國立臺灣大學
National Taiwan University

INTERSPEECH 2024

Singing Voice Graph Modeling for SingFake Detection






Xuanjun Chen Haibin Wu Jyh-Shing Roger Jang Hung-yi Lee

Interspeech 2024 Oral Presentation

2024.09.05

Major Contributions

Propose SingGraph model for SingFake detection task

Analyze how to encode vocal and instrument better

Propose a beat-matching augmentation method

SingGraph model achieves state-of-the-art performance

Outline

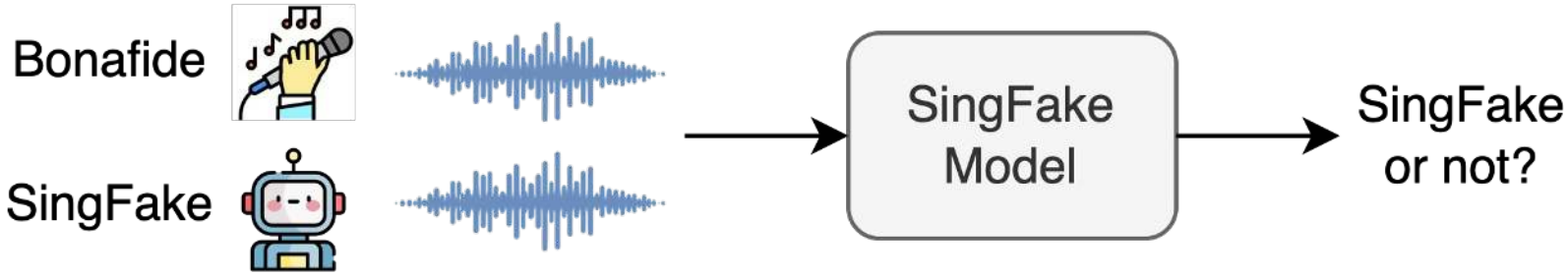
- Introduction
  - What is SingFake detection?
  - Speech anti-spoofing model
  - The difference between speech and singing voice
- SingGraph Model
  - Source separation
  - Vocal and instrumental Encoder
  - AASIST backend
  - Training framework
  - Beat-matching augmentation

Outline

- Experiment Setting
- Research Question
  - Are instrumental music cues helpful?
  - Does it require separate encoders for encoding?
  - Does it require input from vocal and instrumental sources?
  - Is beat-matching augmentation effective?
- Comparison with the state-of-the-art (SoTA) model
- Conclusion

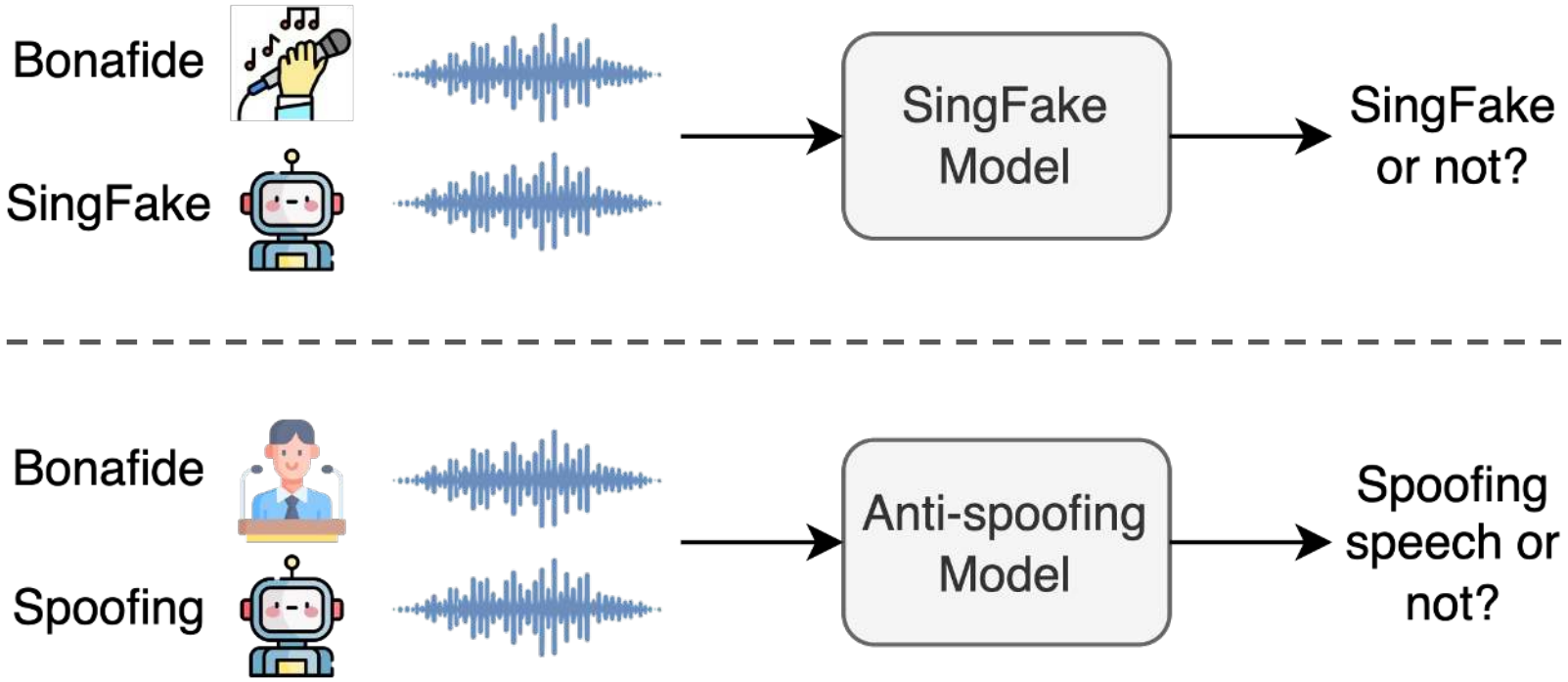
What is SingFake detection?

- Singing voice deepfake (SingFake) detection aims to determine whether the singing voice is SingFake or not



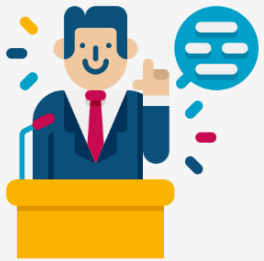
Speech anti-spoofing model

- Speech anti-spoofing aims to determine whether the speech is spoofing sample or not



