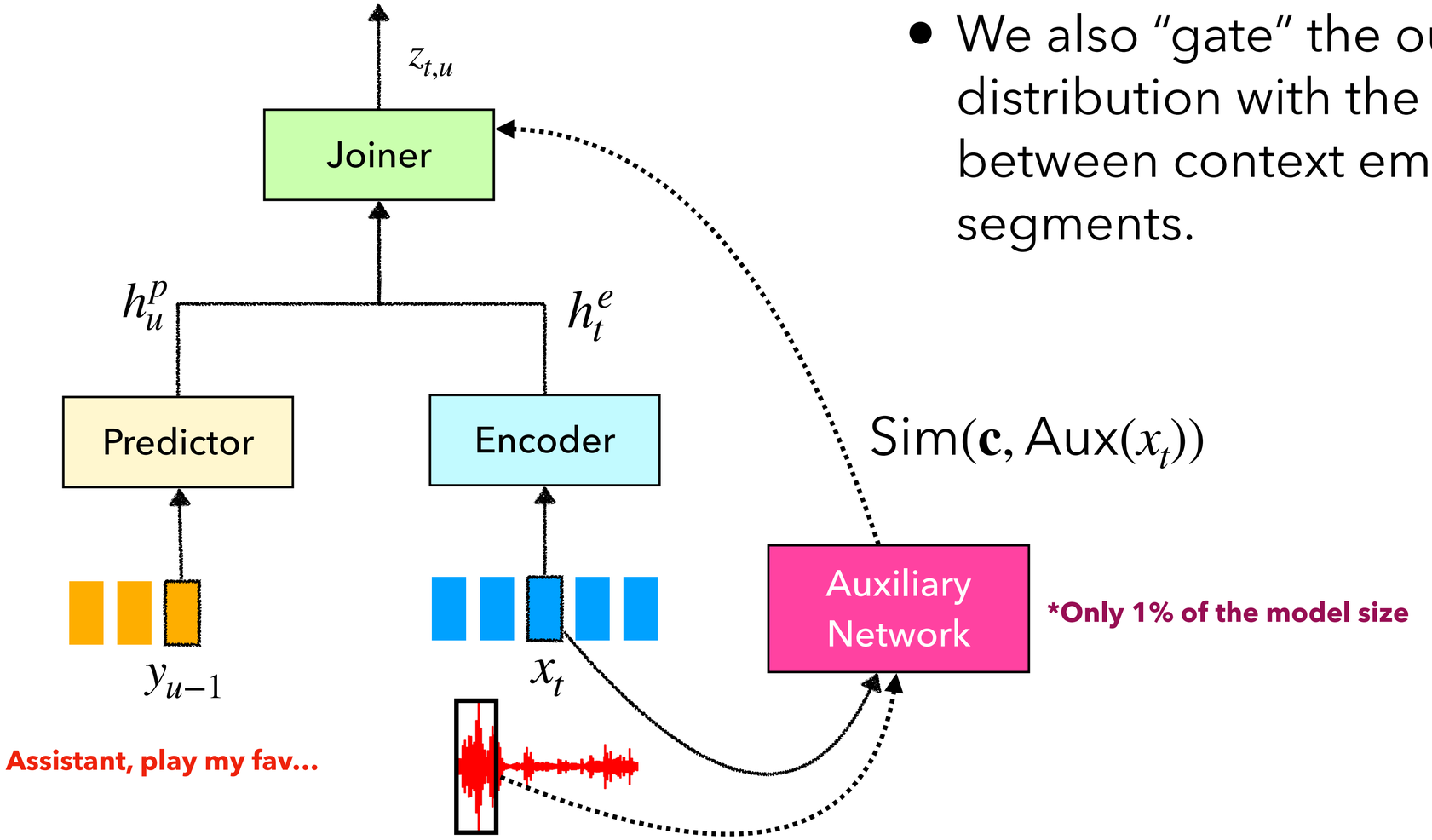


- **Auxiliary network** encodes the wake-word segment into a "context" embedding
- We concat this to input features and project to original dimension

