Integration of speech separation, diarization, and recognition for multi-speaker meetings:

System description, comparison, and analysis

Desh Raj, Pavel Denisov, Zhuo Chen, Hakan Erdogan, Zili Huang, Maokui He, Shinji Watanabe, Jun Du, Takuya Yoshioka, Yi Luo, Naoyuki Kanda, Jinyu Li, Scott Wisdom, John R. Hershey

Multi-speaker meeting transcription

Input: recordings. Output: speaker-attributed transcription

10 min to **1-2 hours**

2-10 speakers

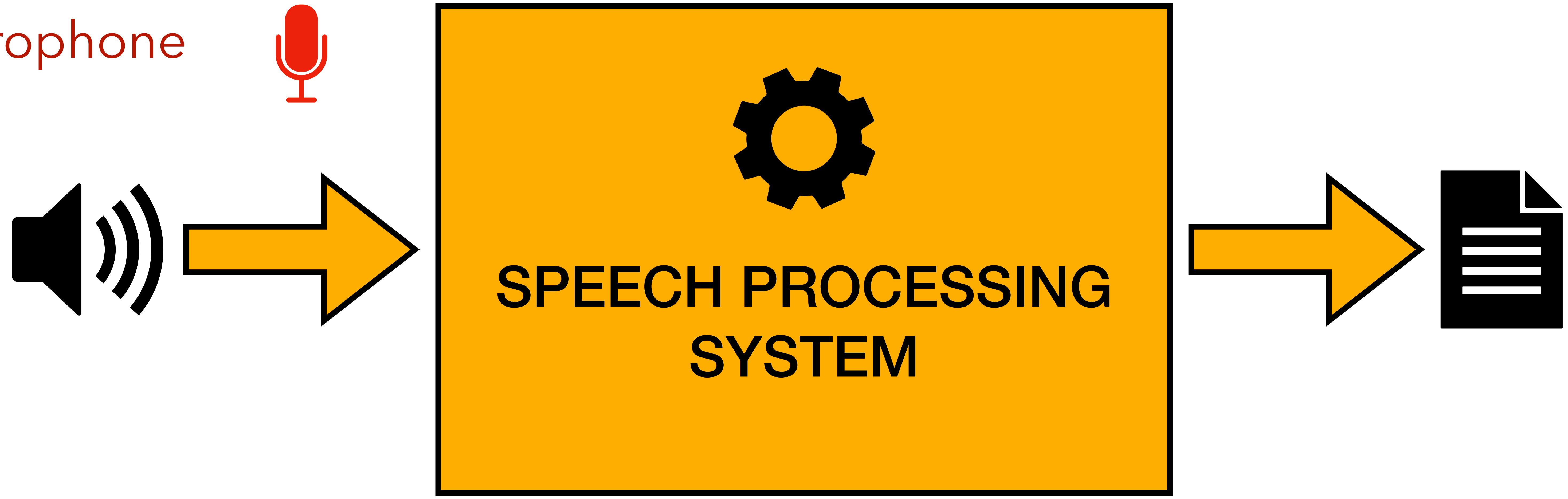
Typically **20% overlap**

Single/multi microphone

WHO said **WHAT** and **WHEN**

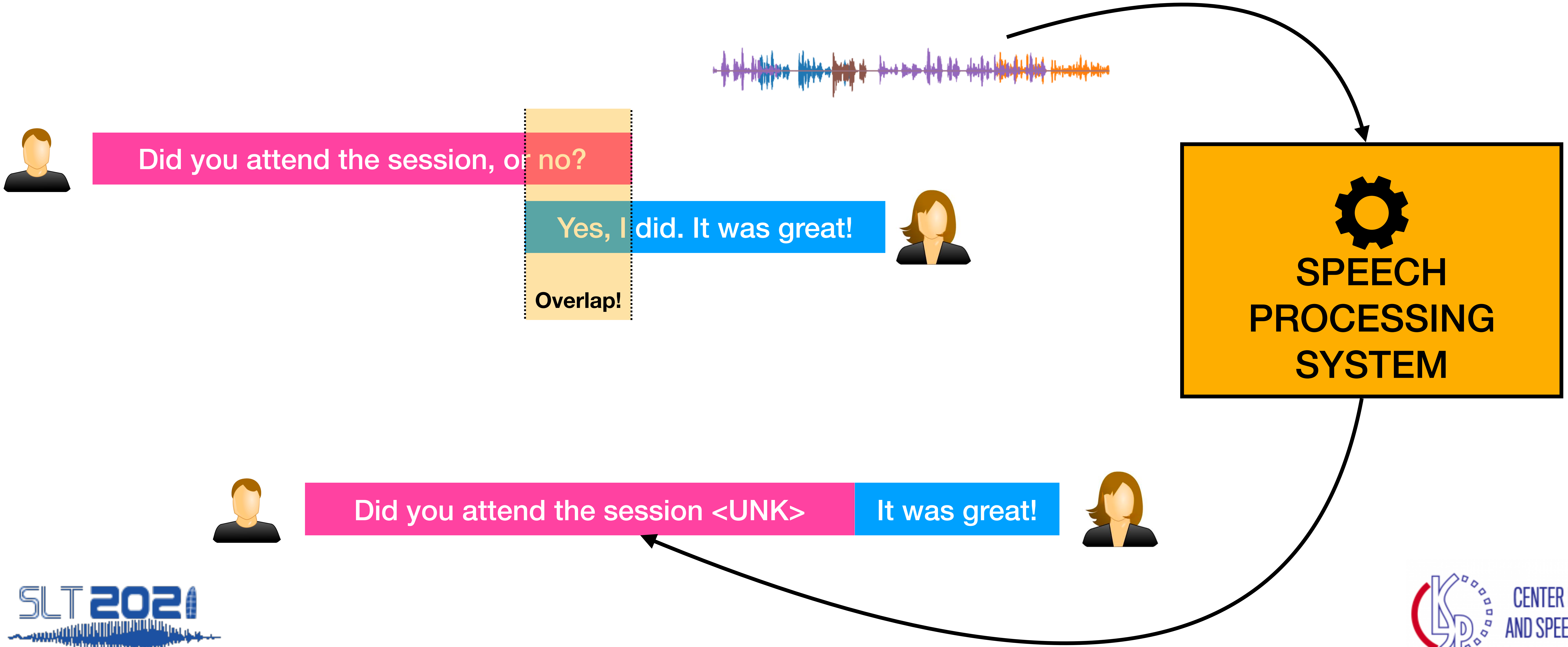
Speech recognition

Speaker diarization



Why is it difficult?