The difference between speech and singing voice

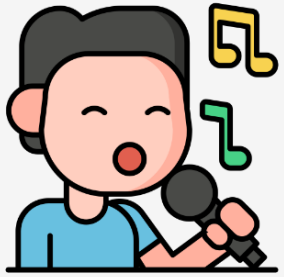
Speech



speech

VS

Singing



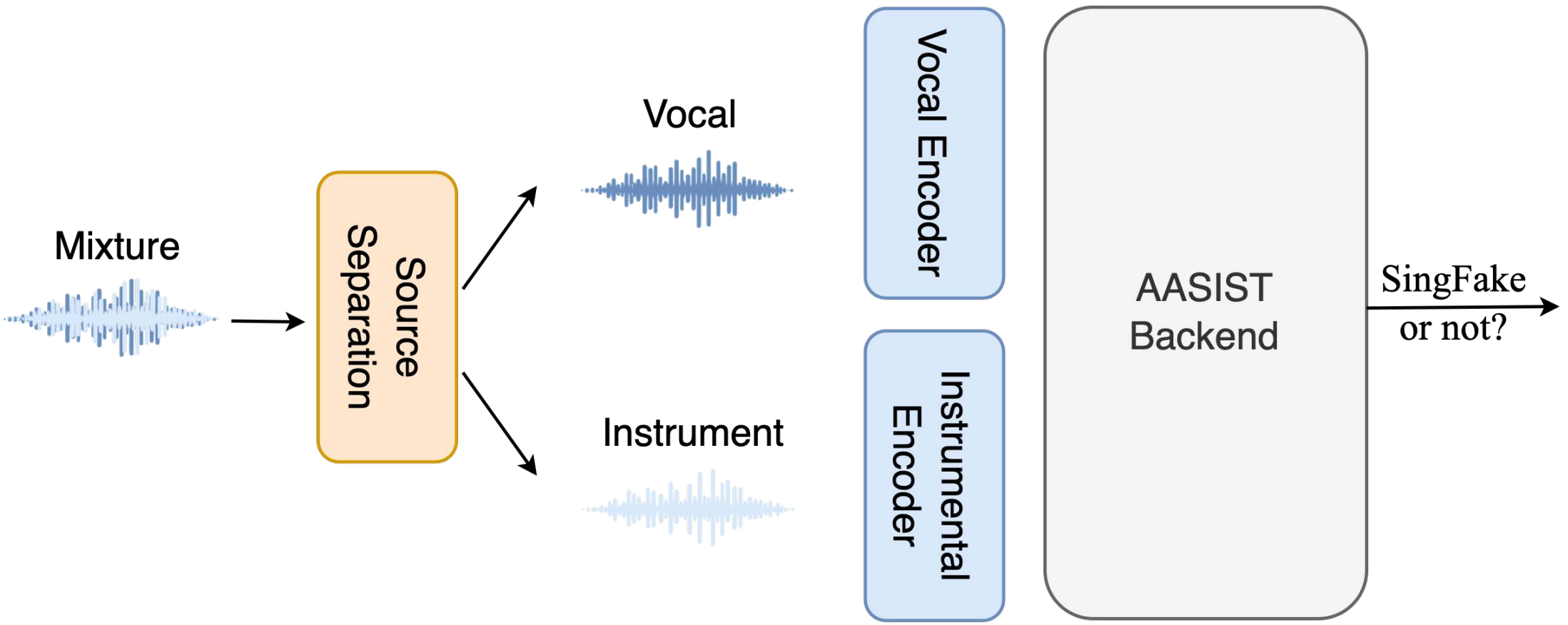
vocal



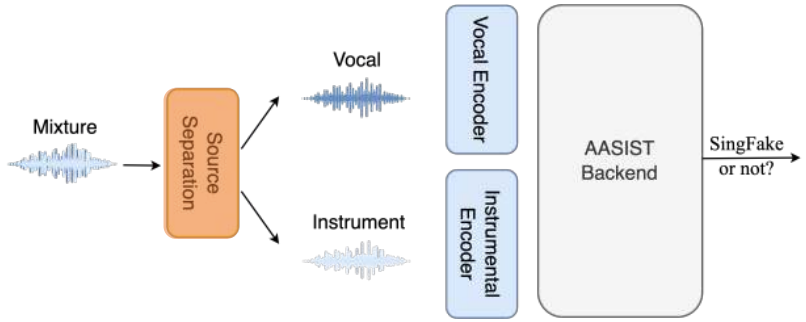
instrument

SingGraph Model

SingGraph Model



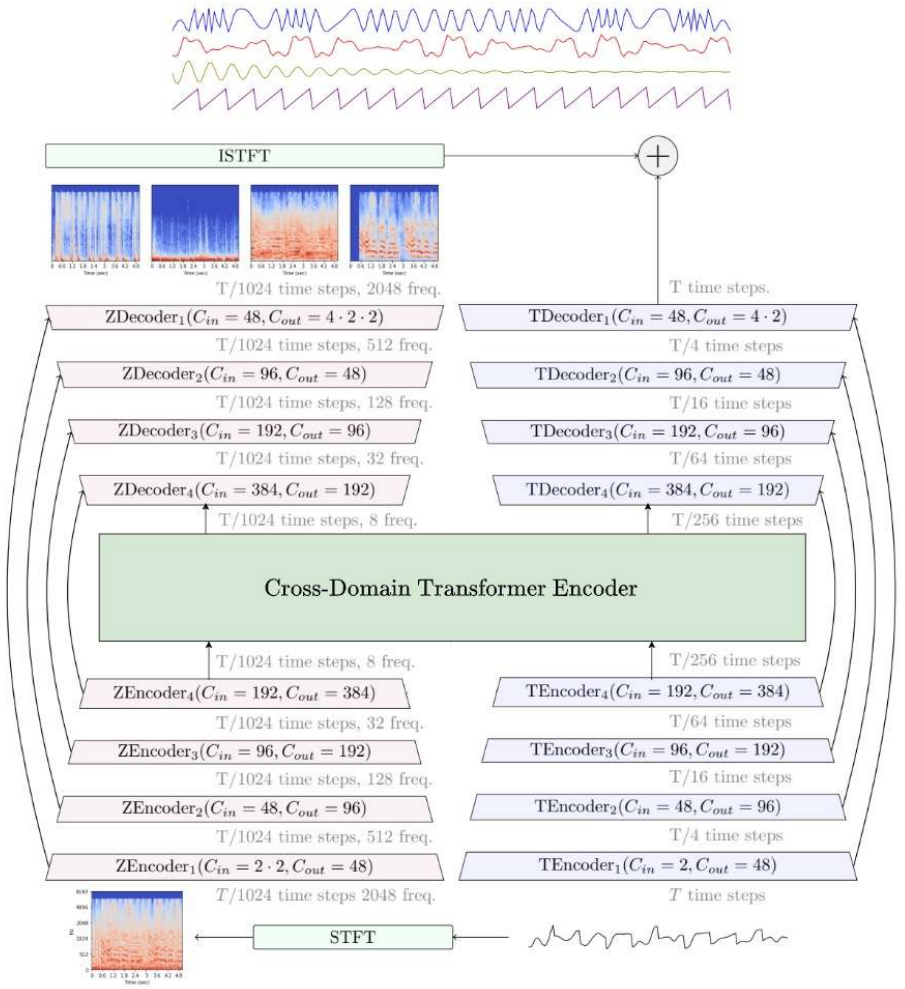
SingGraph: Source separation



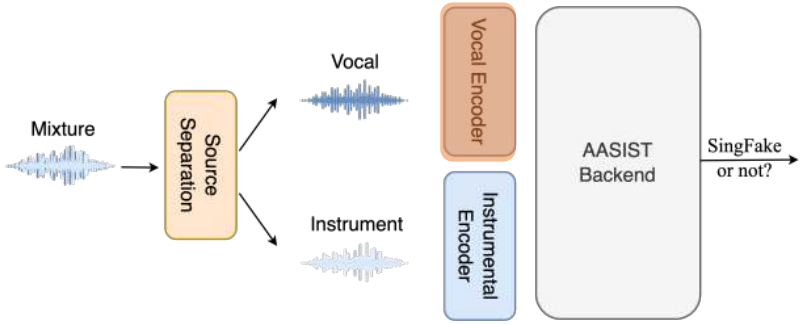
Demucs:

Demucs Music Source Separation

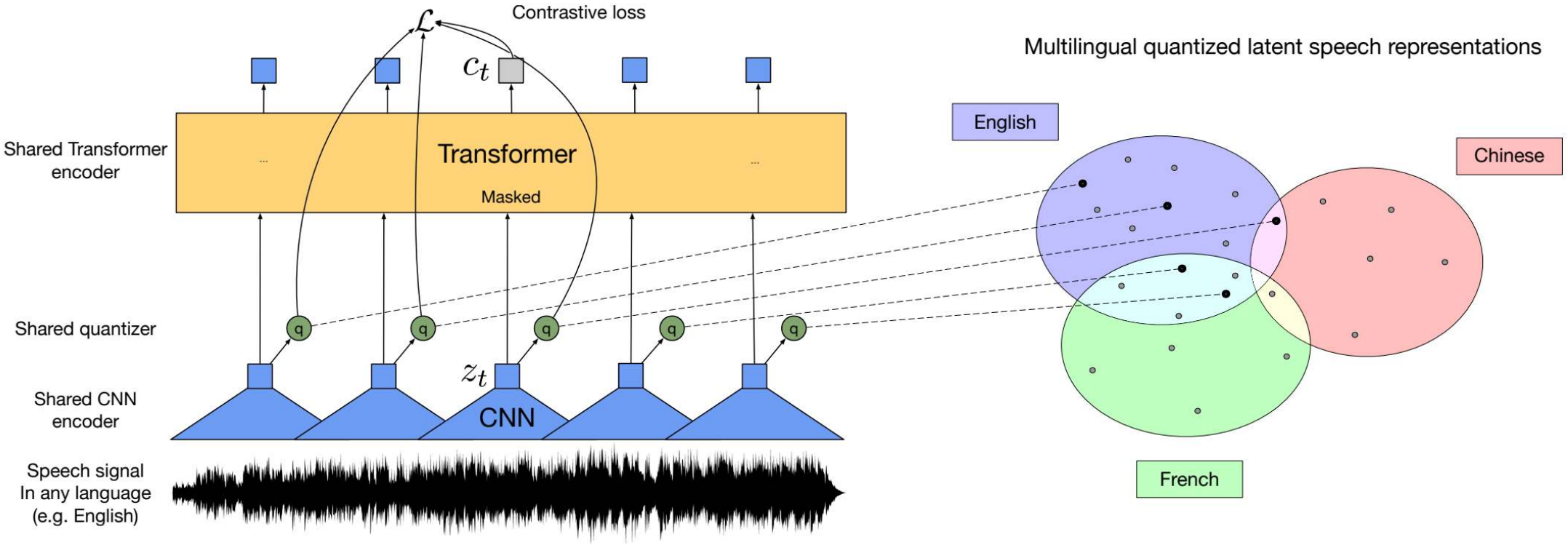
[Rouard, Simon et al., 2023]



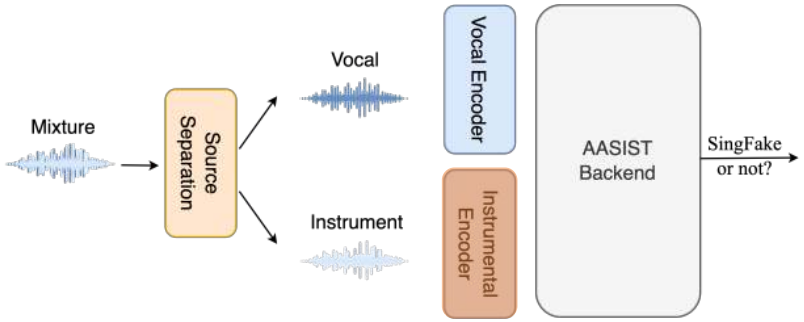
SingGraph: Vocal Encoder



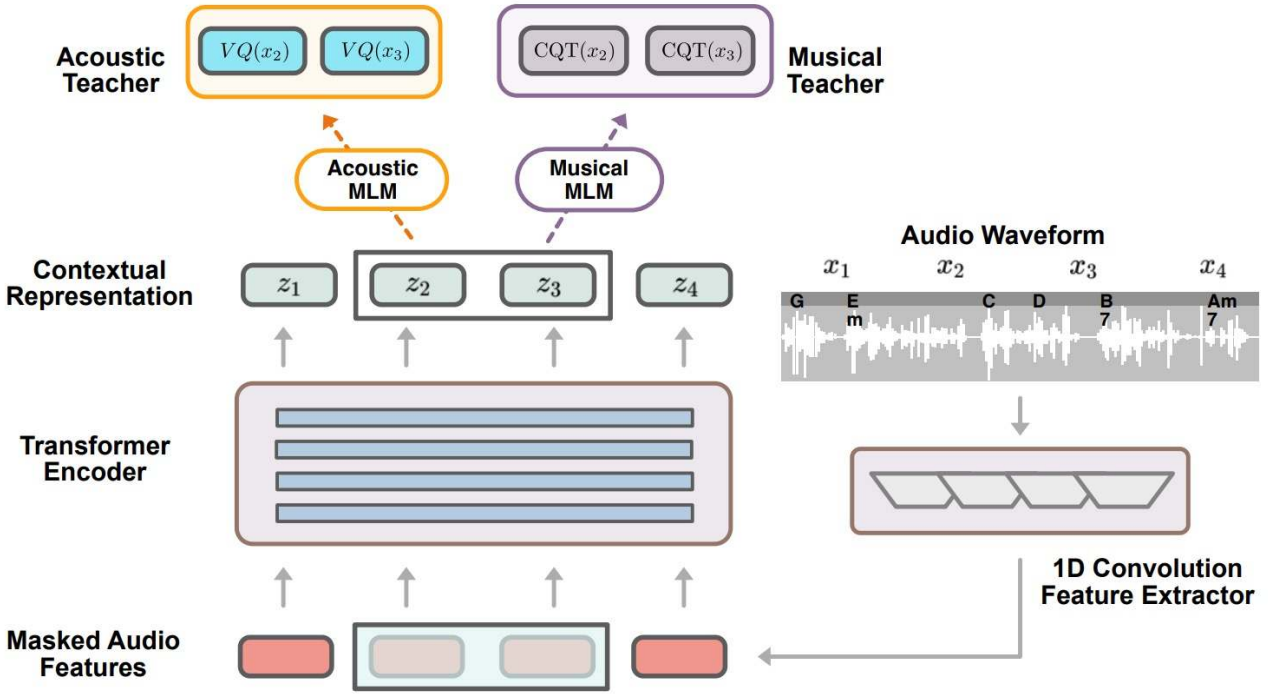
XLSR-Wav2Vec2 (W2V2): Unsupervised Cross-lingual Representation Learning for Speech Recognition [Alexis Conneau et al., 2020]



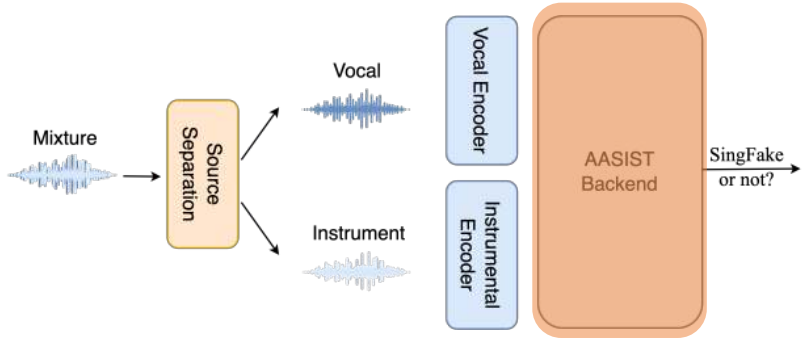
SingGraph: Instrumental Encoder



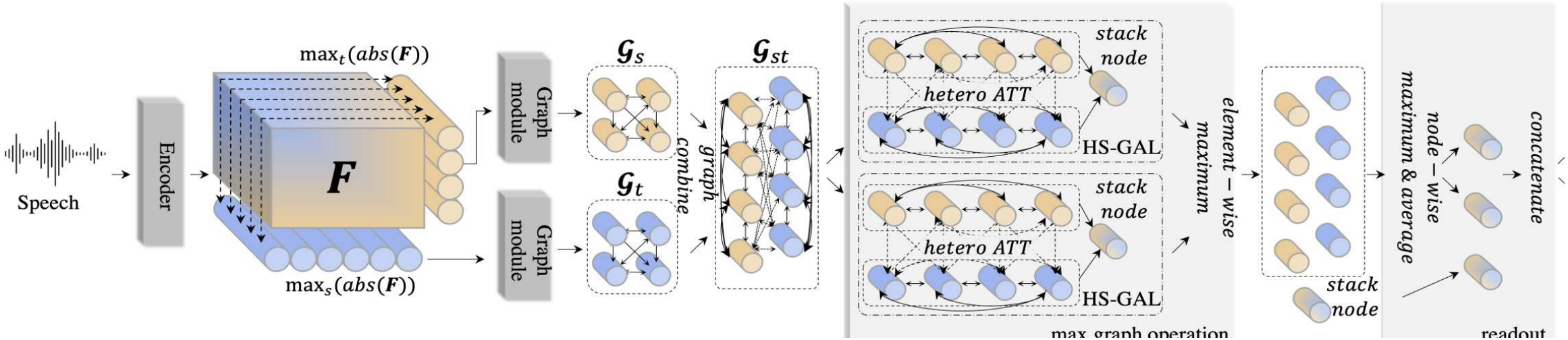
MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training [Li et al., 2024]