Anchoring the transducer

2. Gating the joiner

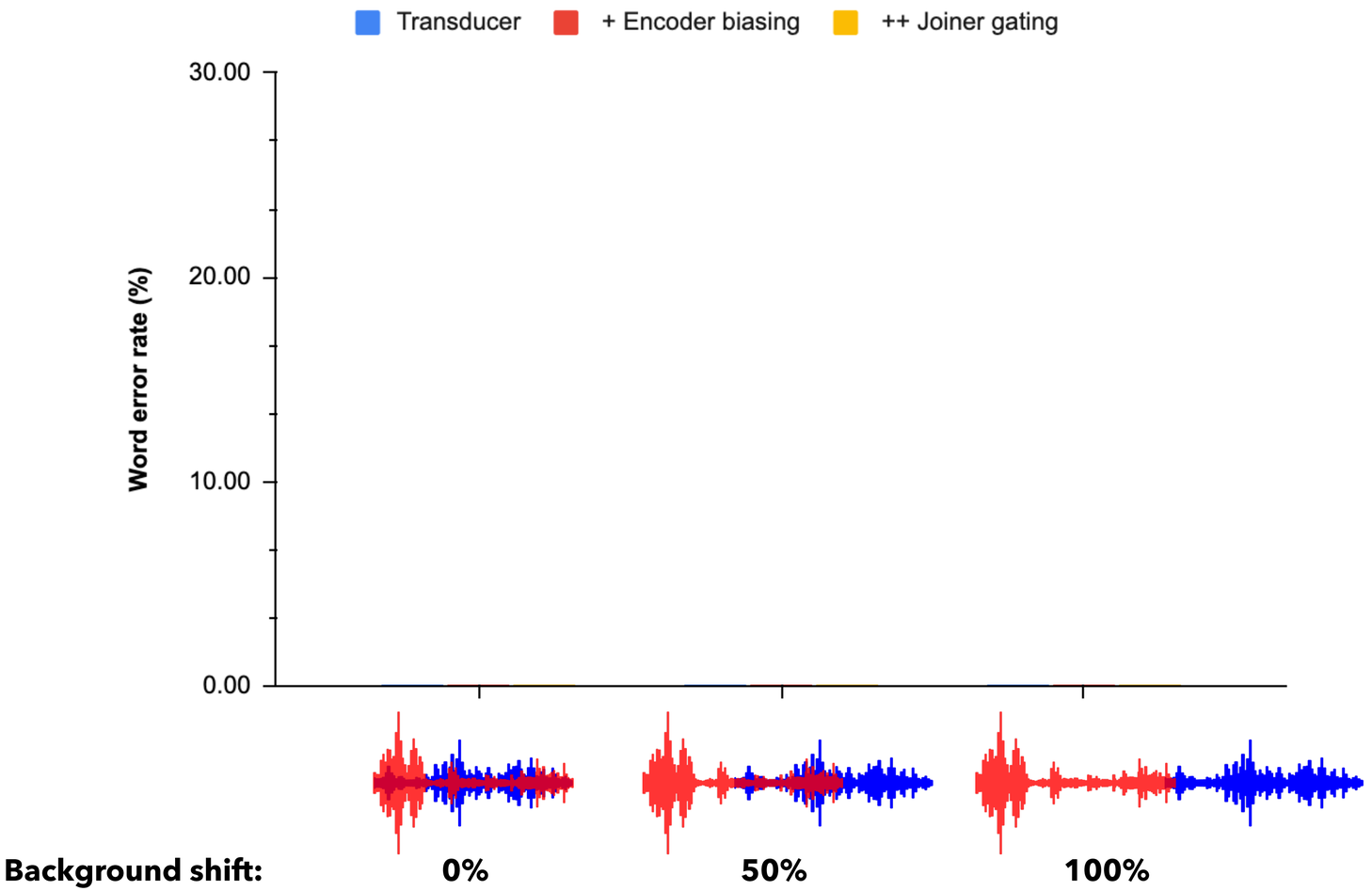
Boost the logits for blank tokens when speaker is different from wake-word segment.

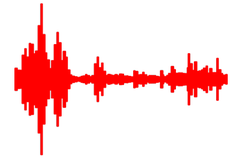


- We also "gate" the output distribution with the similarity score between context embedding and segments.

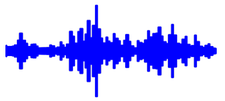
Effect of TS-ASR

WER on LibriSpeech mixtures (average over SNRs 1~20 dB)