Overlapping speech affects both ASR and diarization outputs



SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Why is it difficult?

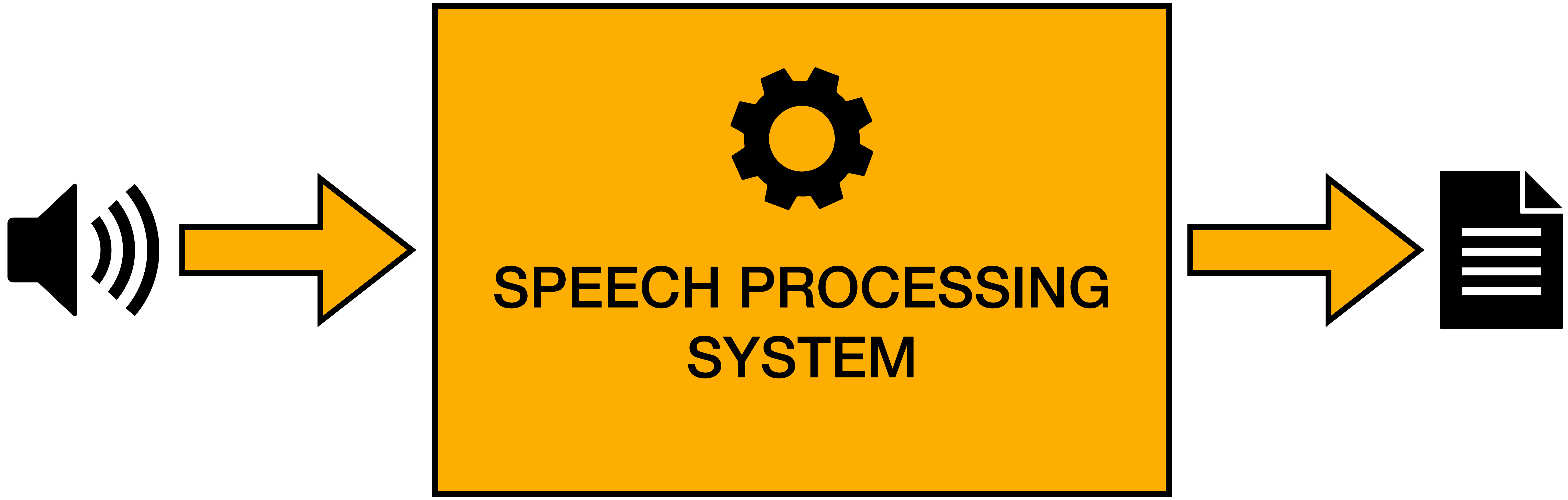
Overlapping speech affects both ASR and diarization outputs

- ASR models are typically trained on single-speaker utterances
- Conventional clustering-based diarization systems assume single-speaker segments

SPEECH PROCESSING SYSTEM

One possible solution

Separate the speech before applying diarization and ASR



We study this integration extensively…

…on the LibriCSS dataset*

10 min “mini-sessions”

8 speakers per recording

0–40% overlap ratio

7-microphone circular array

*MORE ON THIS LATER

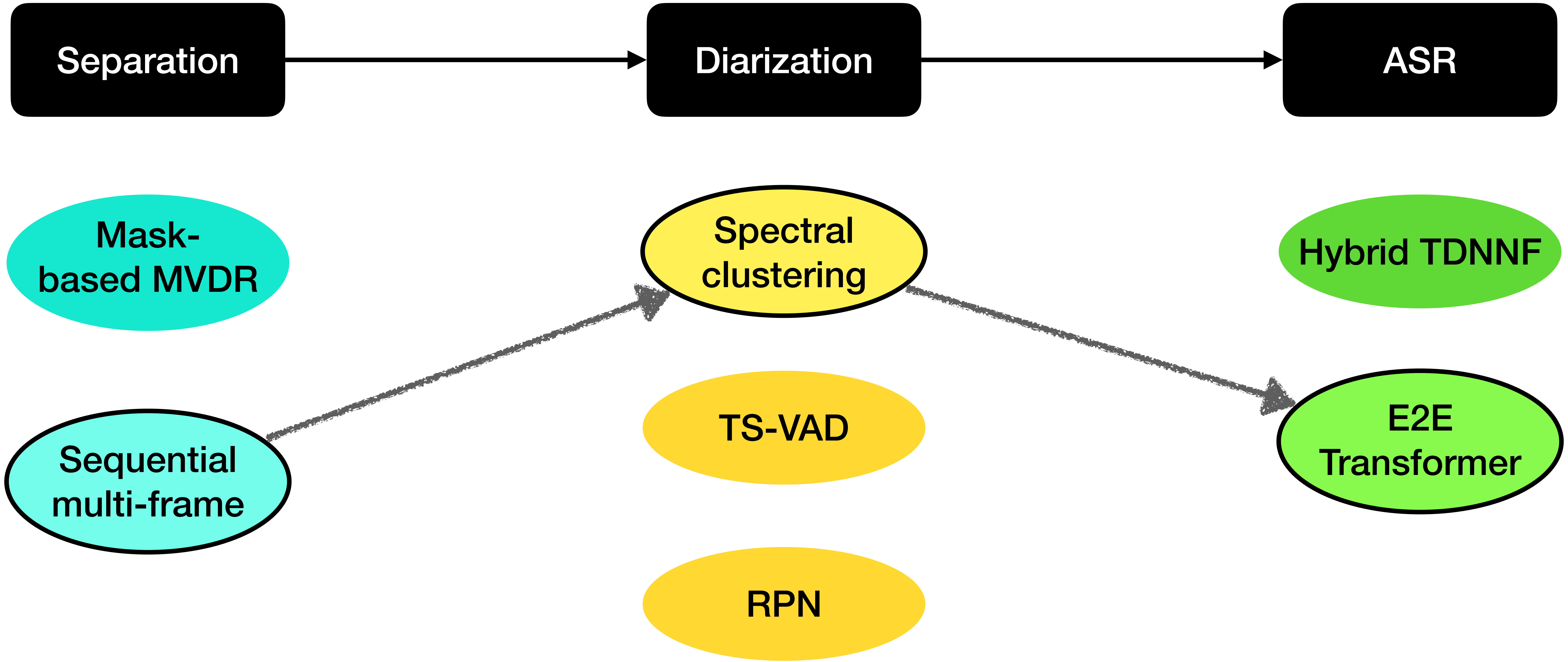
SLT 2021

Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, Zhong Meng, Yi Luo, J. Wu, and Jinyu Li, “Continuous speech separation: Dataset and analysis,” ICASSP 2020.