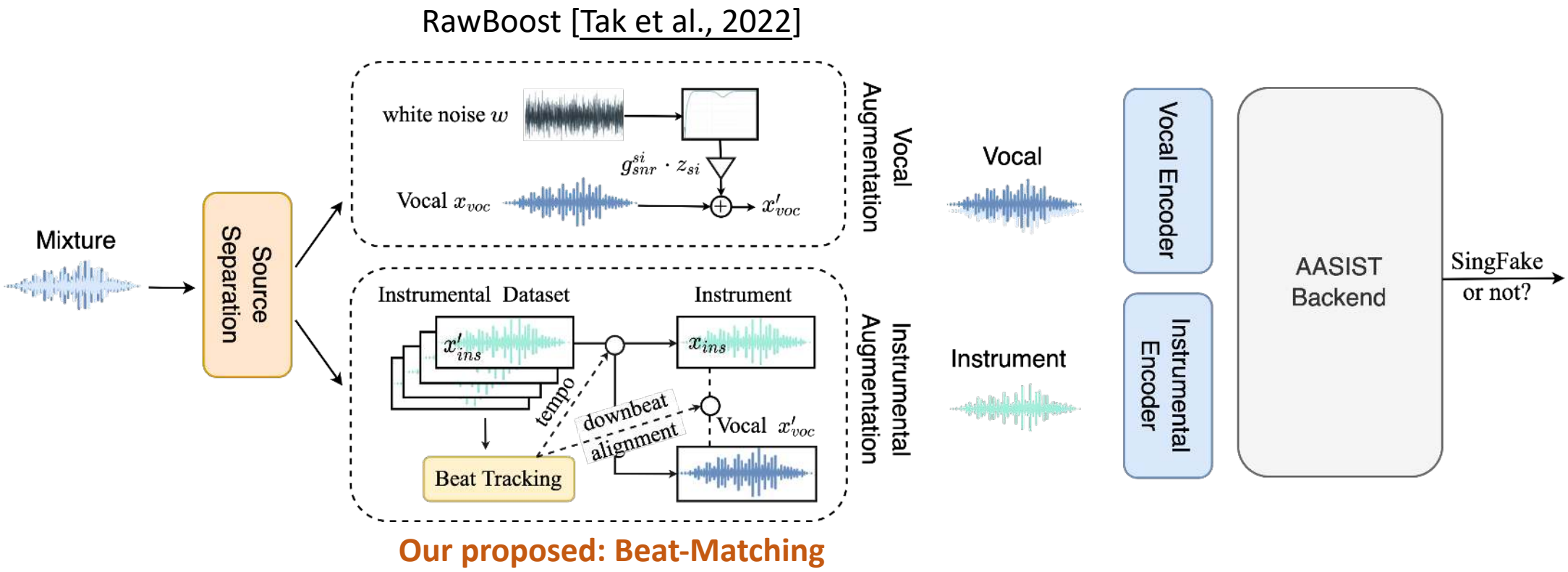
SingGraph: AASIST backend



AASIST: AUDIO ANTI-SPOOFING USING INTEGRATED SPECTRO-TEMPORAL GRAPH ATTENTION NETWORKS [Jung et al., 2021]



SingGraph: Training framework

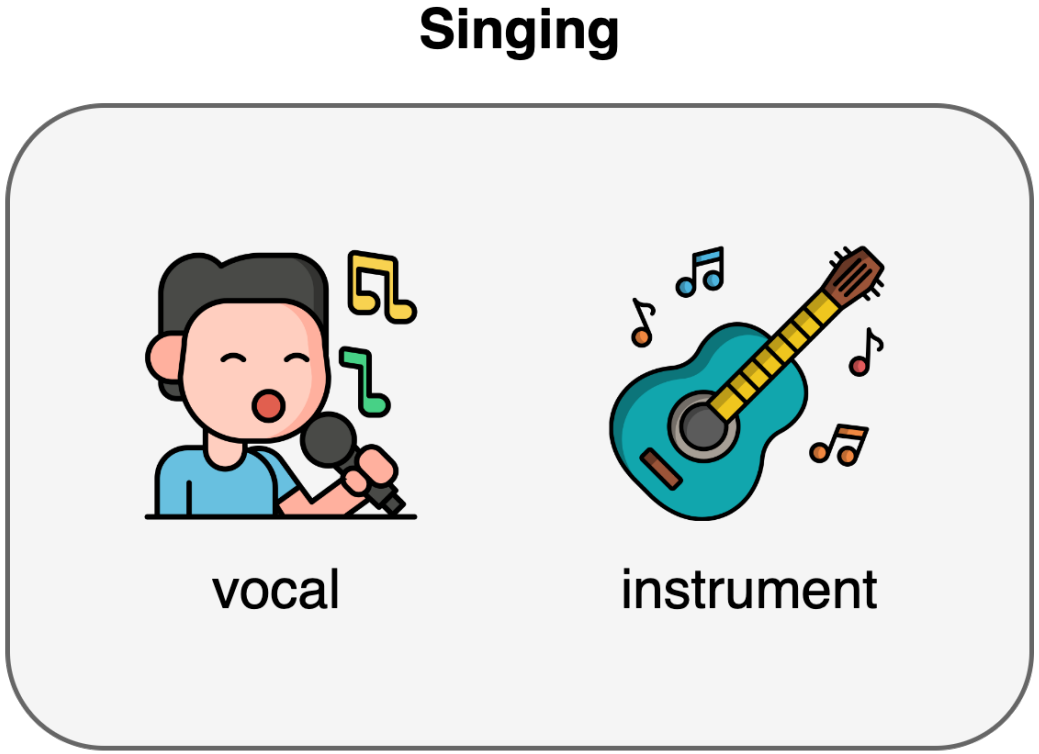


SingGraph: Beat-matching augmentation

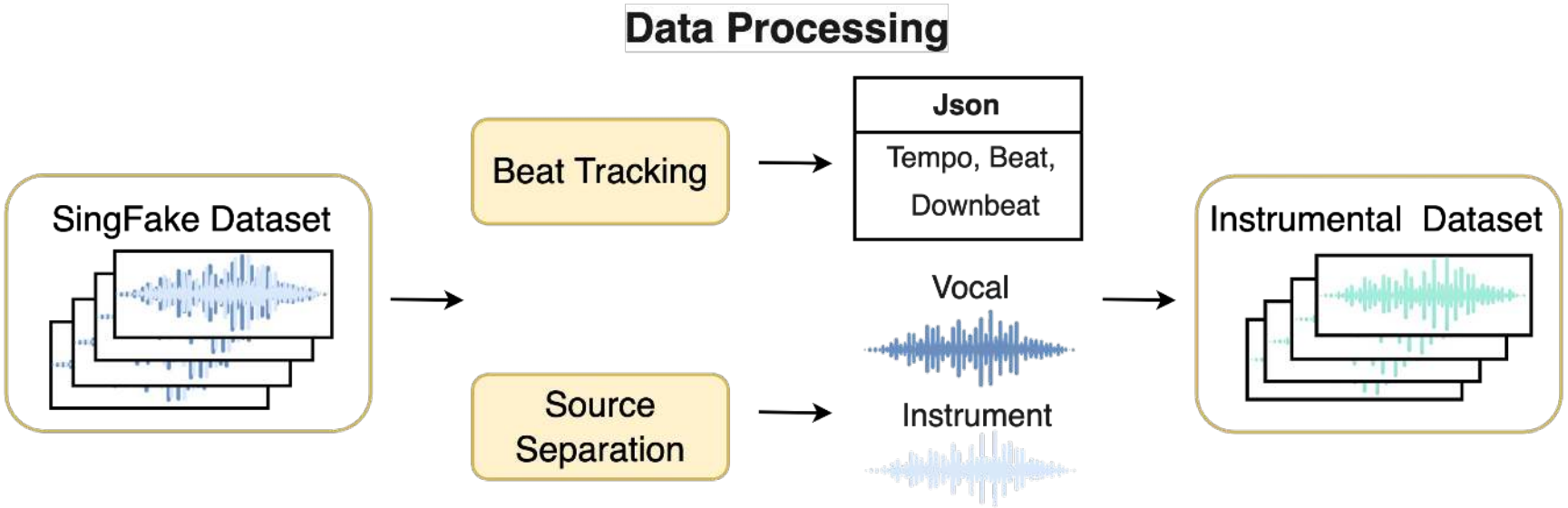
Intuition Idea

Vocals usually need to follow the **rhythm** of the music.

If we can find **instrumental music with the same rhythm**, which means we can construct more instrumental and vocal pairs for training.