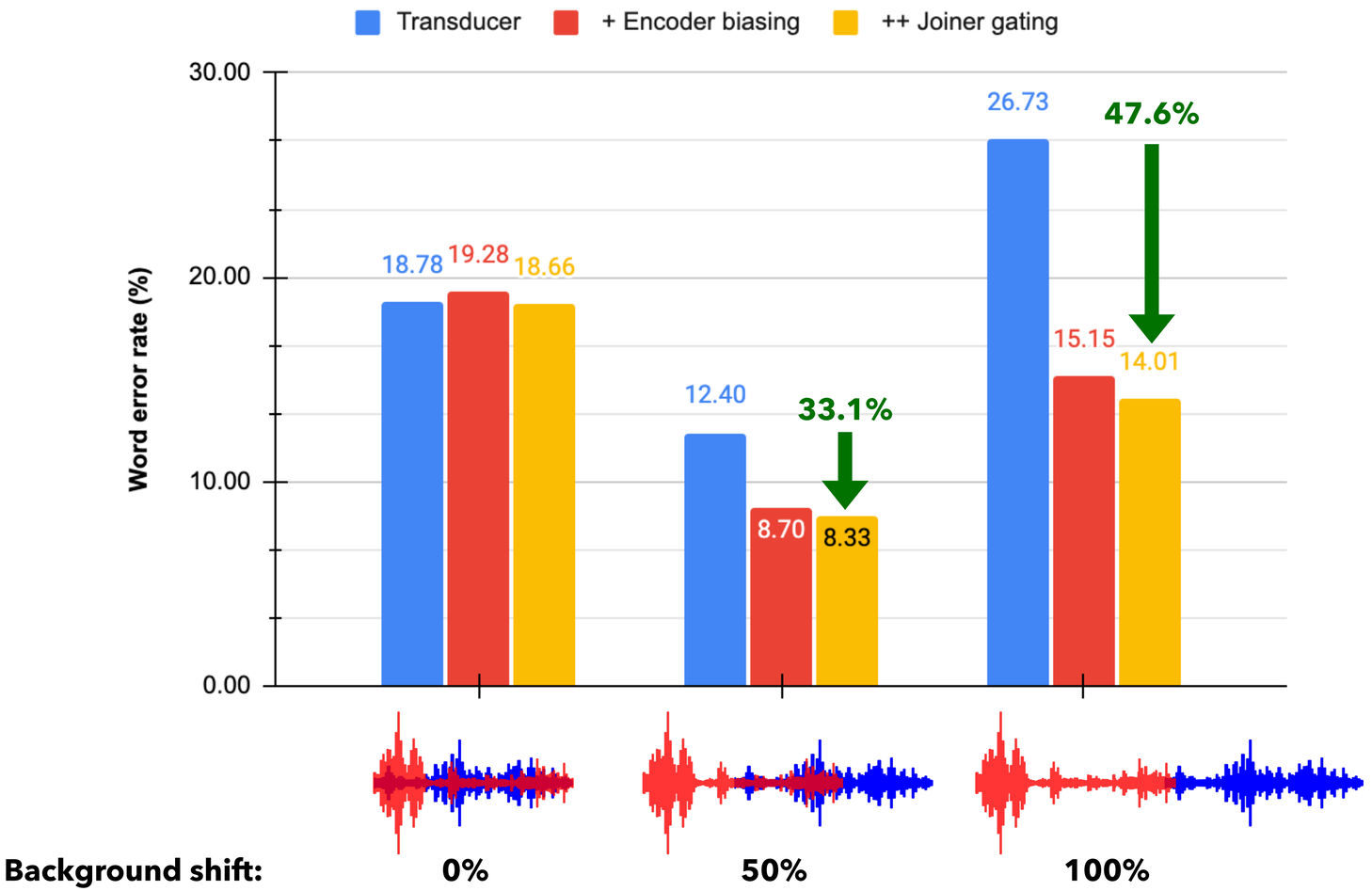
Primary speaker



Background speaker

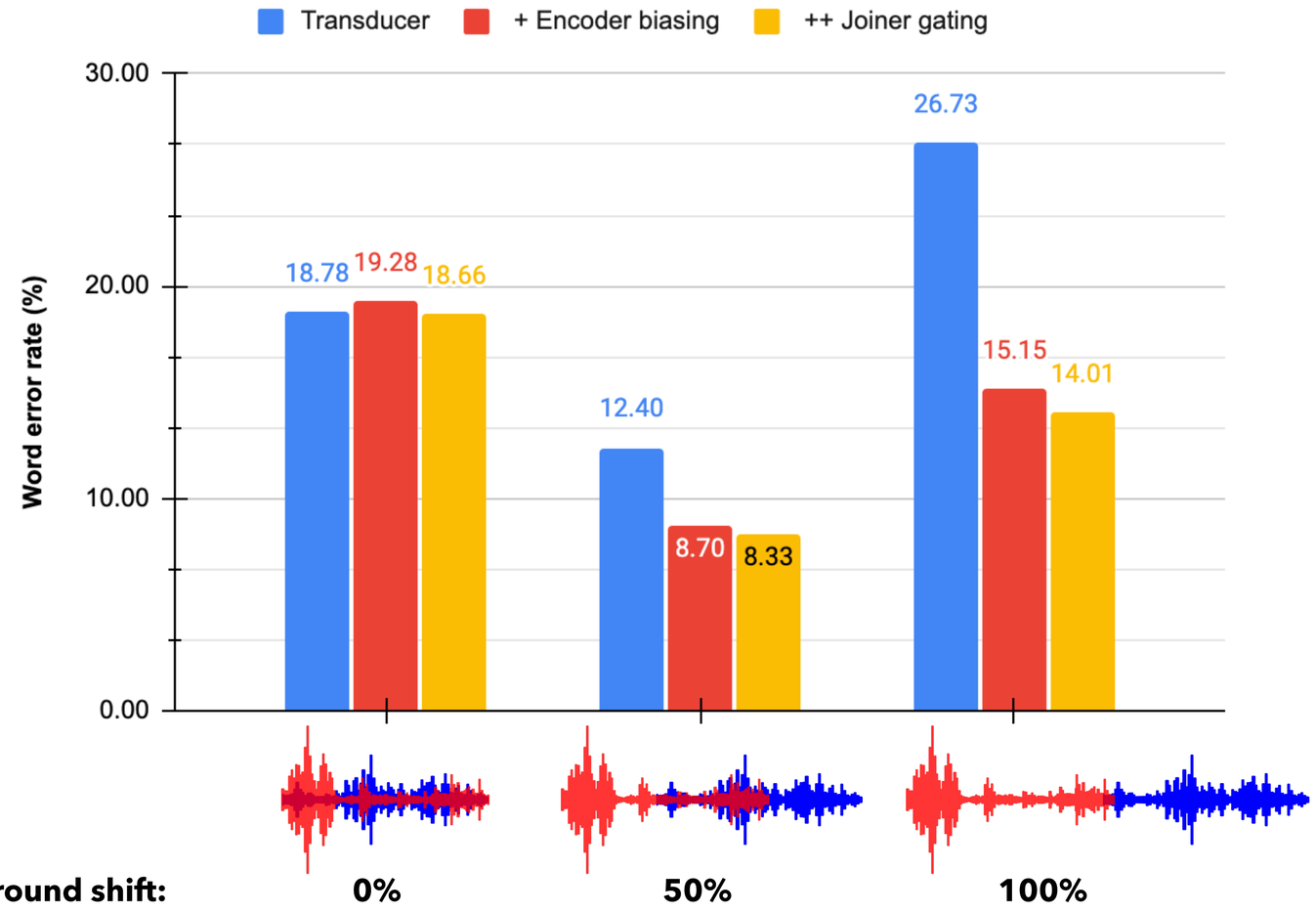
Effect of TS-ASR

WER on LibriSpeech mixtures (average over SNRs 1~20 dB)



Effect of TS-ASR

WER on LibriSpeech mixtures (average over SNRs 1~20 dB)