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Our final pipeline

Final cpWER = 12.7% (compared with 27.1% for "no separation" baseline)



cpWER = concatenated minimum-permutation word error rate

End of Highlight

Overview

- Modular pipeline: Too many options!
- Related Work: Integrated pipelines in literature
- More on the dataset (LibriCSS) and the metric (cpWER)
- Results and Discussion:
    - The separation component
    - The diarization component
    - The ASR component
- Where does your model fit in?

What components should you choose?

Separation → Diarization → ASR

Separation	Diarization	ASR
Mask-based MVDR	Clustering	Hybrid TDNNF
Sequential multi-frame	TS-VAD	E2E Transformer
	RPN	

Different models are optimized for **different objectives**

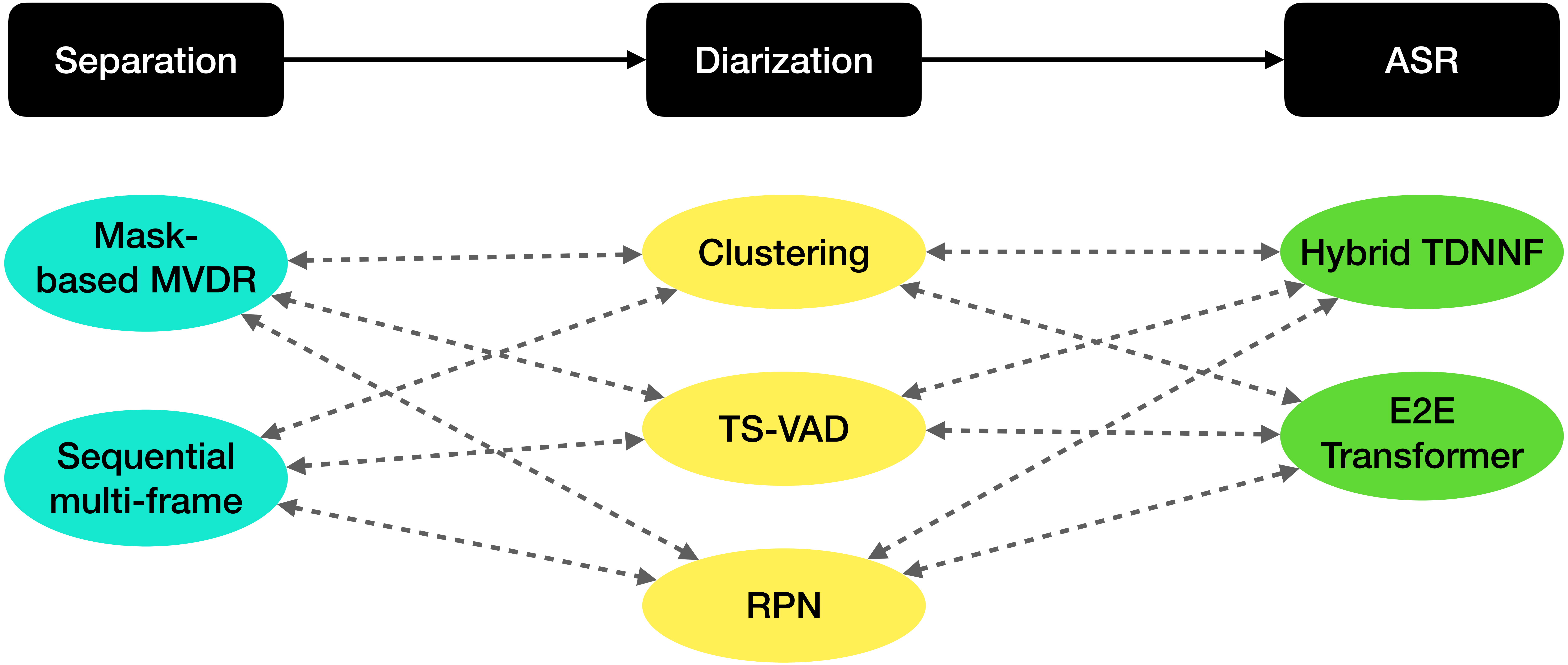
TS-VAD and RPN generate **overlapping** segments

Can a model trained on **clean speech** do well on separated audio?

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

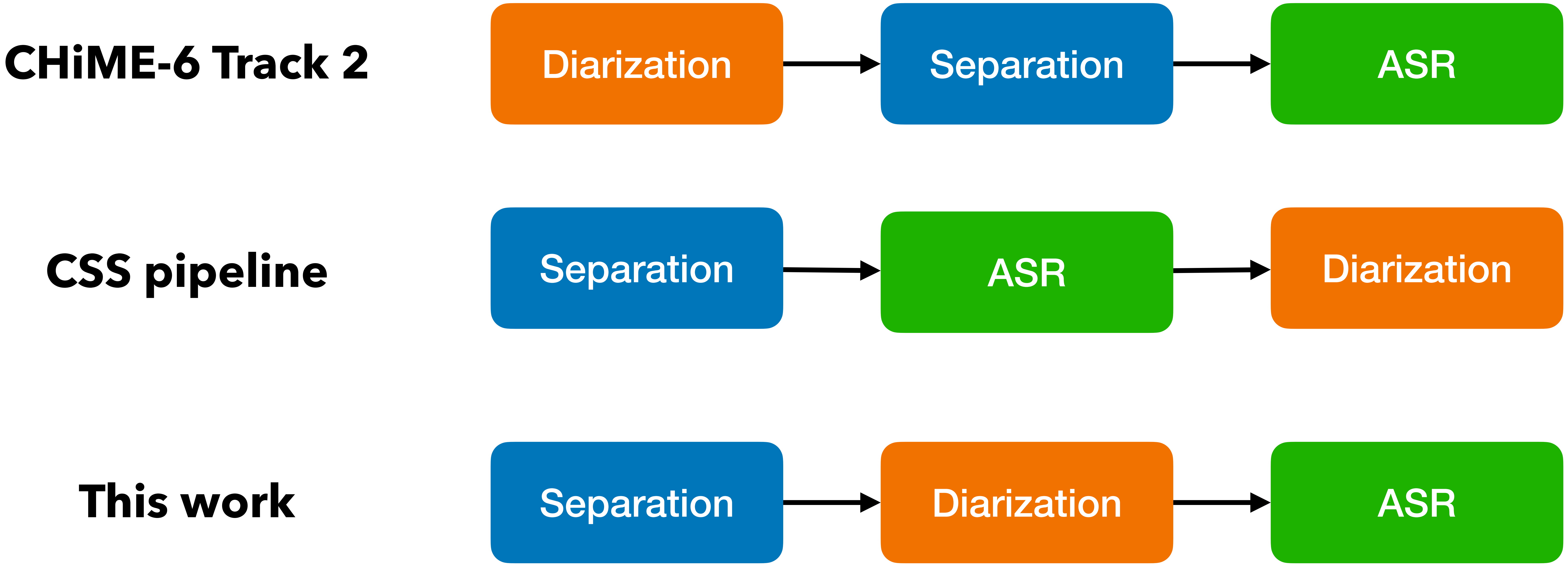
What components should you choose?