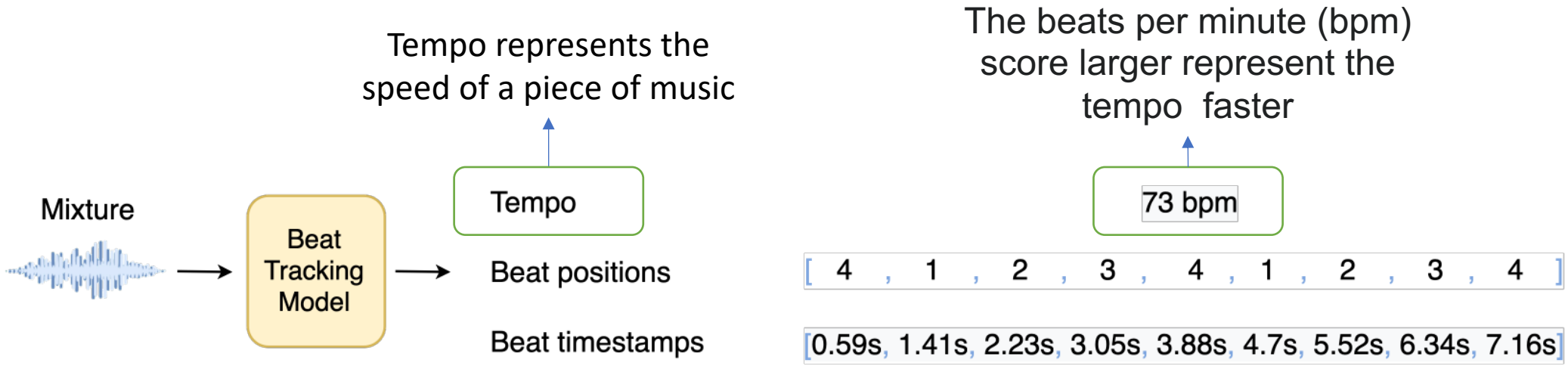


SingGraph: Beat-matching augmentation



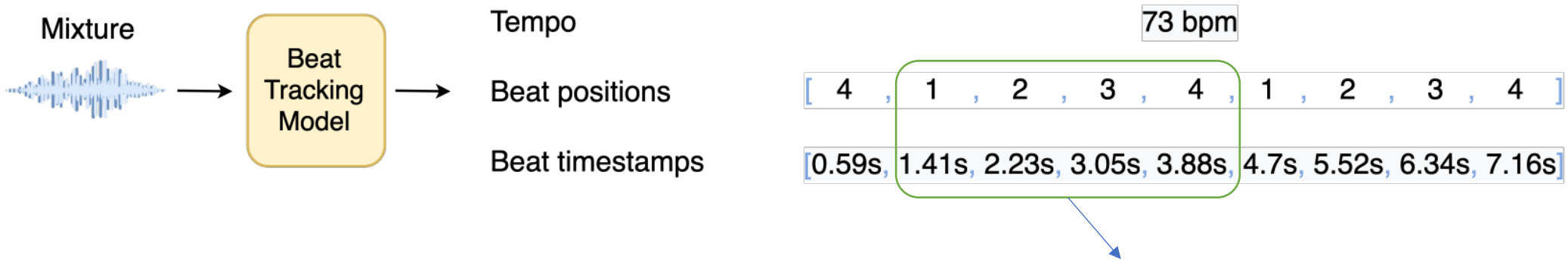
SingGraph: Beat-matching augmentation

Concept of music rhythm



SingGraph: Beat-matching augmentation

Concept of music rhythm

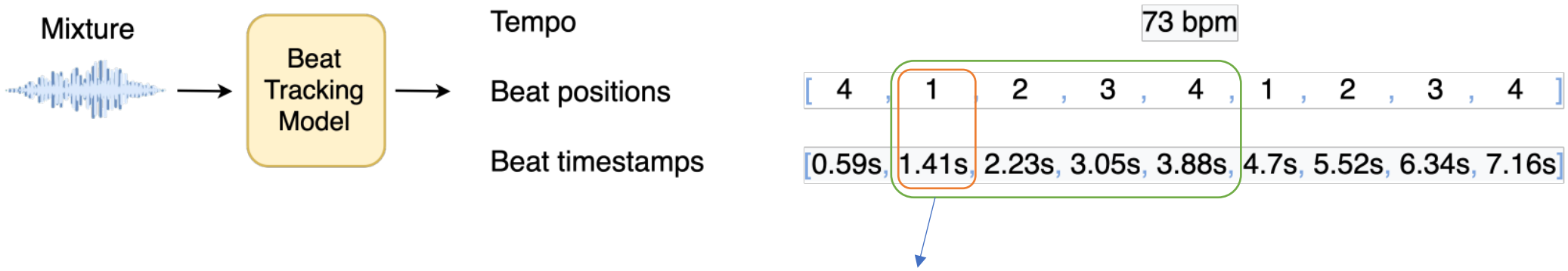


Bar

A unit of time that contains a specific number of beats played at a particular tempo

SingGraph: Beat-matching augmentation

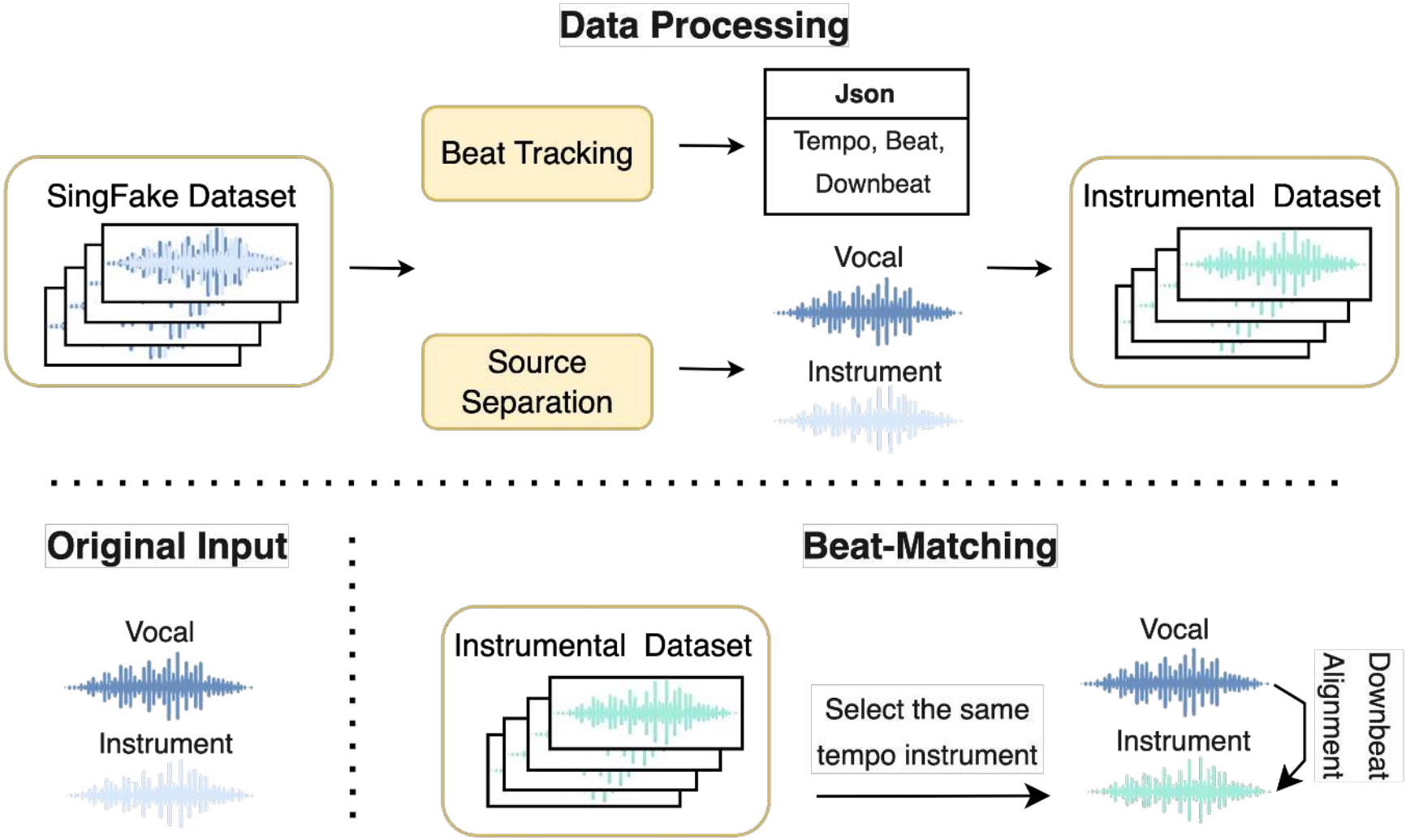
Concept of music rhythm



Downbeat

The downbeat is the rhythmic anchor of the musical phrases

SingGraph: Beat-matching augmentation



Experiment

Experiment Setting

- SingFake dataset [Zang et al., 2023]



Seen Singer



Unseen Singer

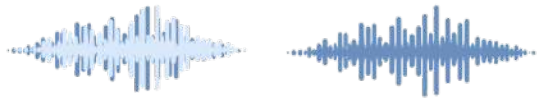


Unseen Singer with Distortion

Simple → Difficult

Research Question

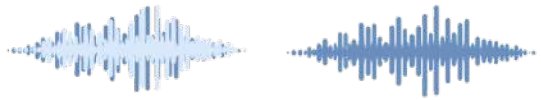
Are instrumental music cues helpful?



Mixture (M) v.s. Vocal (V)

Research Question

Are instrumental music cues helpful?



Mixture (M) v.s. Vocal (V)

Does it require separate encoders for encoding?



Single v.s. Dual Encoder

Research Question

Are instrumental music cues helpful?



Mixture (M) v.s. Vocal (V)

Does it require separate encoders for encoding?



Single v.s. Dual Encoder

Does it require input from vocal and instrumental sources?



Instrument-Vocal (IV)

Research Question

Are instrumental music cues helpful?



Mixture (M) v.s. Vocal (V)

Does it require separate encoders for encoding?



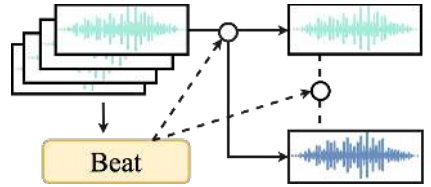
Single v.s. Dual Encoder

Does it require input from vocal and instrumental sources?



Instrument-Vocal (IV)

Is beat-matching augmentation effective?



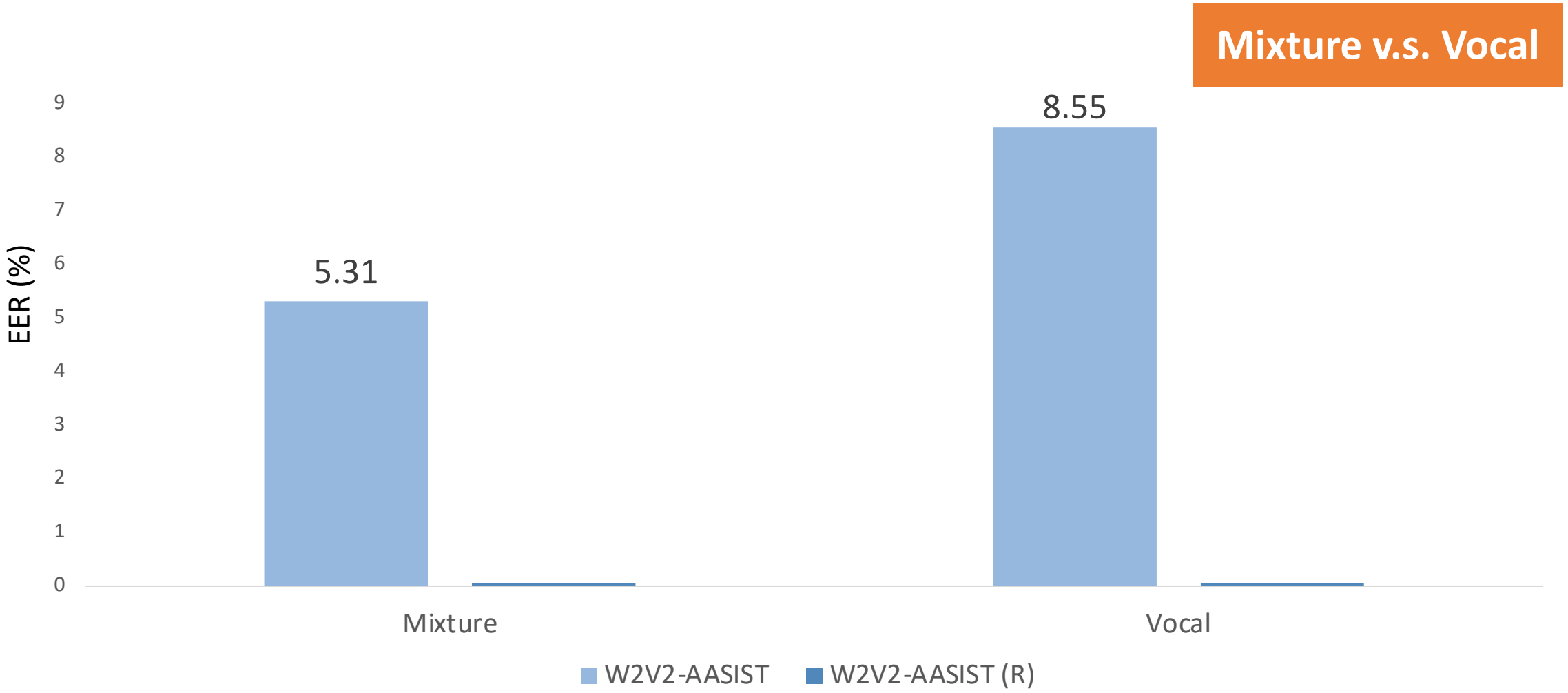
Q1: Are instrumental music cues helpful?