REFERENCE:

Then they seemed to spring from every part of the country

TRANSDUCER:

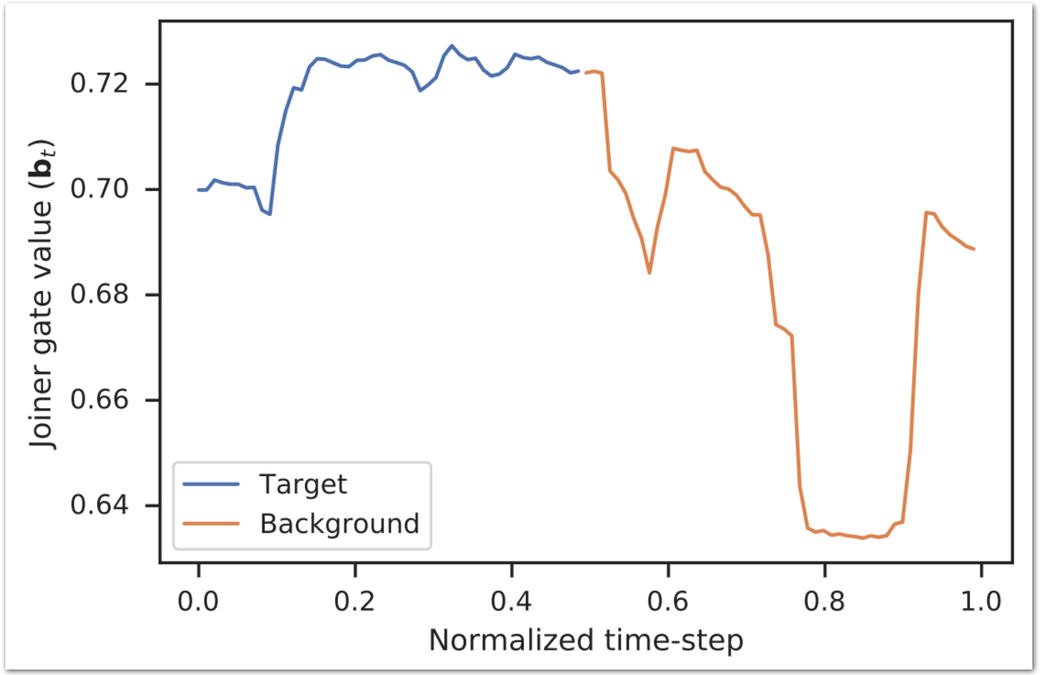
Then they seemed to spring from every part of the country [hastened may be very much modified to dogmas]

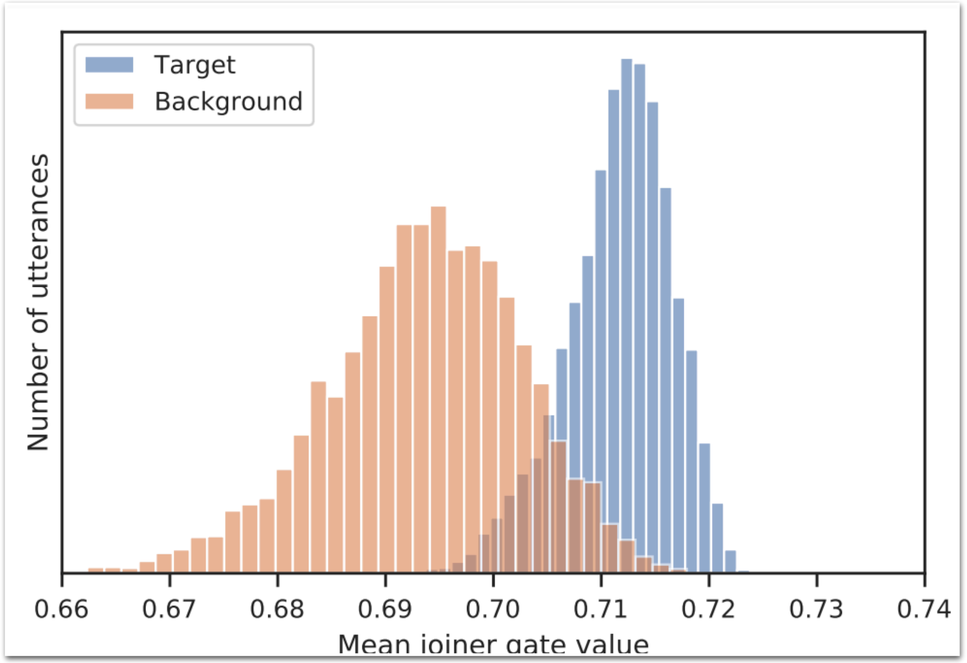
TS—ASR:

Then they seemed to spring from every part of the country

Analysis of joiner gating

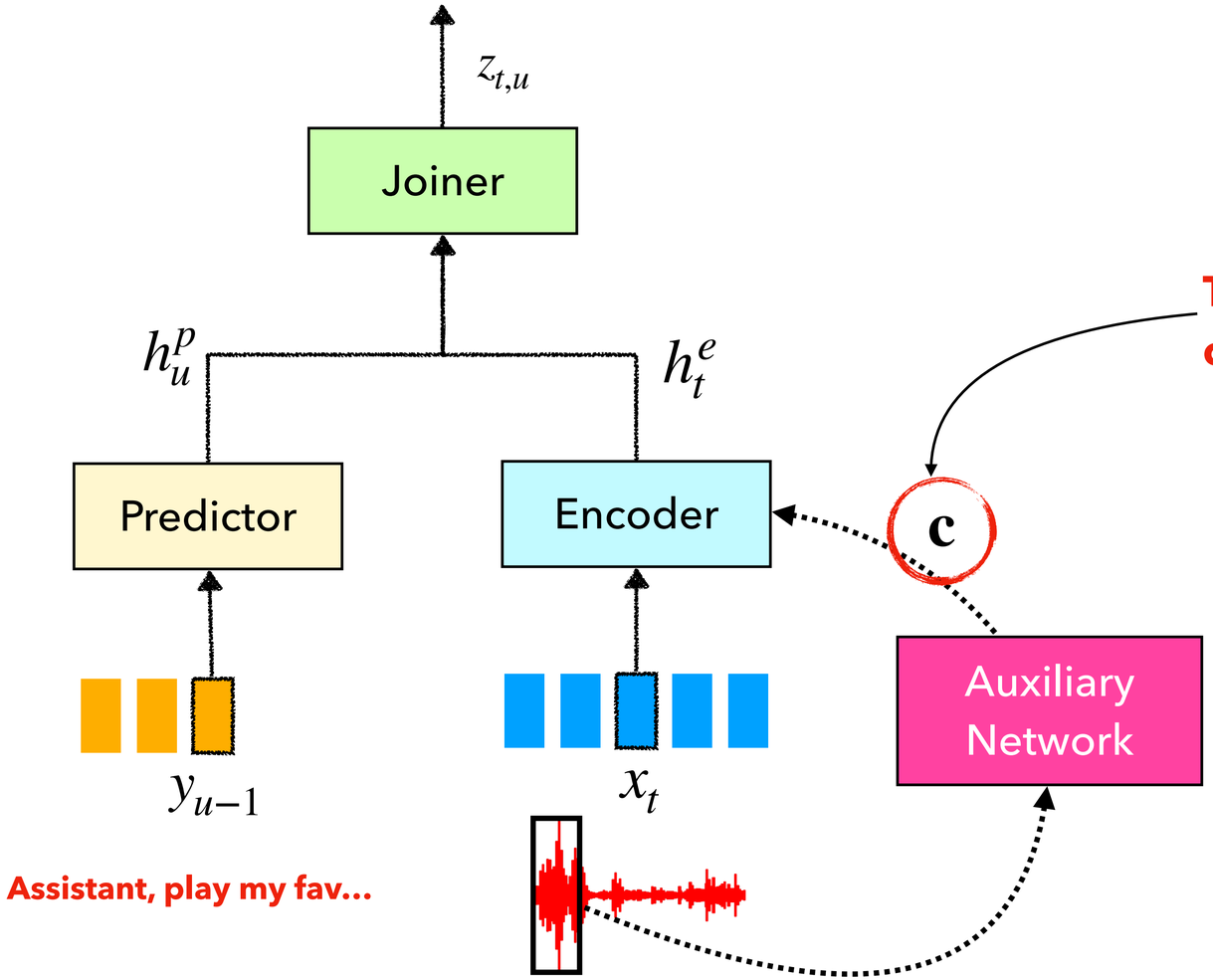
Similarity scores for target and background segments





Let’s think about the context embedding

We want to disentangle “style” from “content”