How do the components **interact** among themselves?

CHiME-6 and CSS

Different orders of the 3 main components



Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, Zhong Meng, Yi Luo, J. Wu, and Jinyu Li, “Continuous speech separation: Dataset and analysis,” ICASSP 2020.

Shinji Watanabe, Michael Mandel, Jon Barker, and Emmanuel Vincent, “CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings,” ArXiv, vol. abs/2004.09249, 2020.

CHiME-6 and CSS

Diarization: Overlap-aware? Across streams?

CHiME-6 Track 2	Diarization (overlap-aware) →	Separation →	ASR
CSS pipeline	Separation →	ASR →	Diarization (segments from ASR)
This work	Separation →	Diarization (cross-stream) →	ASR

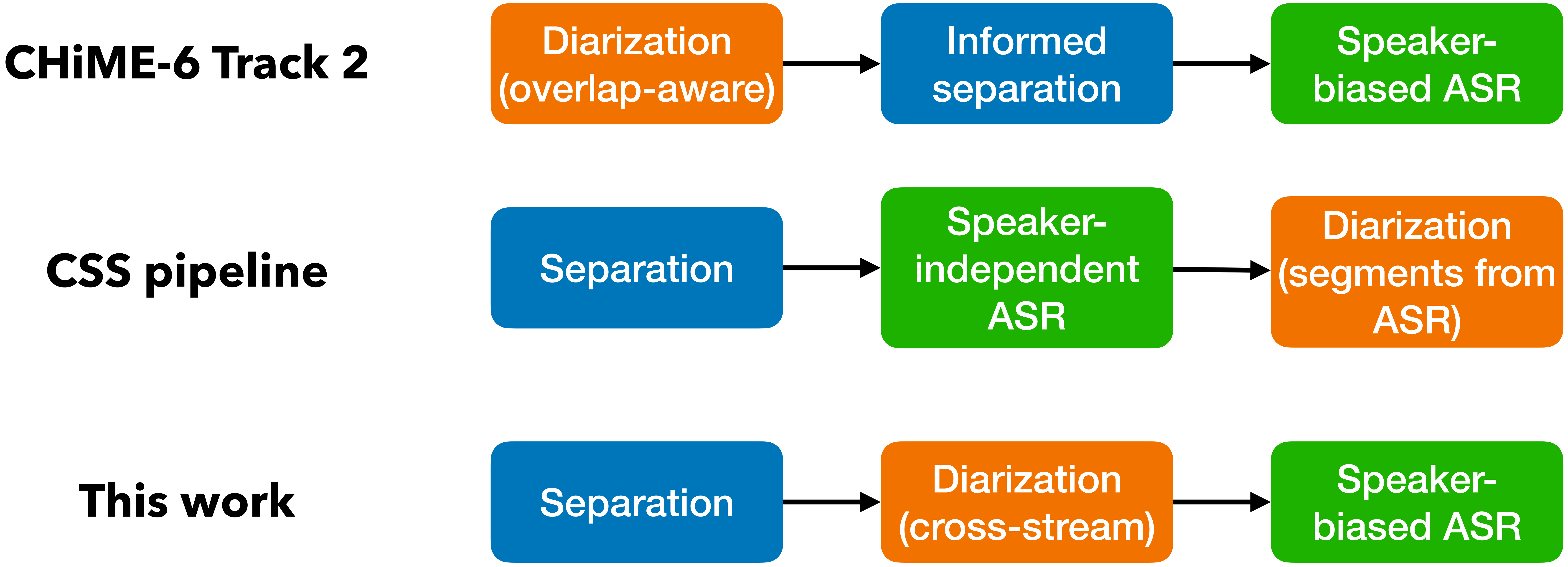
CHiME-6 and CSS

Separation: informed?

CHiME-6 Track 2	Diarization (overlap-aware) →	Informed separation →	ASR
CSS pipeline	Separation →	ASR →	Diarization (segments from ASR)
This work	Separation →	Diarization (cross-stream) →	ASR