- Evaluation: Seen singer
- Model
  - W2V2-AASIST
- Input setting
  - Mixture
  - Vocal
- Augmentation
  - RawBoost (R)



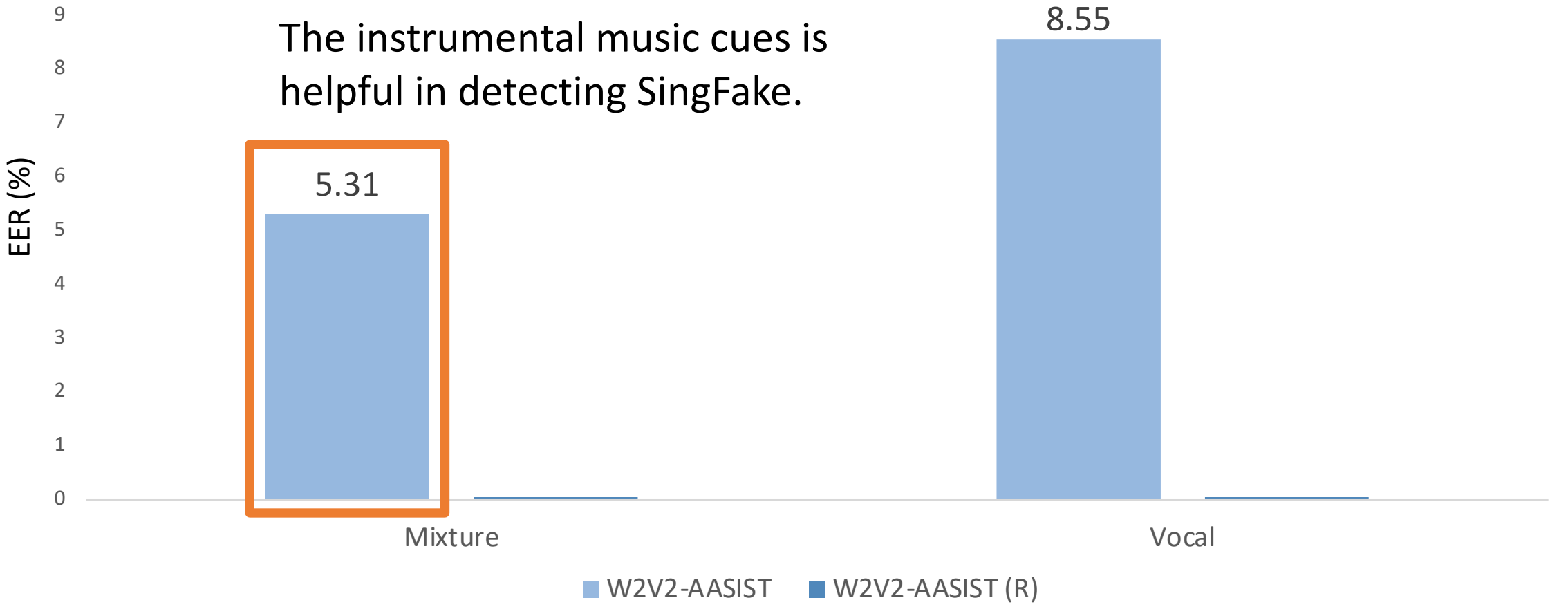
W2V2-AASIST Model

Q1: Are instrumental music cues helpful?



Q1: Are instrumental music cues helpful?

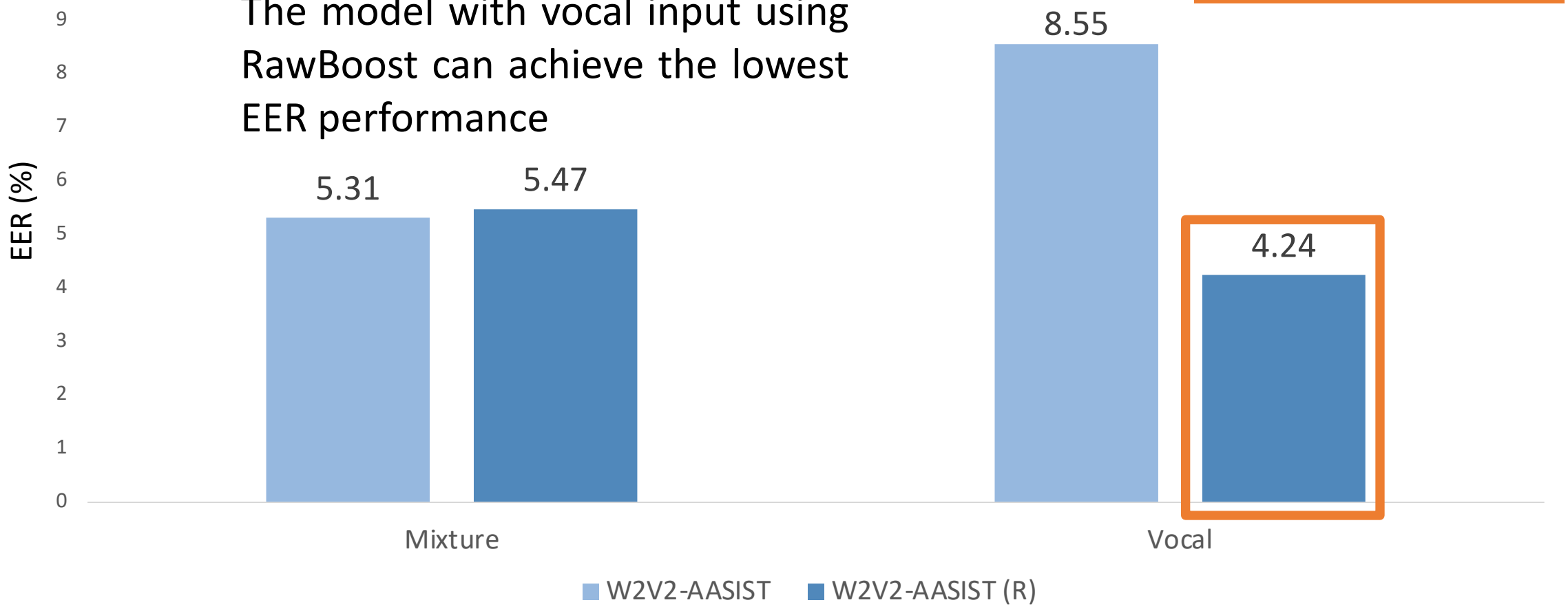
Mixture v.s. Vocal



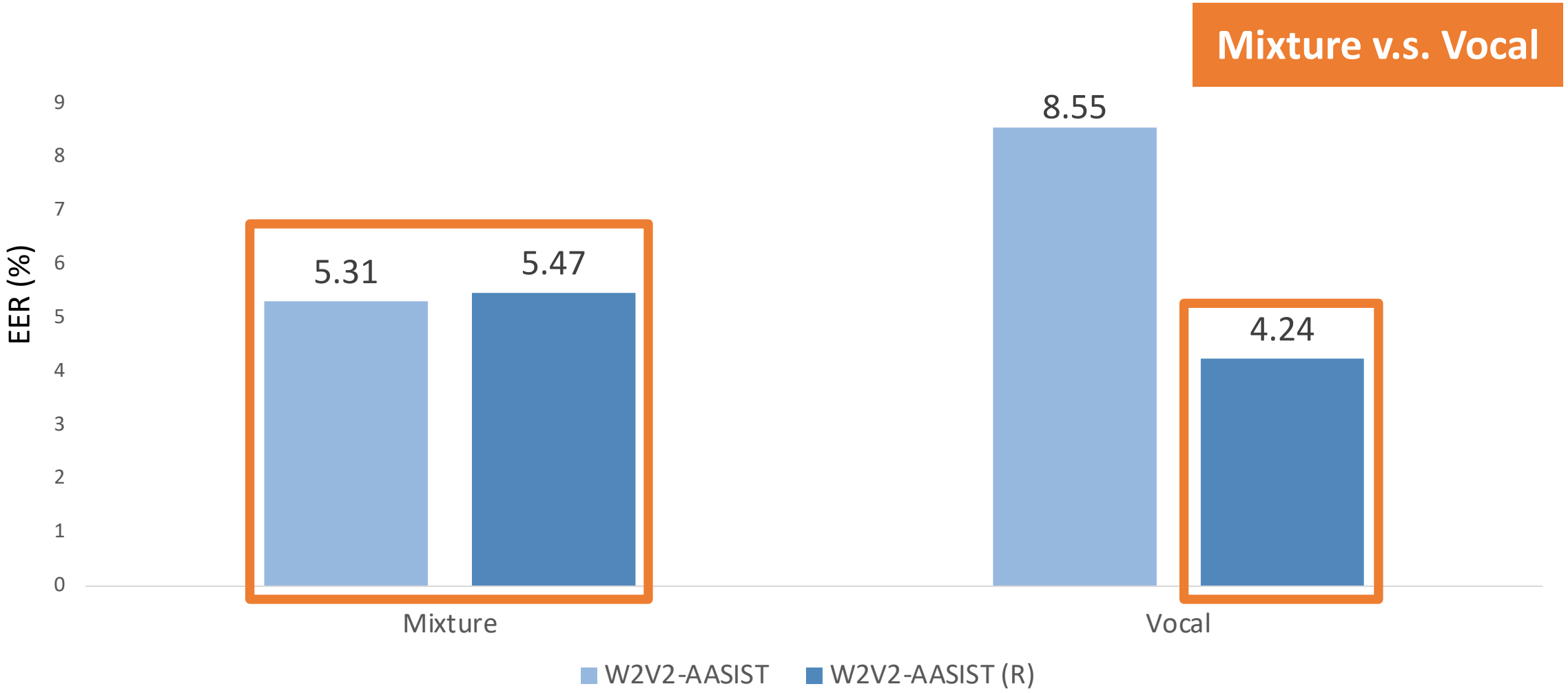
Q1: Are instrumental music cues helpful?

Mixture v.s. Vocal

The model with vocal input using RawBoost can achieve the lowest EER performance



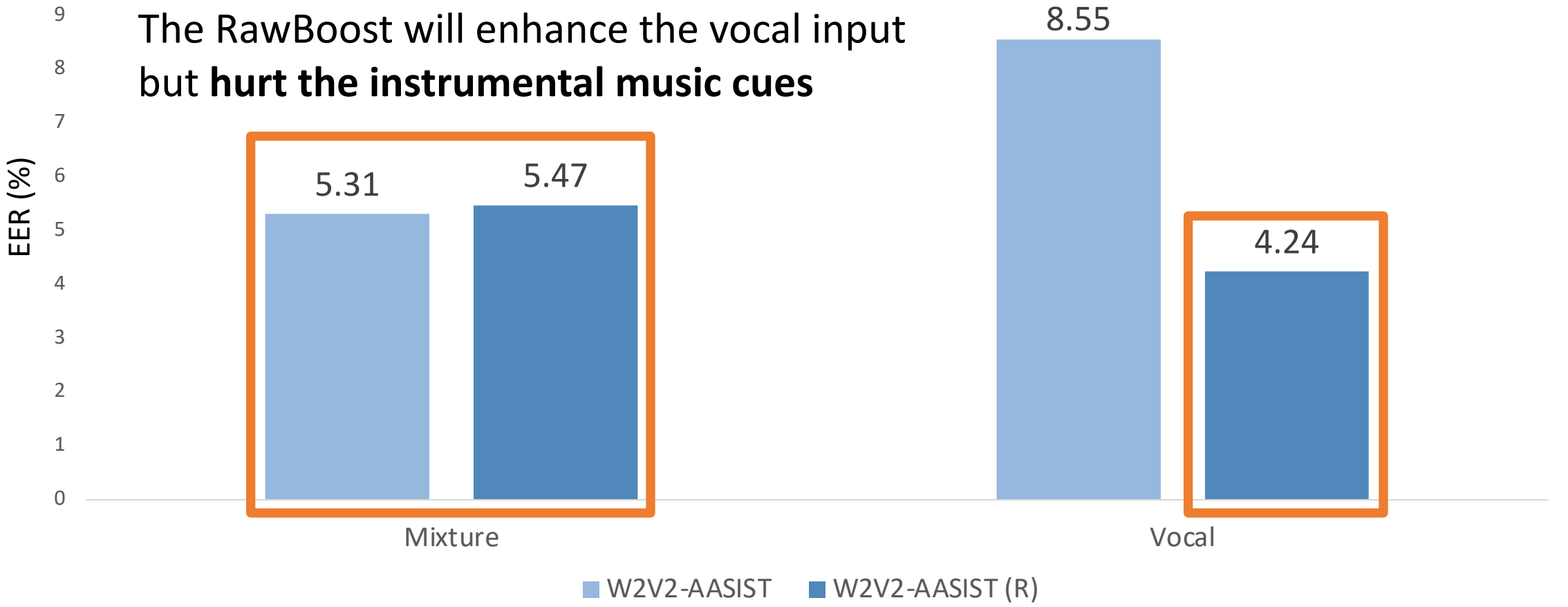
Q1: Are instrumental music cues helpful?