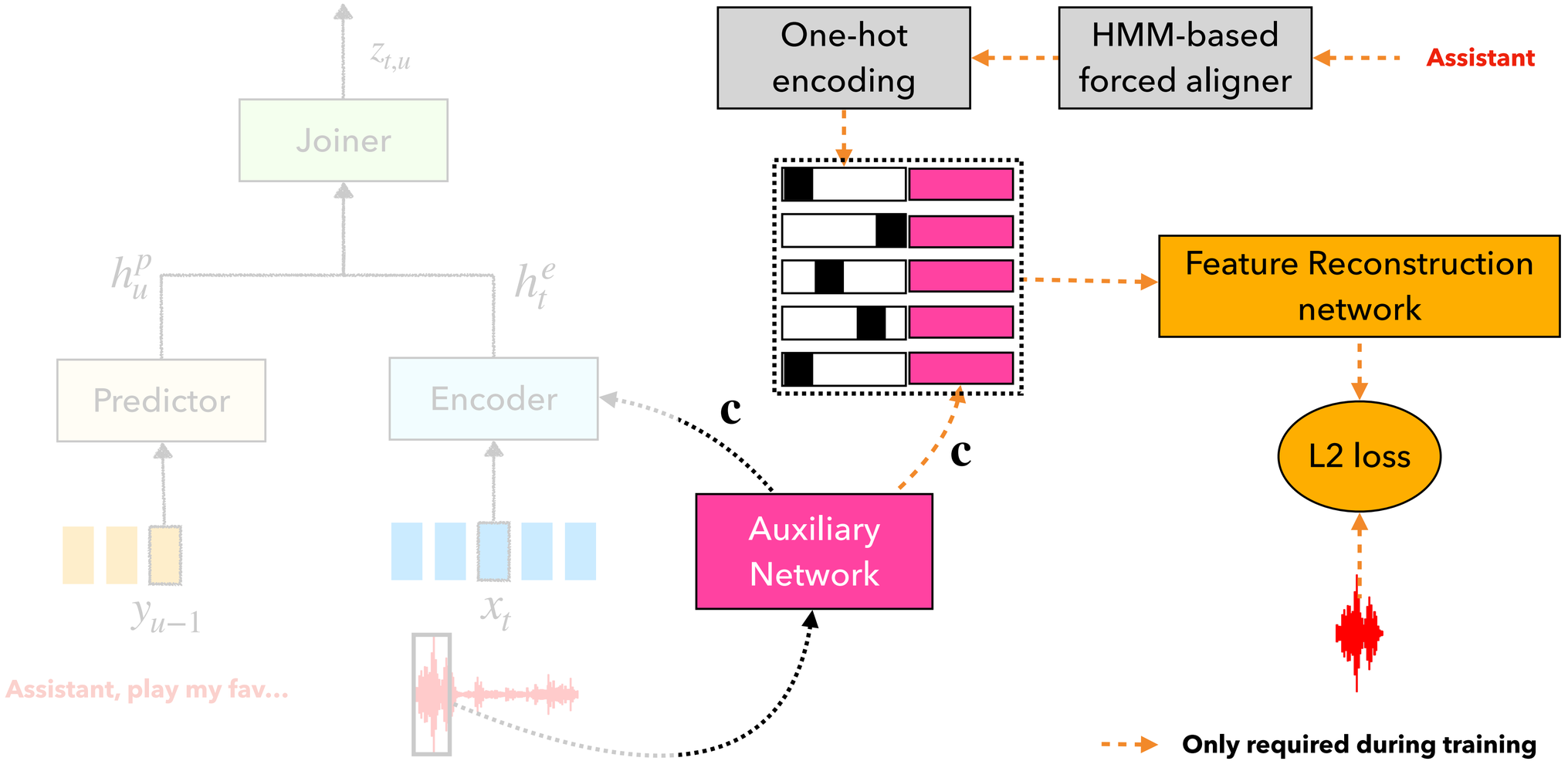


The context embedding should only contain target-speaker information

Auxiliary Network

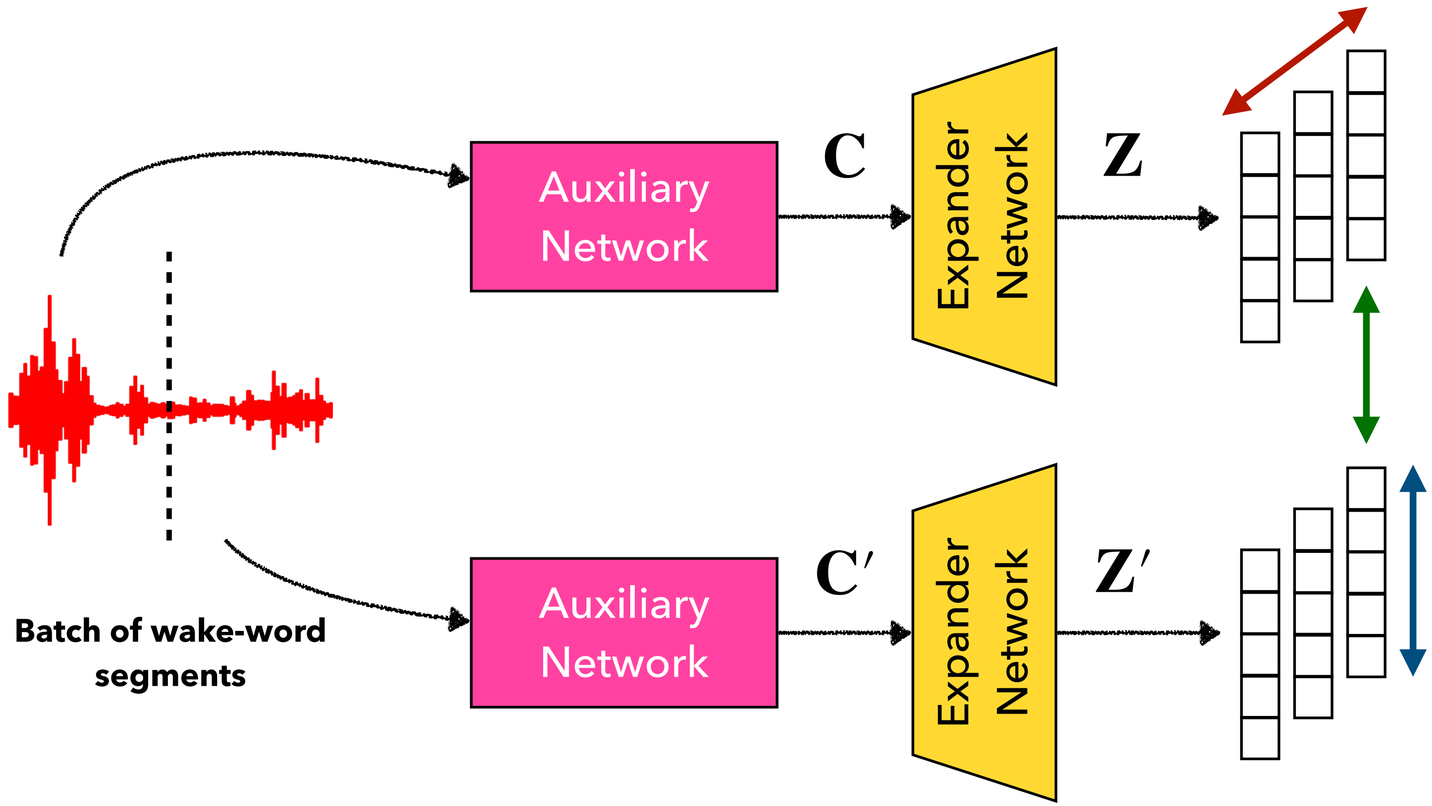
Disentangling “style” from “content”

Method 1: Feature Reconstruction



Disentangling “style” from “content”

Method 2: VIC Regularization



Variance prevents representations from collapsing to a constant