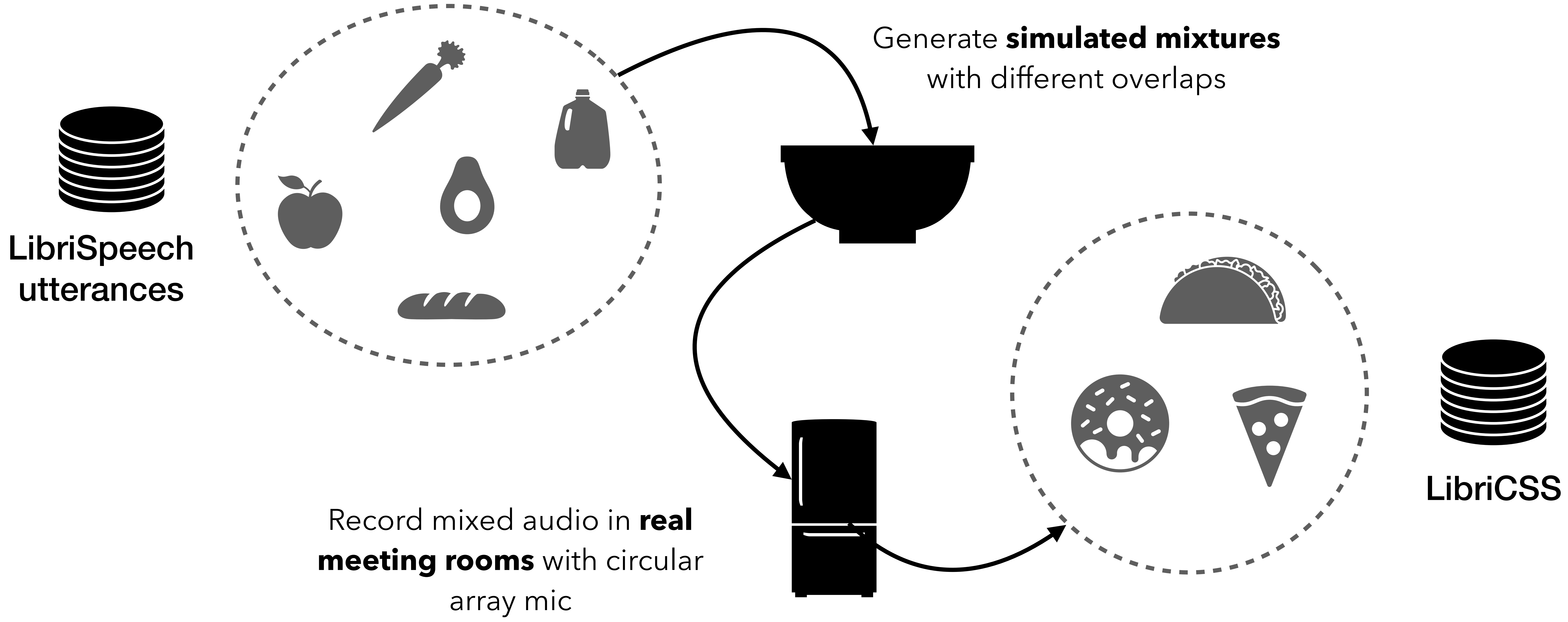
CHiME-6 and CSS

ASR: using speaker information?



More about LibriCSS

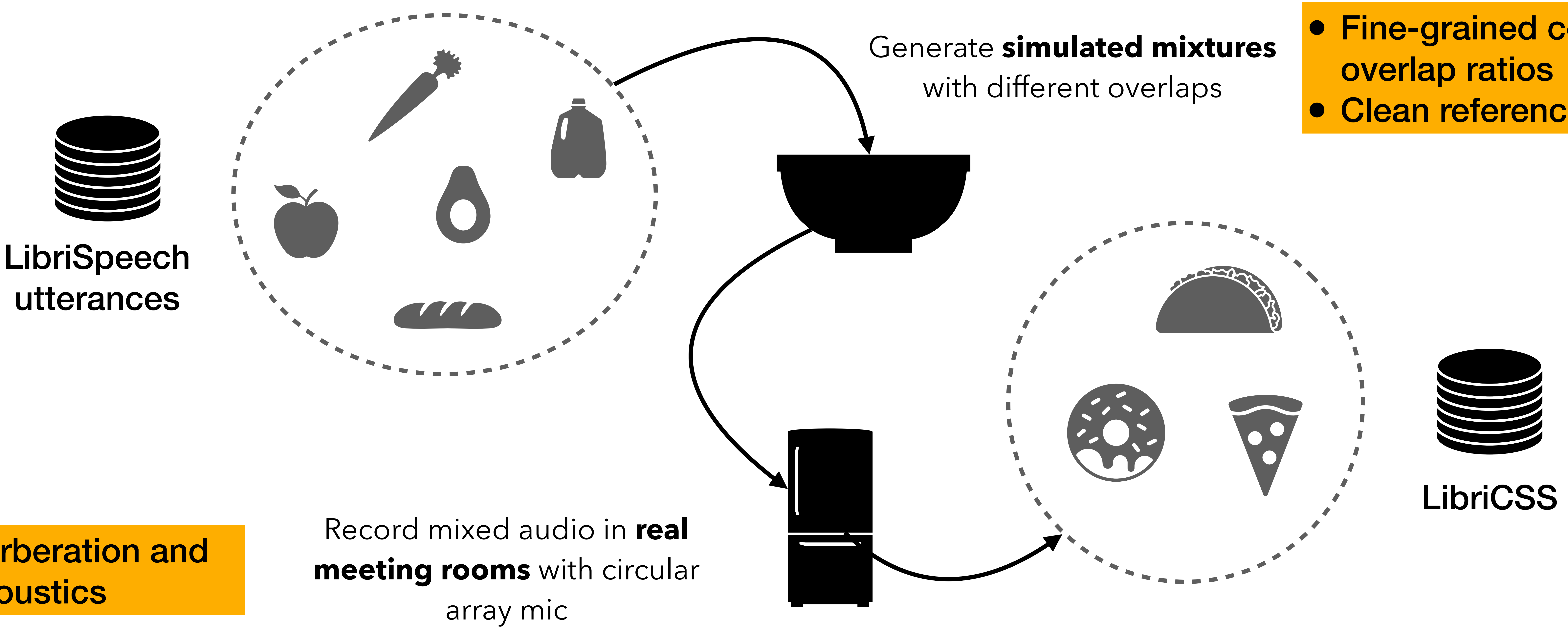
“Real recordings of simulated conversations”



Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” ICASSP 2015.

More about LibriCSS

Why is it useful?