Q1: Are instrumental music cues helpful?

Mixture v.s. Vocal

The RawBoost will enhance the vocal input but **hurt the instrumental music cues**

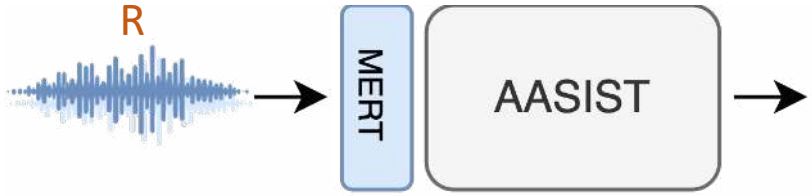


Q2: Does it require separate encoders for encoding?

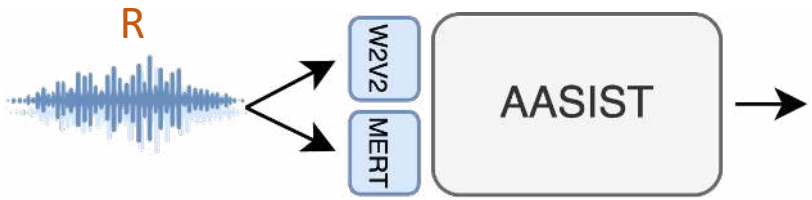
- Evaluation: Seen singer
- Single Encoder Model
  - W2V2-AASIST
  - MERT-AASIST
- Dual Encoder Model
  - W2V2-MERT-AASIST
- Input setting
  - Mixture
  - Vocal
- Augmentation
  - RawBoost (R)



W2V2-AASIST Model

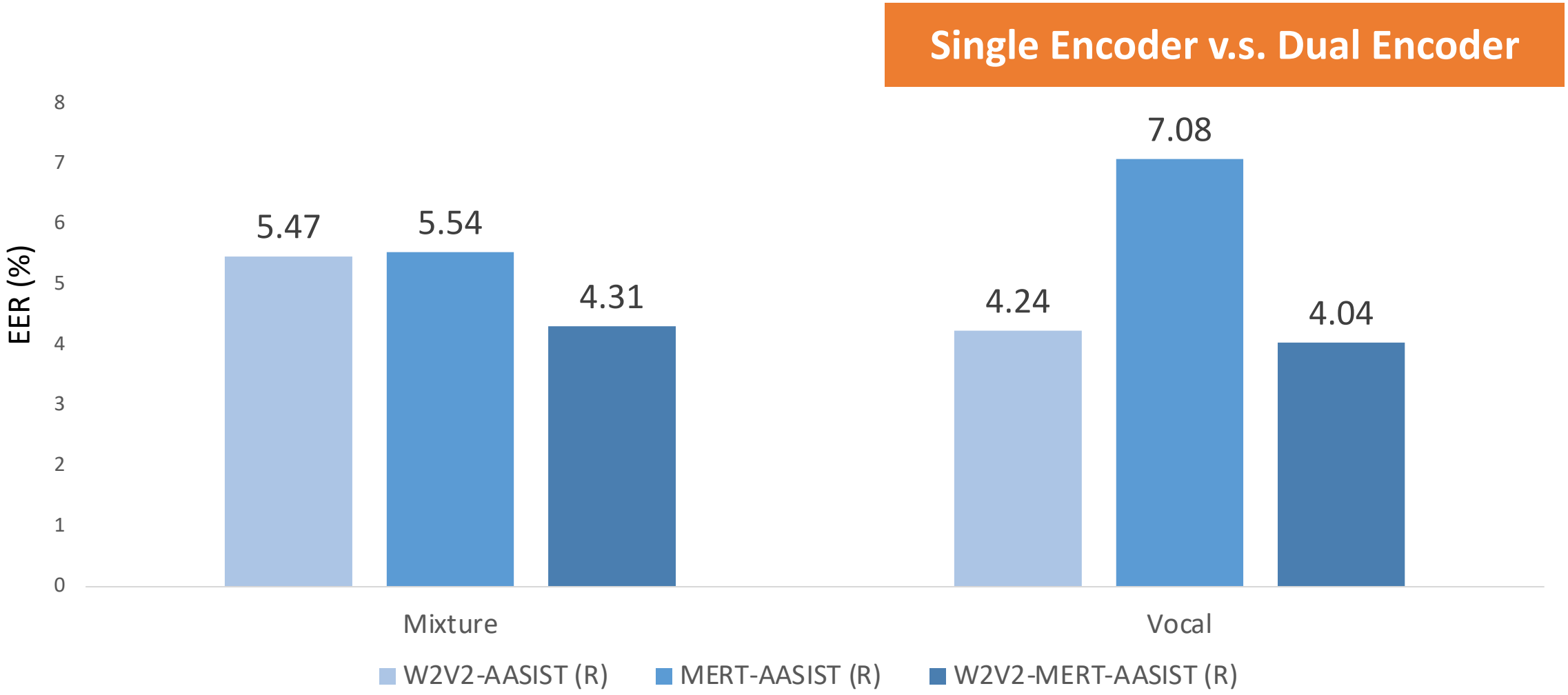


MERT-AASIST Model

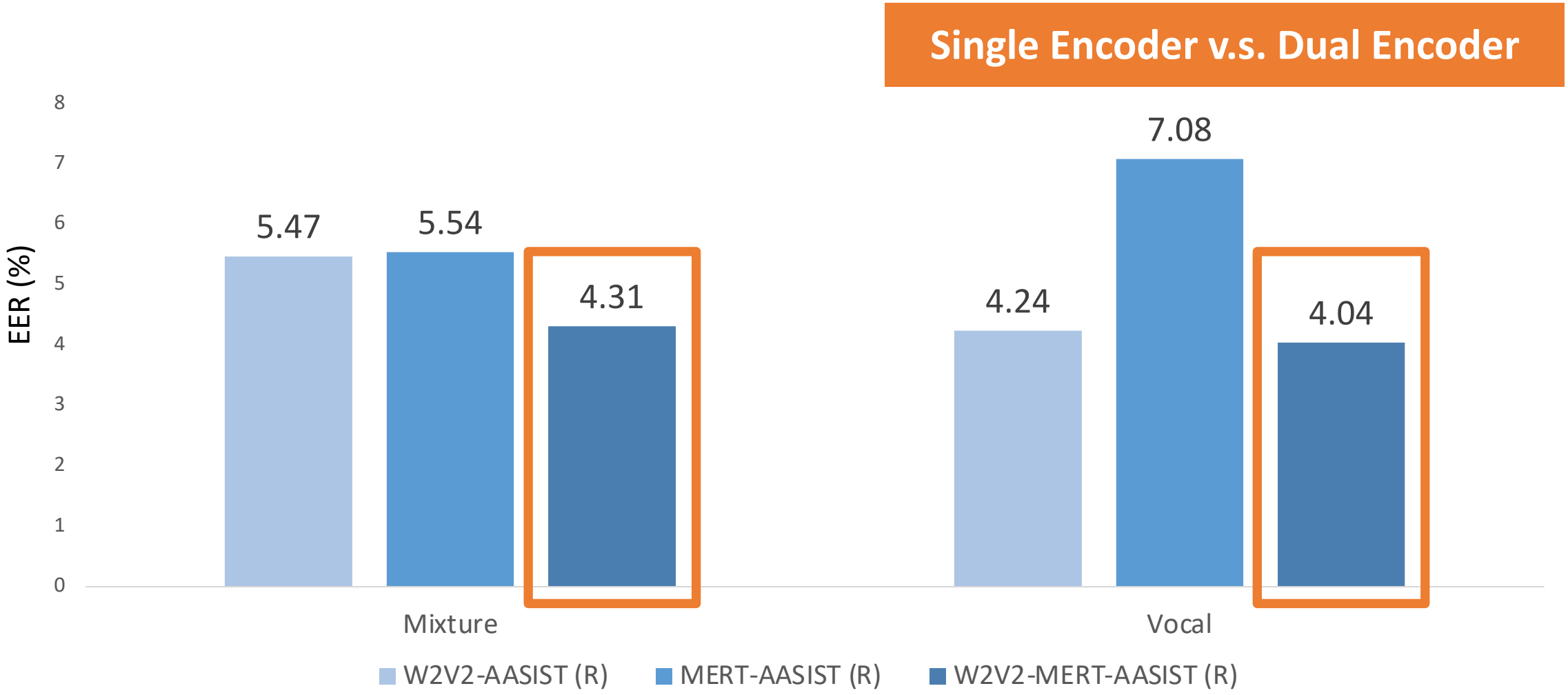


W2V2-MERT-AASIST Model

Q2: Does it require separate encoders for encoding?



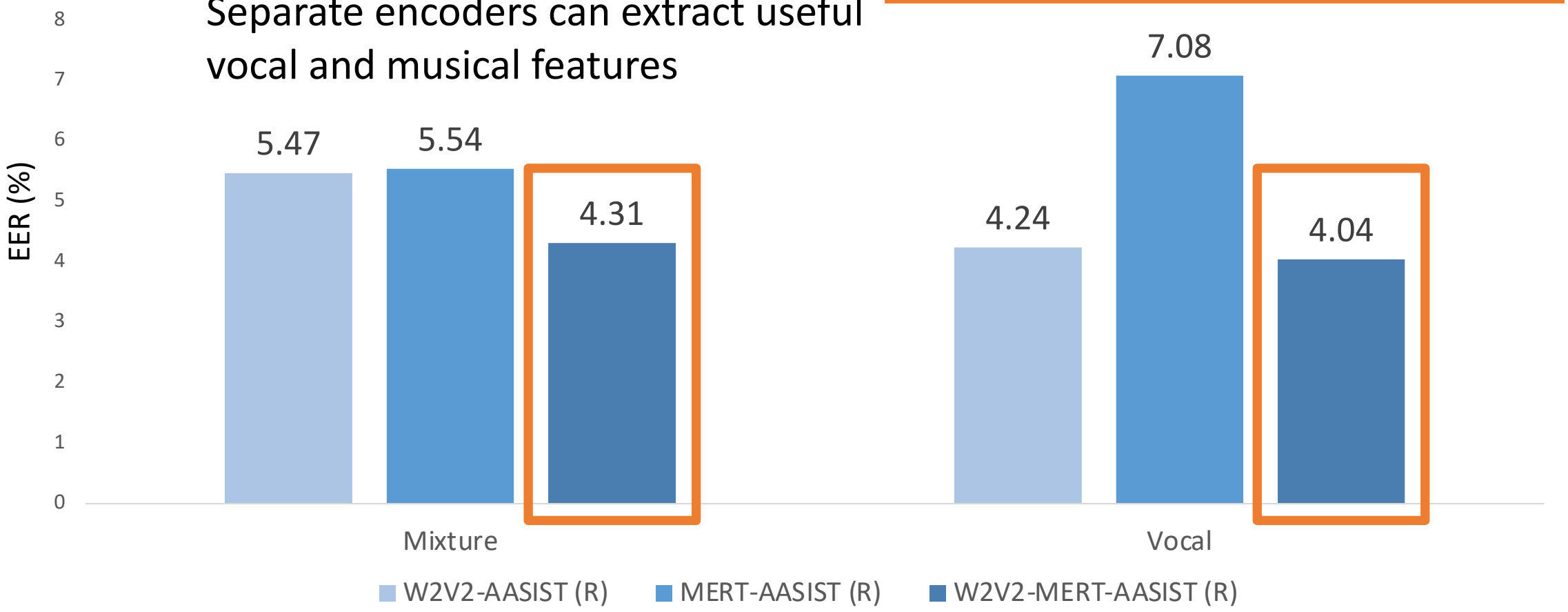
Q2: Does it require separate encoders for encoding?