$$v(\mathbf{Z}) = \frac{1}{\mathfrak{D}} \sum_{j=1}^{\mathfrak{D}} \max\left(0, 1 - \sqrt{\mathrm{Var}(\mathbf{z}^j)}\right)$$

Invariance ensures that representations are similar for the two halves

$$s(\mathbf{Z}, \mathbf{Z}') = \mathrm{MSE}(\mathbf{Z}, \mathbf{Z}')$$

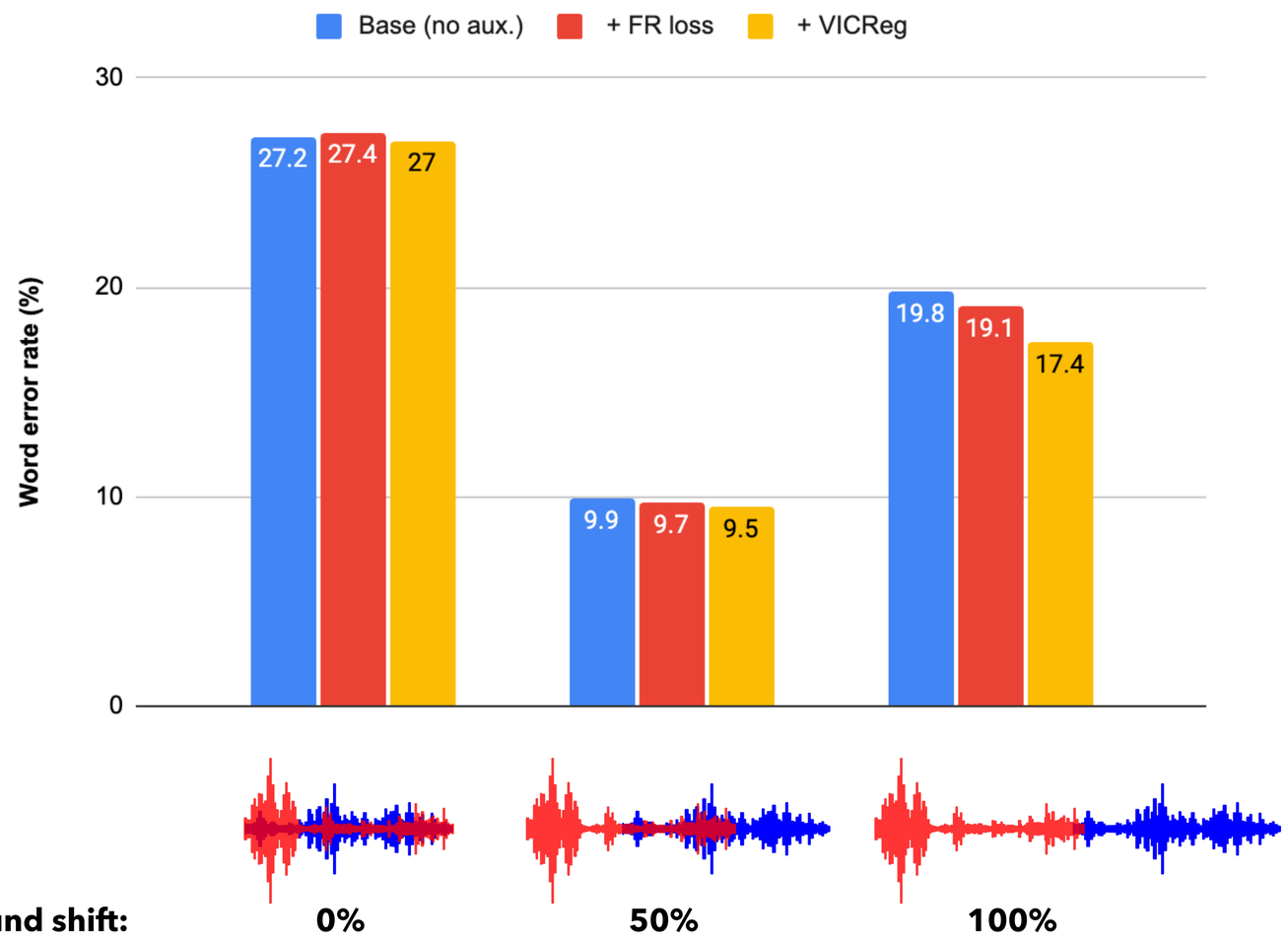
Covariance reduces correlation among dimensions to increase information content