- Fine-grained control on overlap ratios
- Clean references available

Real reverberation and acoustics

More about cpWER

Metric for "who spoke what"

cpWER = concatenated minimum-permutation word error rate

Concatenate all utterances of a speaker in reference and hypothesis

Score all pairs of reference and hypothesis speakers

Find permutation that **minimizes the total WER**
(linear sum assignment)

Speech separation results

SDR is not related to cpWER results

Separation → Spectral clustering → Hybrid TDNNF

		SDR	DER	cpWER
No separation		-	18.3	31.0
Mask-based MVDR	2.4s chunks; 2 streams	5.8	**13.9**	22.8
Sequential multi-frame	10s chunks; 3 streams	**14.1**	14.1	**19.3**

Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, and Fil Alleva, "Multi-microphone neural speech separation for farfield multi-talker speech recognition," ICASSP 2018

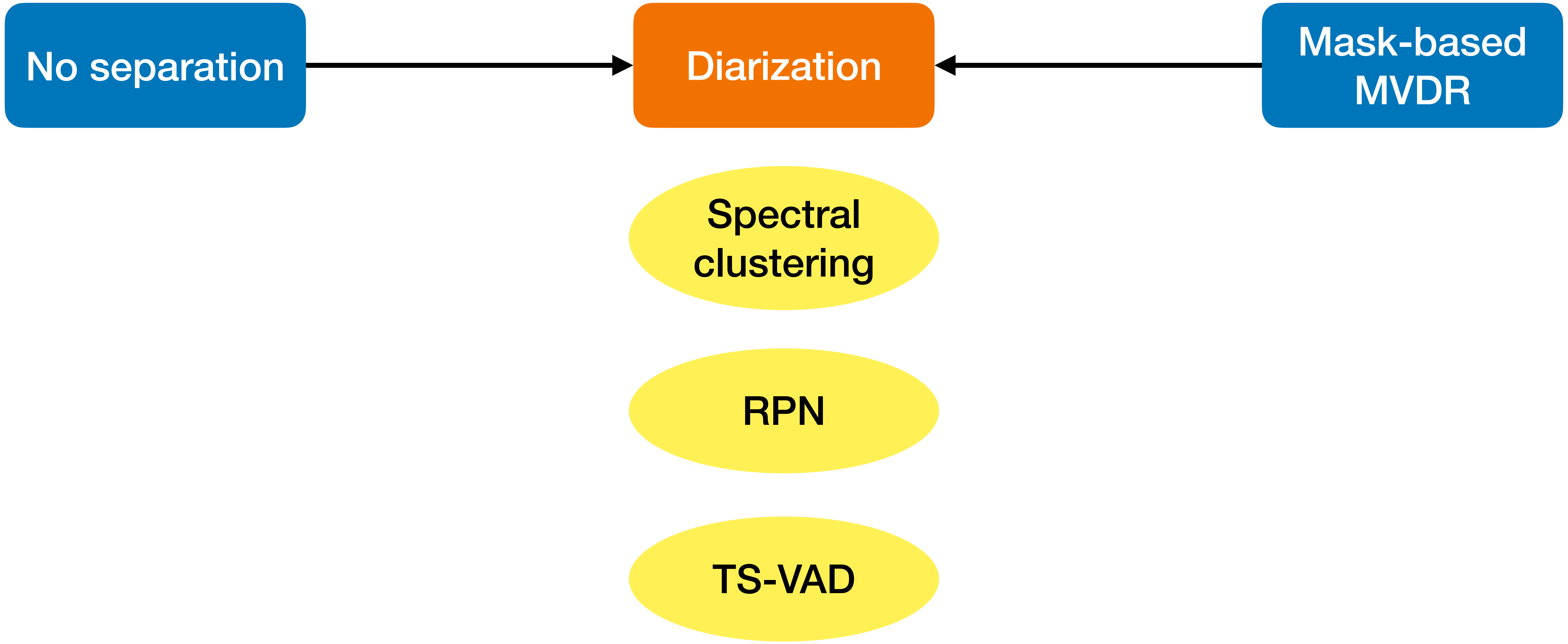
Zhong-Qiu Wang, Hakan Erdogan, Scott Wisdom, Kevin Wilson, Desh Raj, Shinji Watanabe, Zhuo Chen, and John R. Hershey, "Sequential multi-frame neural beamforming for speech separation and enhancement," IEEE SLT 2021.

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Diarization results

Clustering-based vs. supervised methods



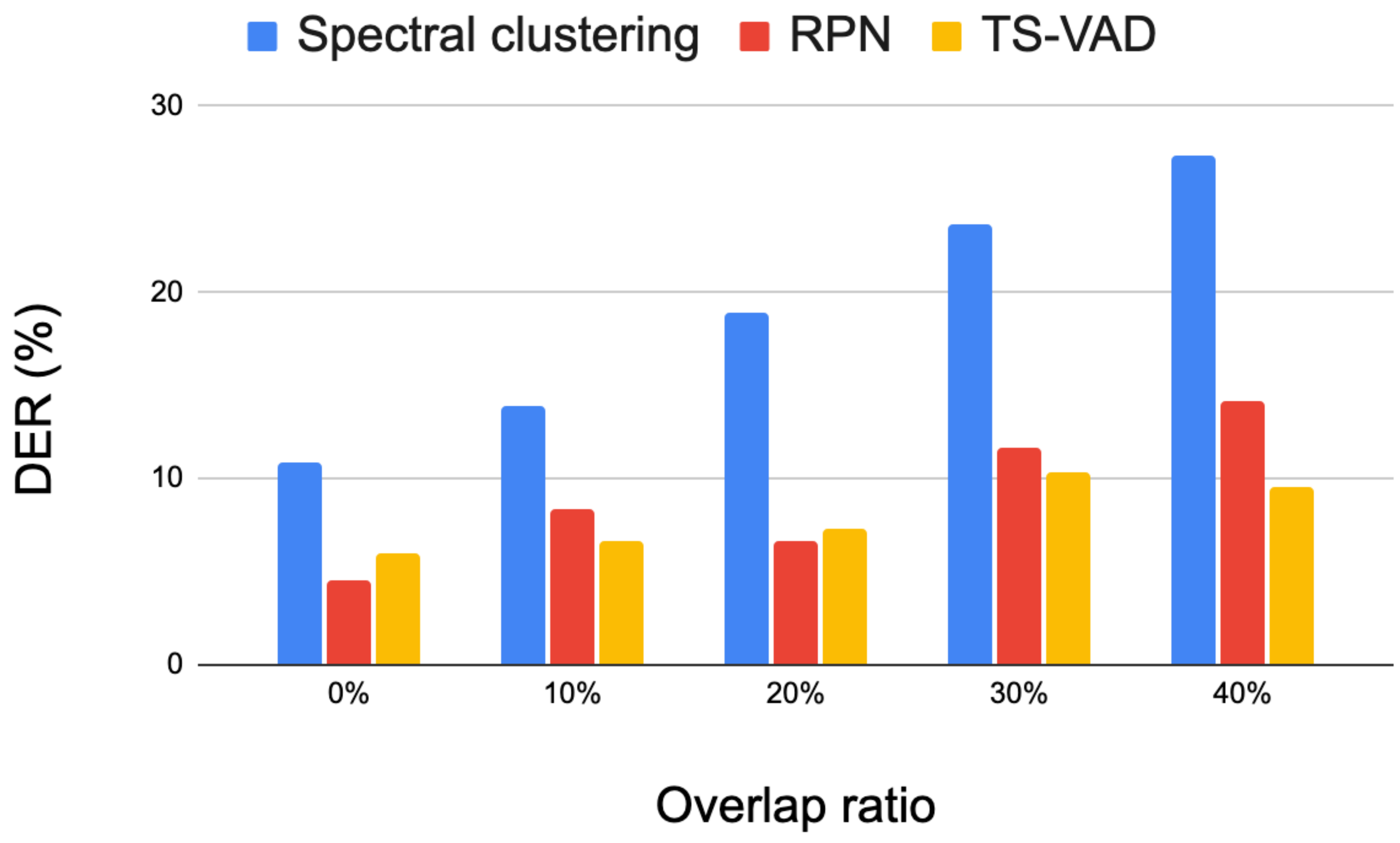
Park et al., “Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap,” IEEE Signal Processing Letters, 2020.