Q2: Does it require separate encoders for encoding?

Single Encoder v.s. Dual Encoder

Separate encoders can extract useful vocal and musical features



Q3: Does it require input from vocal and instrumental sources?

- Evaluation: Seen singer
- Single Source Input
  - Model
    - W2V2-MERT-AASIST
    - W2V2-AASIST
    - MERT-AASIST
  - Configuration
    - Input: Mixture or vocal
    - Augmentation: RawBoost (R)
- Dual Source Input
  - W2V2-MERT-AASIST (IV)
    - Augmentation for vocal: RawBoost (R)



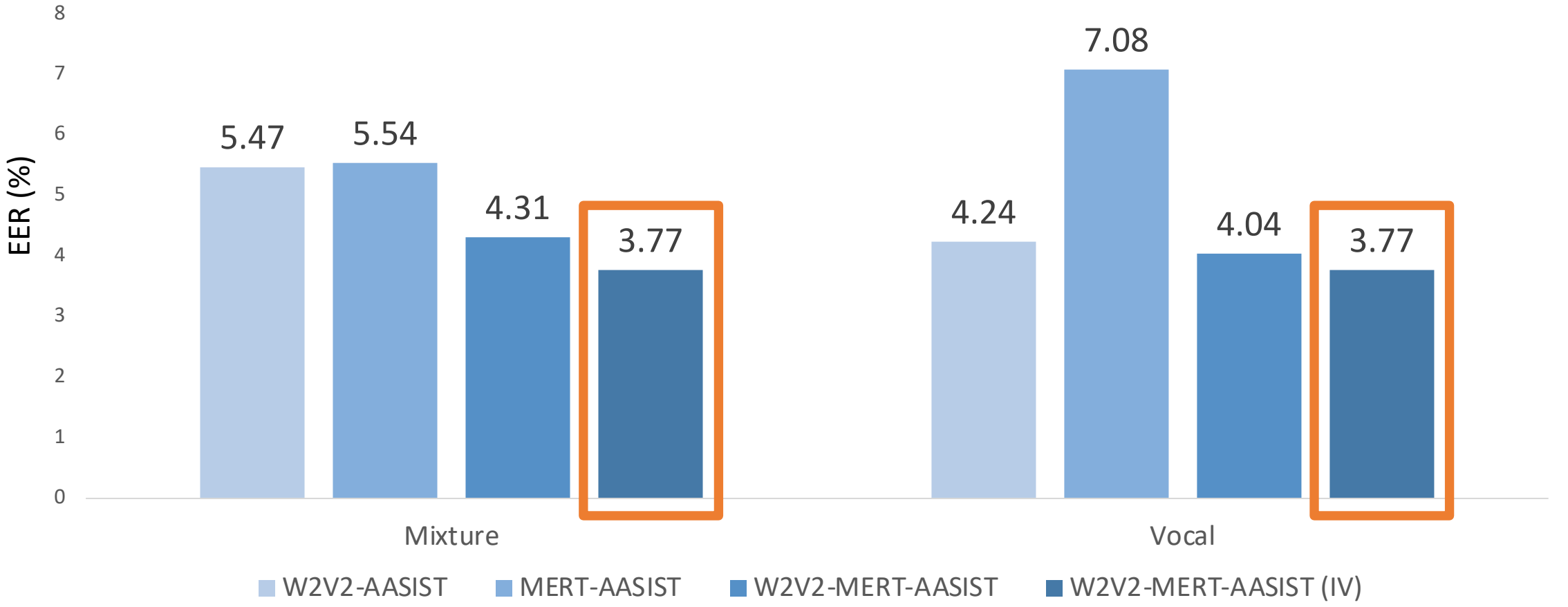
W2V2-MERT-AASIST Model



W2V2-MERT-AASIST (IV) Model

Q3: Does it require input from vocal and instrumental sources?

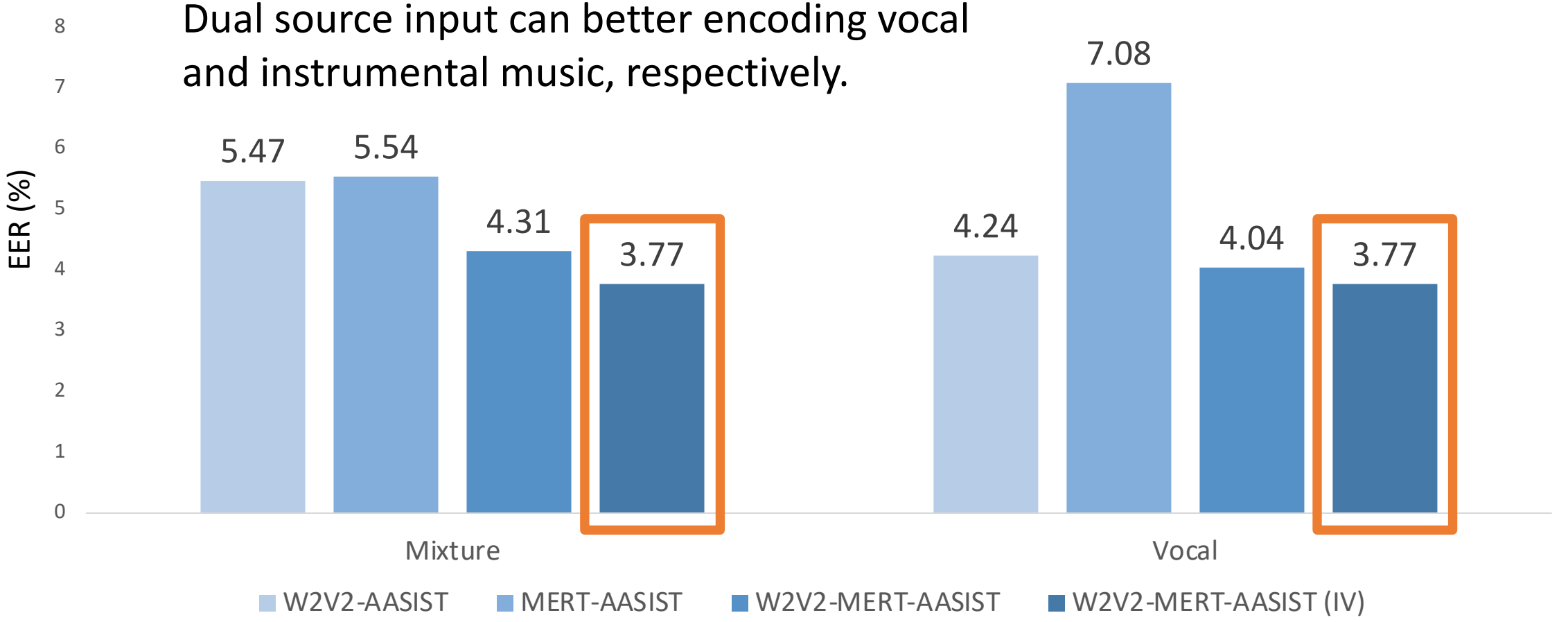
Instrumental and vocal sources



Q3: Does it require input from vocal and instrumental sources?

Instrumental and vocal sources

Dual source input can better encoding vocal and instrumental music, respectively.



Q4: Is beat-matching augmentation effective?

- Evaluation: Three scenarios
- Single Source Input
  - Model
    - W2V2-MERT-AASIST
    - W2V2-AASIST
    - MERT-AASIST
  - Configuration
    - Input: Mixture (M) or Vocal (V)
    - **Augmentation:** RawBoost (R)
- Dual Source Input
  - W2V2-MERT-AASIST (IV)
    - **Augmentation for vocal:** RawBoost (R)
  - W2V2-MERT-AASIST (IV)
    - **Augmentation for vocal:** RawBoost (R)
    - **Augmentation for instrument music:** Beat-matching (**B**)