$$c(\mathbf{Z}) = \frac{1}{\mathfrak{D}} \sum_{i \neq j} [\mathrm{cov}(\mathbf{Z})]_{i,j}^2$$

Bardes, Adrien et al. “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning.” *ICLR, 2022.*

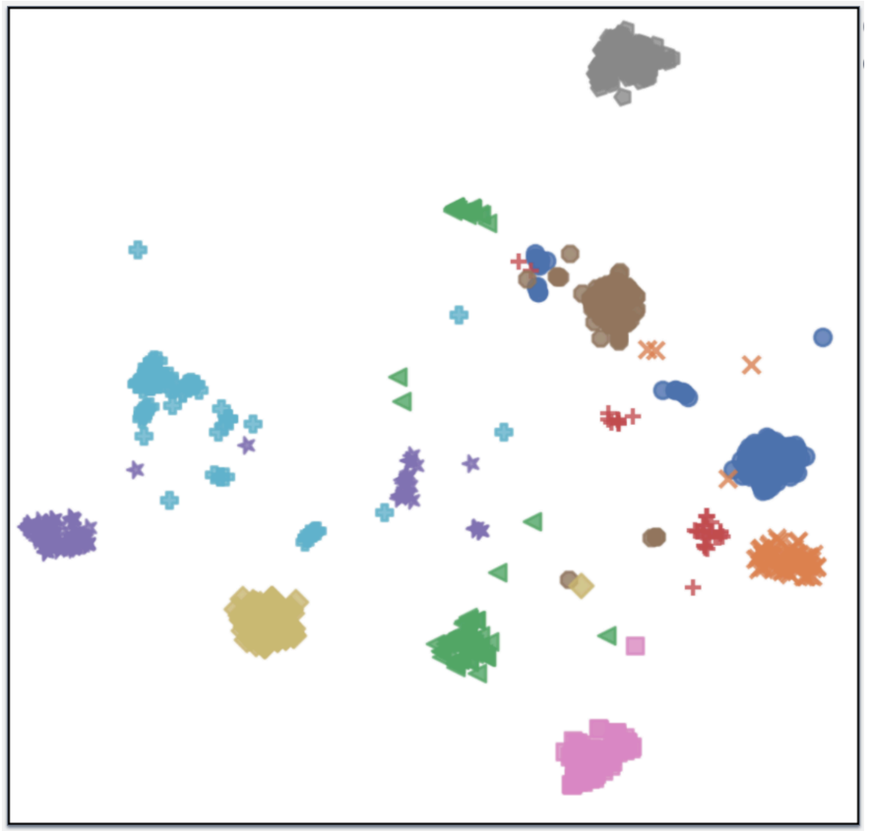
Effect of auxiliary objectives

WER on LibriSpeech mixtures (average over SNRs 1~10 dB)



Context embeddings capture speaker characteristics

T-SNE clustering of context embeddings



Summary

- We can use **wake-words** to bias neural transducers to transcribe only the device-directed speech, reducing WER by 19.6%.
- **Encoder biasing** and **joiner gating** are complementary in suppressing background speech.
- **Auxiliary objectives** help to retain speaker information in the context embedding and remove content information.

For more information…

- Join us at our **oral** presentation:
  - Date/time: June 7, 2023, 10:50 AM (EEST)
  - Session name: Multi-speaker ASR
- Or contact me at:
  - E-mail: r.desh26@gmail.com
  - Twitter: @rdesh26