Huang et al., “Speaker diarization with region proposal network,” ICASSP 2020.

Medennikov, et al., “Target speaker voice activity detection: a novel approach for multispeaker diarization in a dinner party scenario,” Interspeech 2020.

Diarization results

Dichotomy between performance on mixed and separated audio



ASR results

Hybrid TDNNF and End-to-end Transformer models

Mask-based MVDR → Spectral clustering → ASR

KALDI **TDNNF**

ESPnet **Transformer**

Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, “Semiorthogonal low-rank matrix factorization for deep neural networks,” Interspeech 2018.

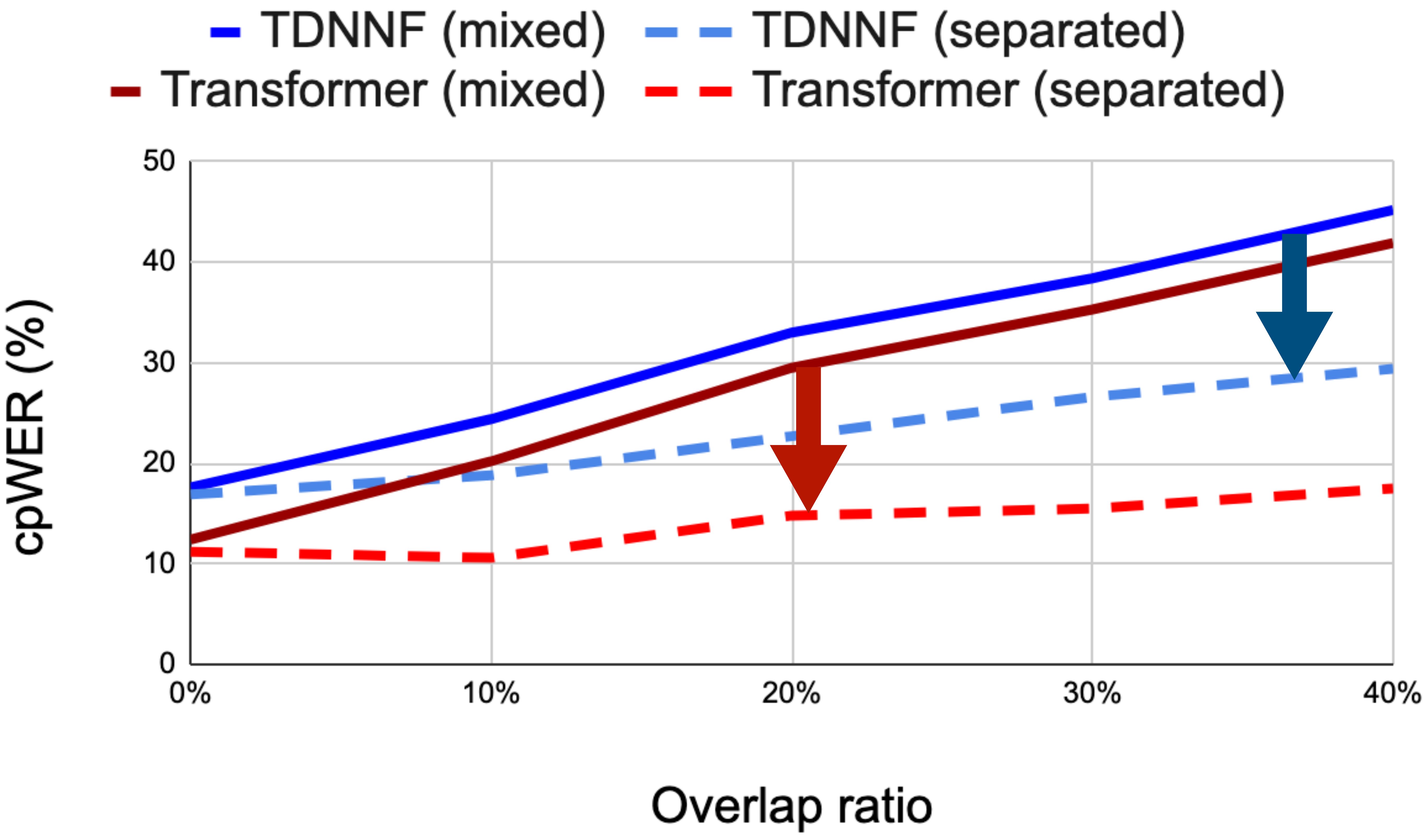
Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al., “A comparative study on transformer vs RNN in speech applications,” IEEE ASRU 2019.