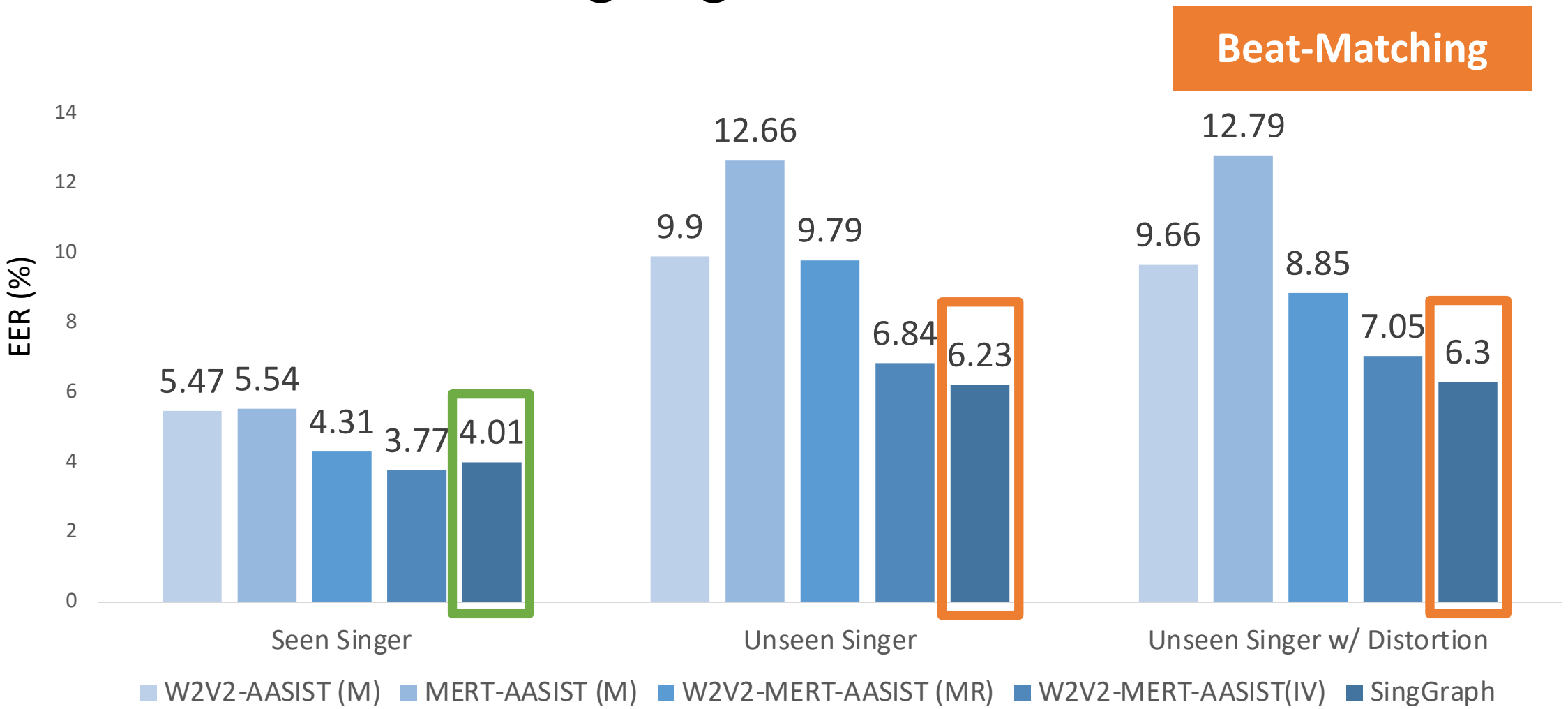
W2V2-MERT-AASIST (IV) Model



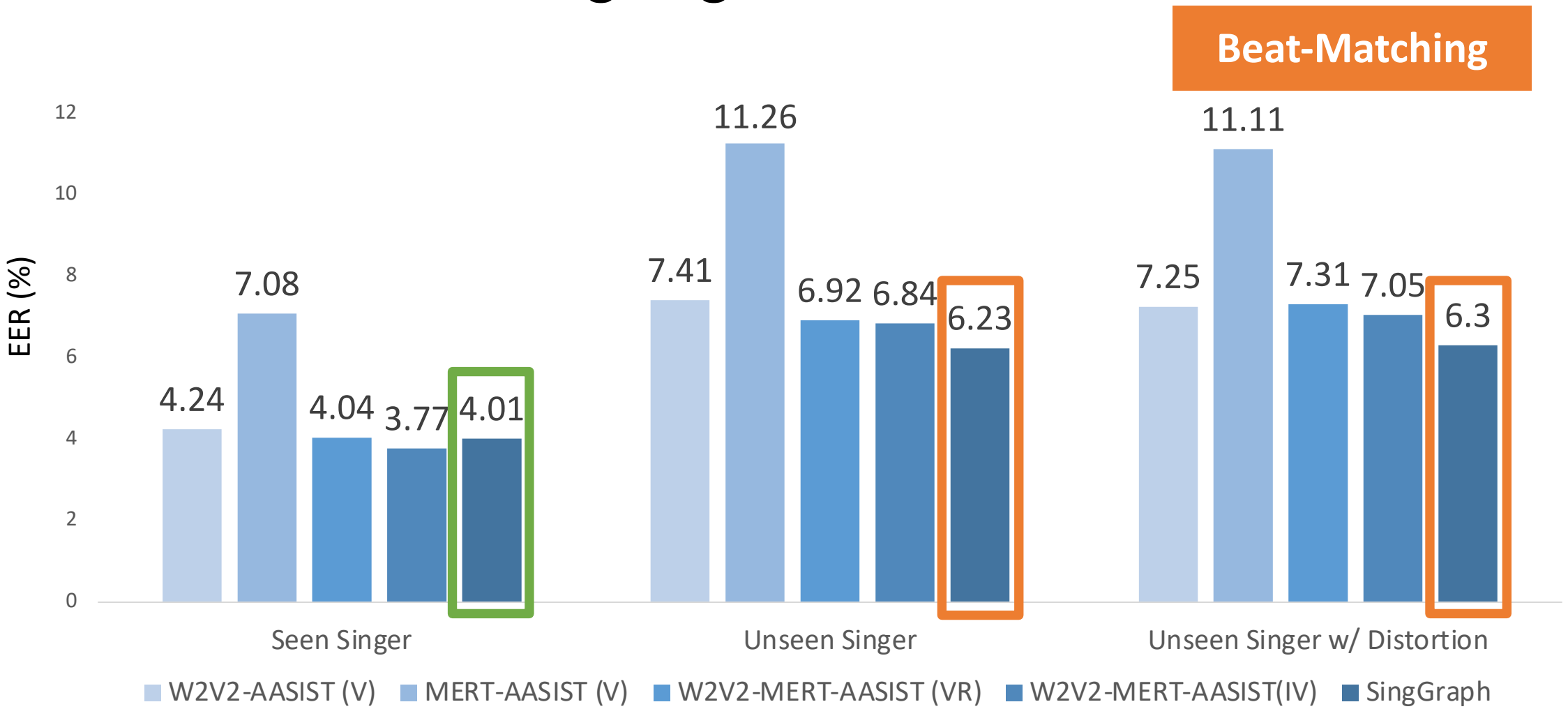
SingGraph Model

SingGraph model

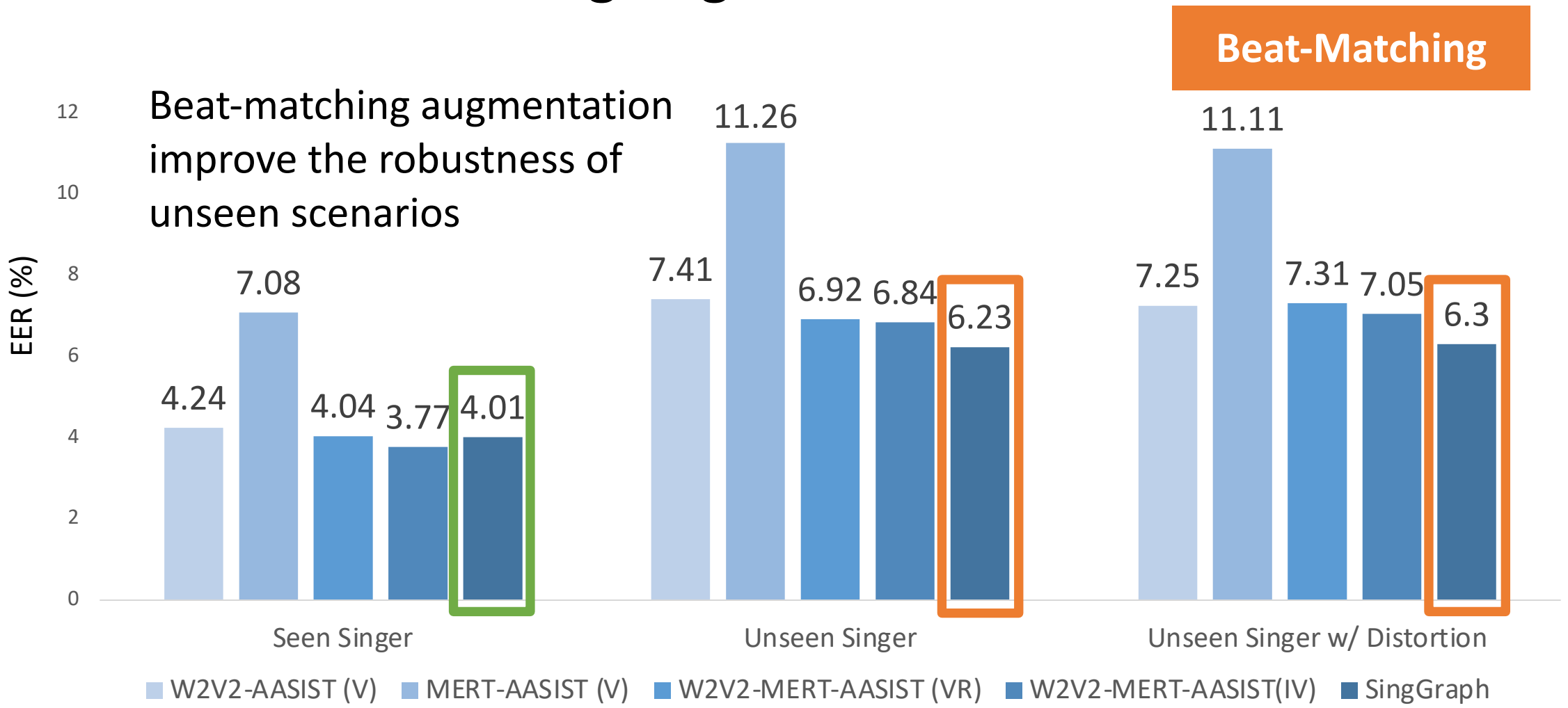
Q4: Is beat-matching augmentation effective?



Q4: Is beat-matching augmentation effective?



Q4: Is beat-matching augmentation effective?



Comparison with the state-of-the-art (SoTA) model

Method	Setup	Seen Singer EER (%)	Unseen Singer EER (%)	Unseen Singer w/ Distortion EER (%)
AASIST Model [3]	Mixture	7.29	11.54	17.29
	Vocals	8.37	10.65	13.07
Spec-ResNet [14]	Mixture	14.88	22.59	24.15
	Vocals	11.86	19.69	21.54
LFCC-ResNet [15]	Mixture	21.35	32.40	31.85
	Vocals	15.88	22.56	23.62
W2V2-AASIST [16]	Mixture	**4.62**	**8.23**	13.62
	Vocals	5.39	9.10	**10.03**

Comparison with the state-of-the-art (SoTA) model

Method	Setup	Seen Singer EER (%)	Unseen Singer EER (%)	Unseen Singer w/ Distortion EER (%)
AASIST Model [3]	Mixture	7.29	11.54	17.29
	Vocals	8.37	10.65	13.07
Spec-ResNet [14]	Mixture	14.88	22.59	24.15
	Vocals	11.86	19.69	21.54
LFCC-ResNet [15]	Mixture	21.35	32.40	31.85
	Vocals	15.88	22.56	23.62
W2V2-AASIST [16]	Mixture	**4.62**	**8.23**	13.62
	Vocals	5.39	9.10	**10.03**
SingGraph (Our)	IV	**4.01** (+13.2)	**6.23** (+24.3)	**6.30** (+37.1)

Conclusion

- Vocal and instrumental music should be encoded separately
  - We encode vocal and instrument via Wav2vec2 and MERT, respectively.
- Beat-matching augmentation is helpful in boosting SingFake detection
- Our SingGraph model achieves the SoTA SingFake detection results

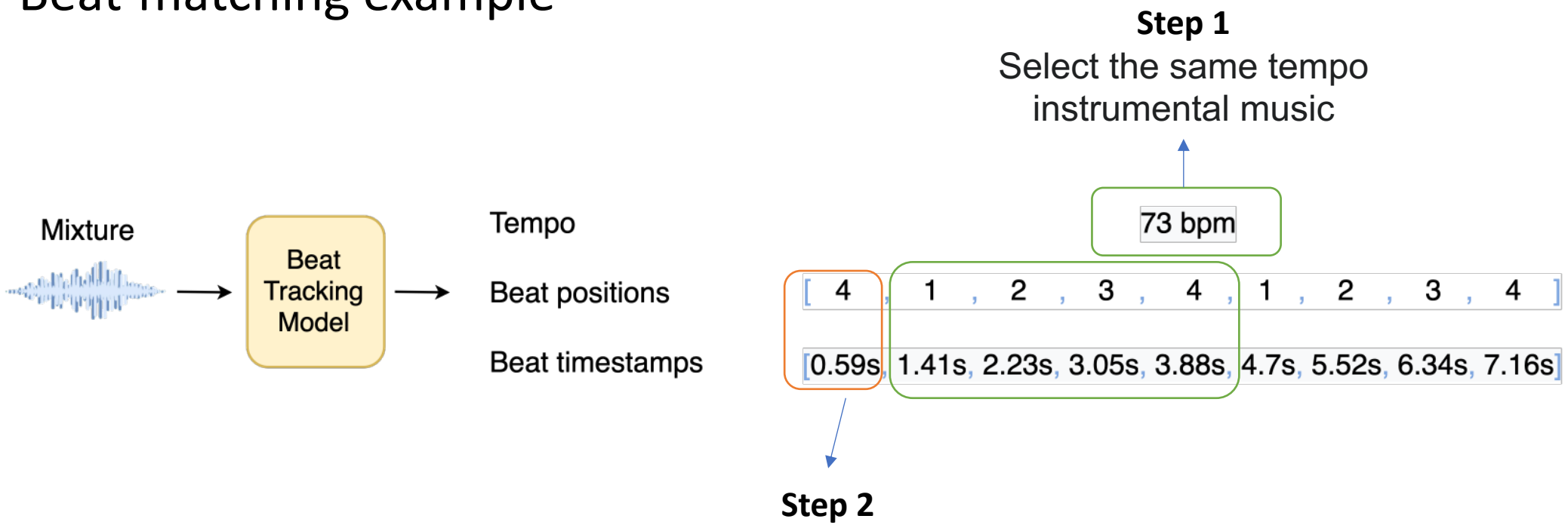
Paper

Released Code

Personal Website

Appendix

Beat-matching example



1. The first position is not always the downbeat
2. Need **downbeat alignment** between two sequences