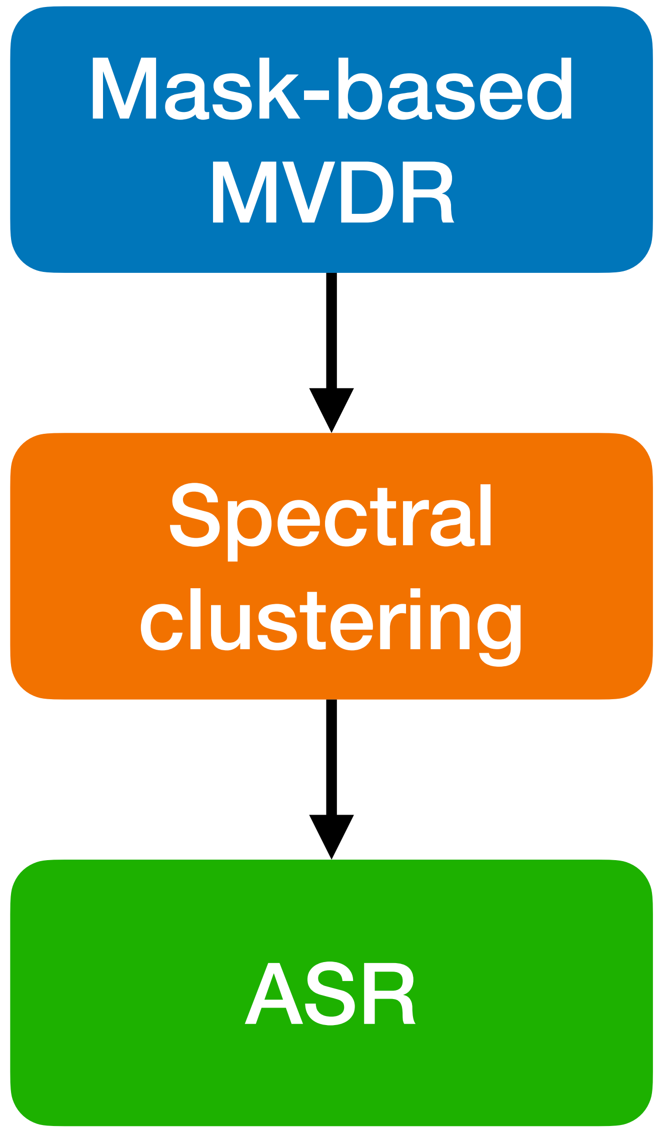
ASR results

Performance on clean and separated audio are correlated

Performance on LibriSpeech:

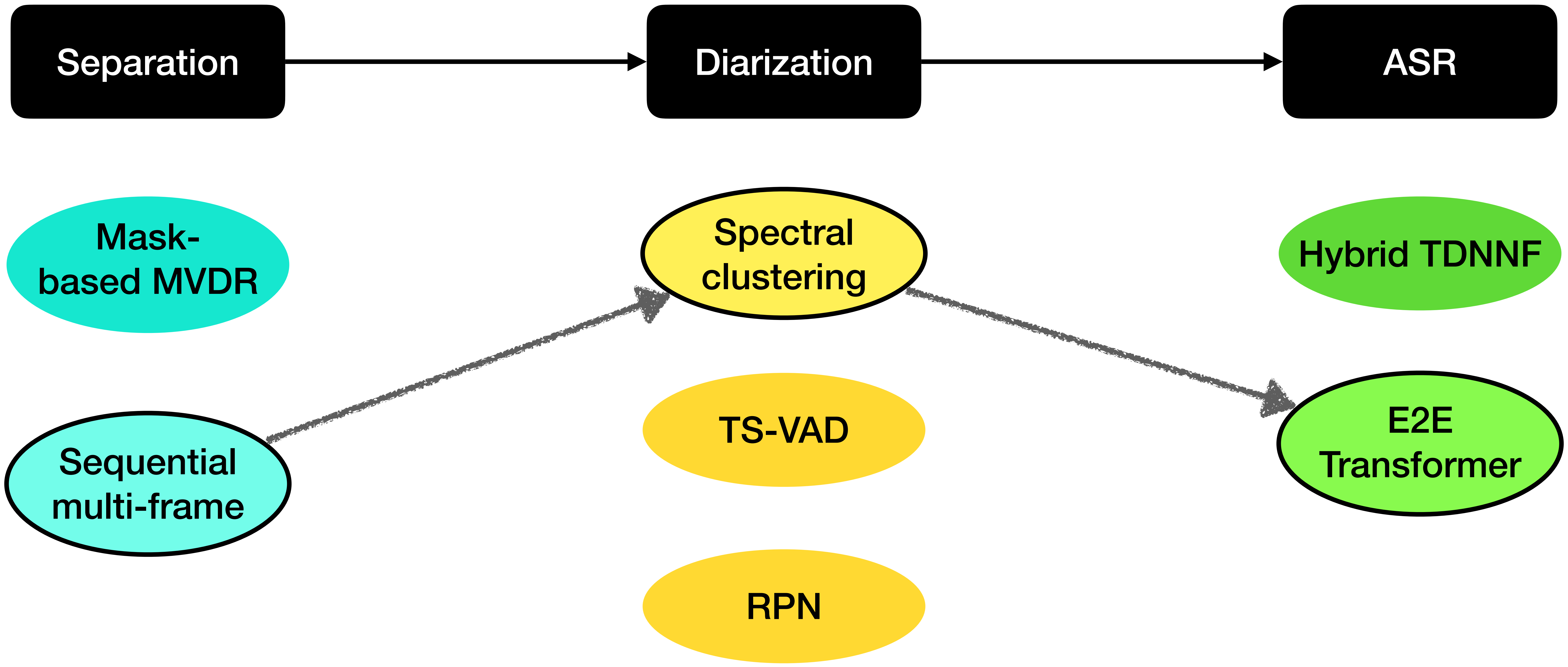
TDNNF	**3.8%**
Transformer	**2.2%**





Our final pipeline

Final cpWER = 12.7% (compared with 27.1% for “no separation” baseline)



cpWER = concatenated minimum-permutation word error rate

How to use this research?

Plug in your component in the pipeline to get cpWER results

Separation → Diarization → ASR

Separation	Diarization	ASR
✓ Mask-based MVDR	✓ Clustering	✓ Hybrid TDNNF
✓ Sequential multi-frame	TS-VAD	✓ E2E Transformer
	RPN	

KALDI

ESPnet

Separated audio and full recipes available!

How to use this research?

Details available on project page

Scan me!

Acknowledgments:

The work reported here was started at JSALT 2020 at JHU, with support from Microsoft, Amazon, and Google.

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING