The JHU Multi-Microphone Multi-Speaker ASR System for the CHiME-6 Challenge

*Ashish Arora, *Desh Raj, *Aswin Shanmugam Subramanian, *Ke Li, Bar Ben-Yair, Matthew Maciejewski, Piotr Żelasko, Paola Garcia, Shinji Watanabe, Sanjeev Khudanpur

Challenge Overview

Track 1

- ASR only
- Oracle speaker segments provided

Track 2

- Diarization + ASR
- No speaker segments

Challenge Overview

Track 1

- ASR only
- Oracle **speaker segments provided**

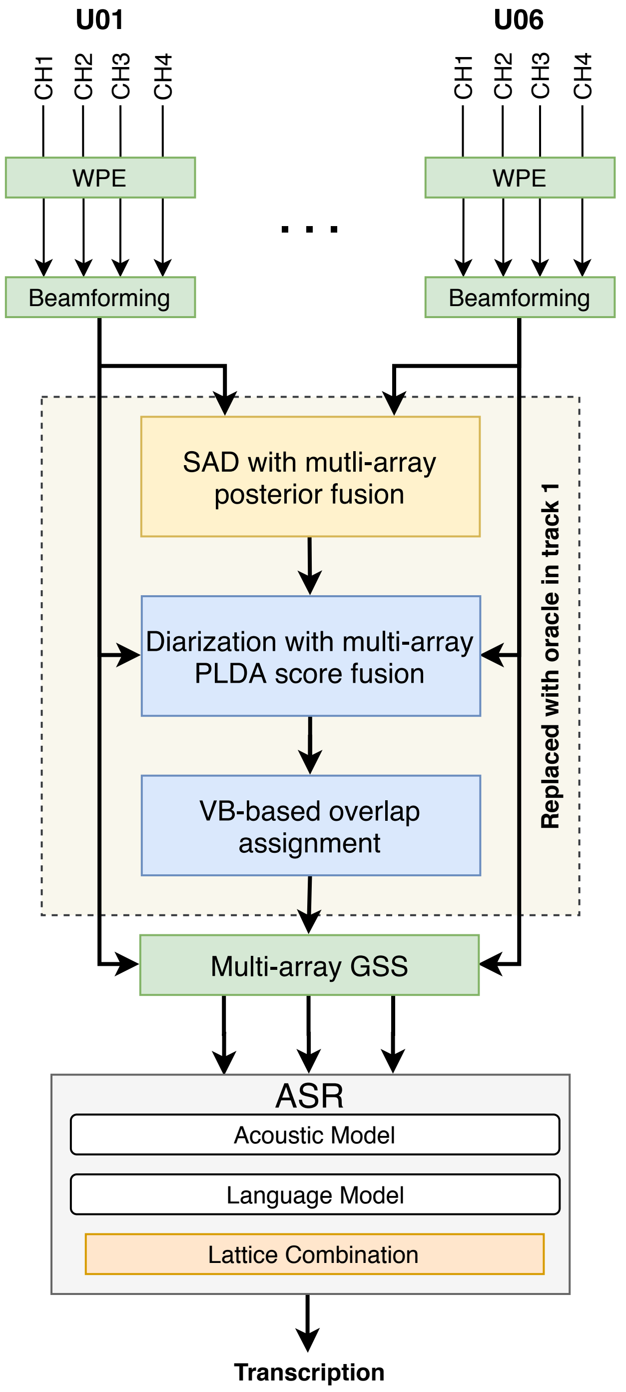
Track 2

- Diarization + ASR
- No speaker segments

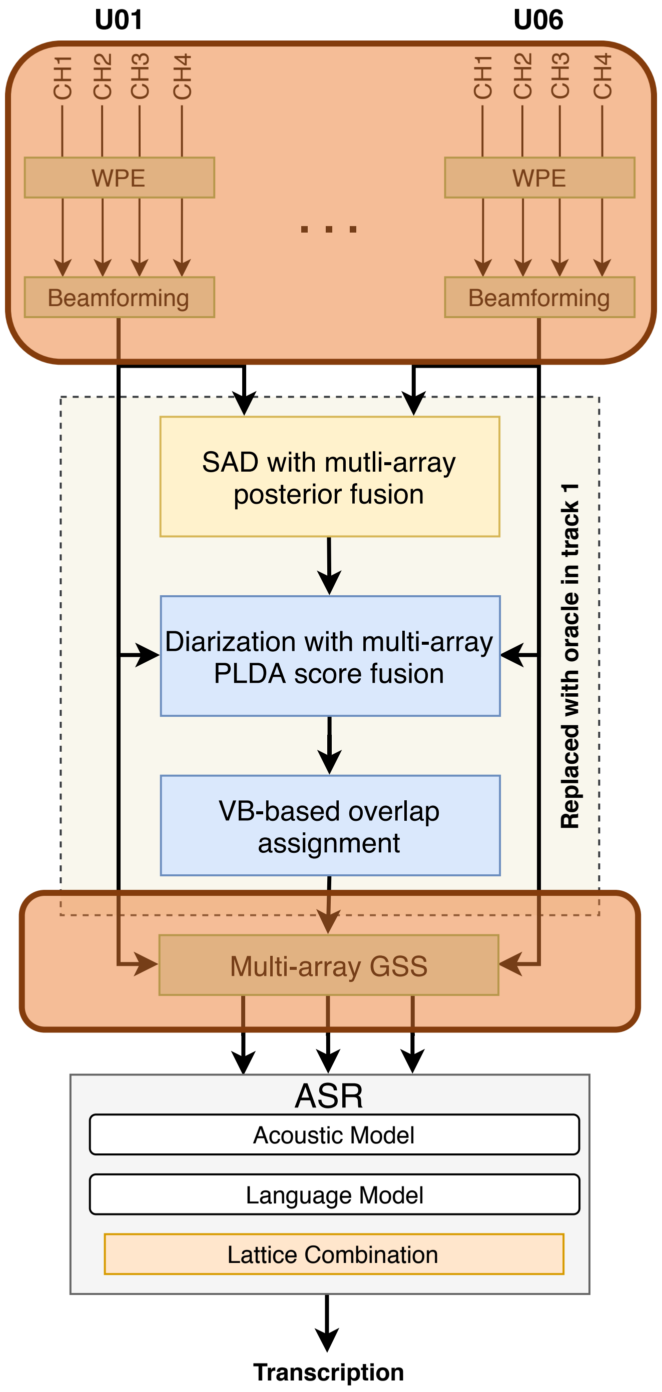
We will present our Track-2 system.

Same ASR model used in both tracks.

Our Track-2 Pipeline



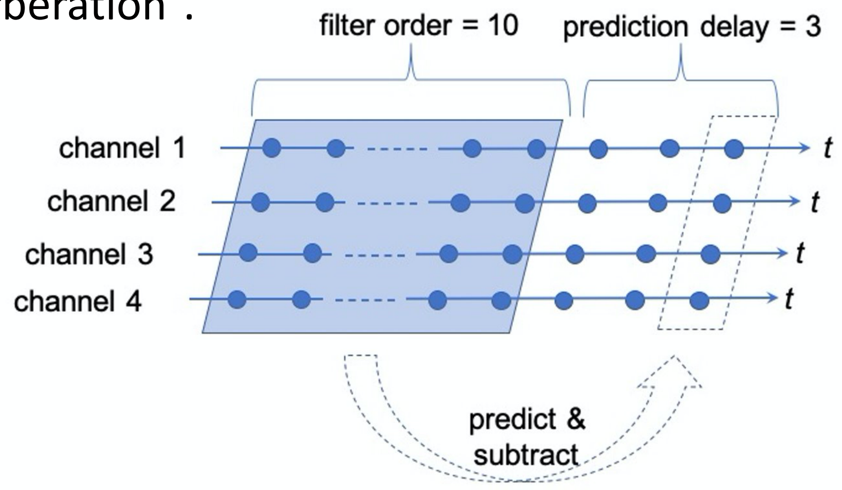
Speech Enhancement



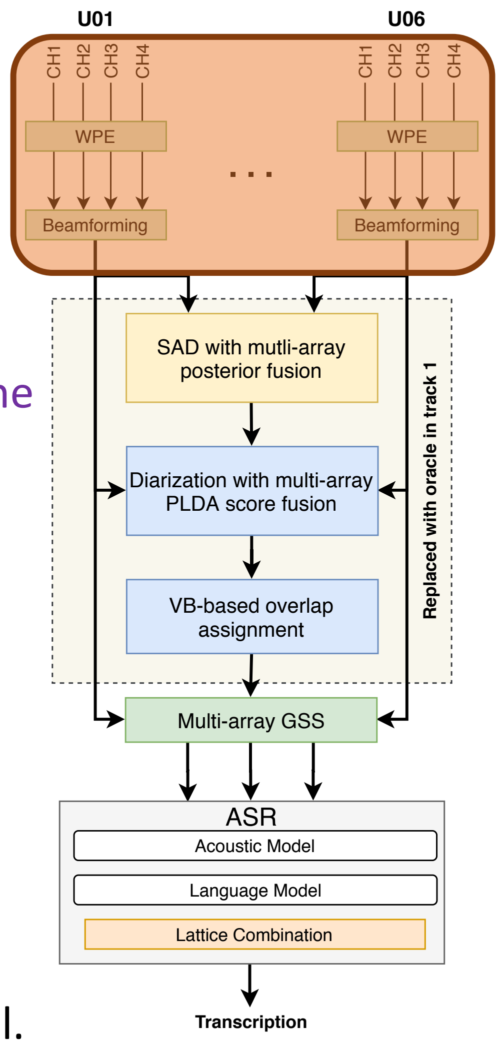
Speech Enhancement

WPE & BeamformIt

- The first preprocessing step performed was weighted prediction error (WPE*) based online **multi-channel** dereverberation^.



- Subsequently a delay and sum beamformer (BeamformIt+) was used to denoise the signal.



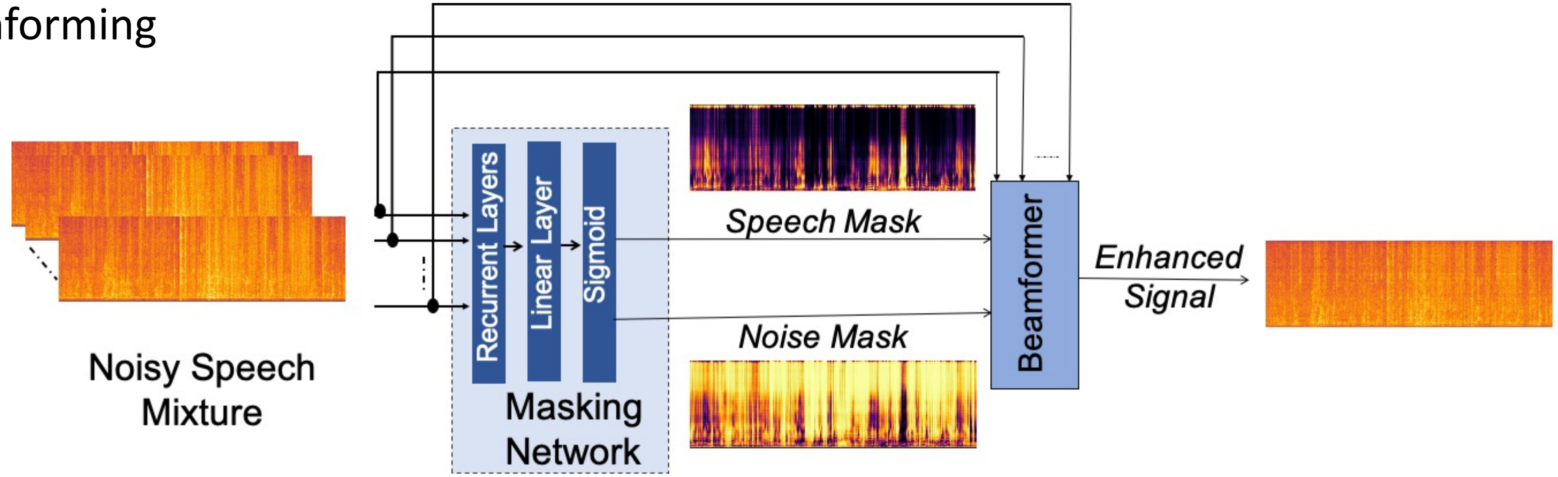
* Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang. "Speech dereverberation based on variance-normalized delayed linear prediction", *IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 7, pp. 1717-1731, Sep. 2010*.

^ L. Drude, J. Heymann, C. Boeddeker and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," *13th ITG-Symposium, 2018*.

\+ Xavier Anguera, Chuck Wooters and Javier Hernando. " Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, pp. 2011-2023, Sep. 2007*.

Denoising Alternative

Neural Beamforming



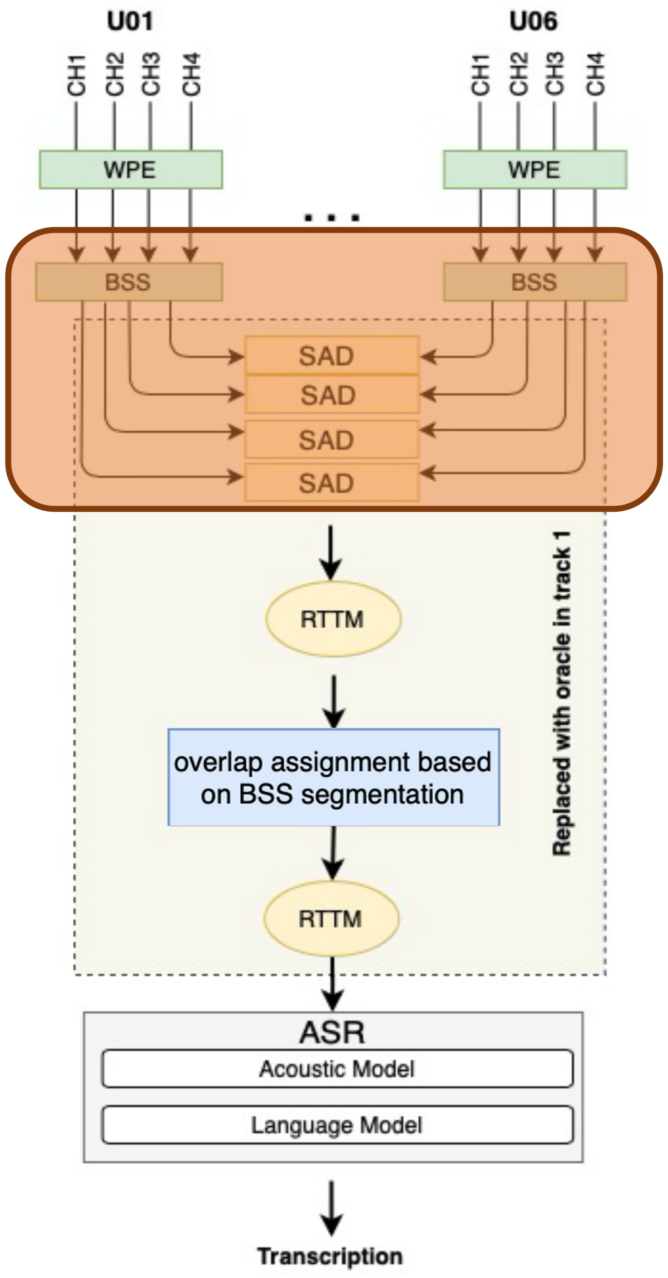
Enhancement	Array	Dev WER (%)	Eval WER (%)
BeamformIt	U06	76.0	72.6
Neural Beamforming	U06	76.2	73.6

- Trained a neural network to separate noise from speech mixtures.
- Used voxceleb data for simulation with CHiME-6 noises (mixtures of 1-4 speakers).
- Perceptually seemed better in very noisy segments but didn't help final performance. So it was not used in our final system.

Other Ideas We Tried

BSS & TaSNet

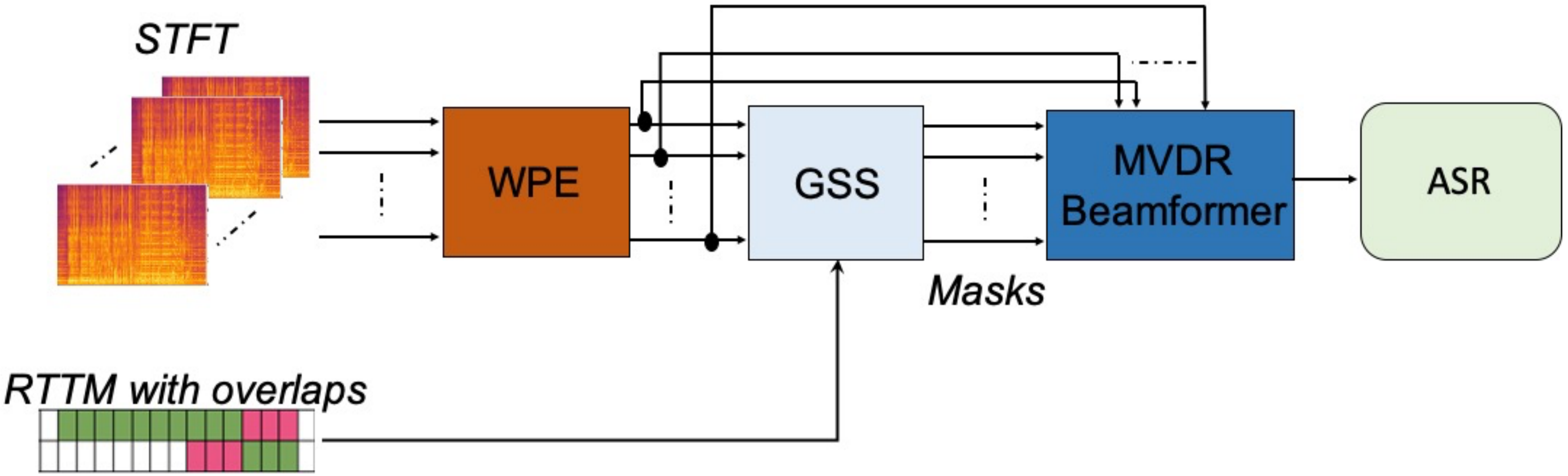
- If blind speech separation (BSS) can be performed before speaker segmentation we won’t require speaker diarization.
- We tried two methods: (1) independent vector analysis (IVA)* and (2) TaSNet+
- We were not able to obtain good results with this approach, so it was not used in our final system.

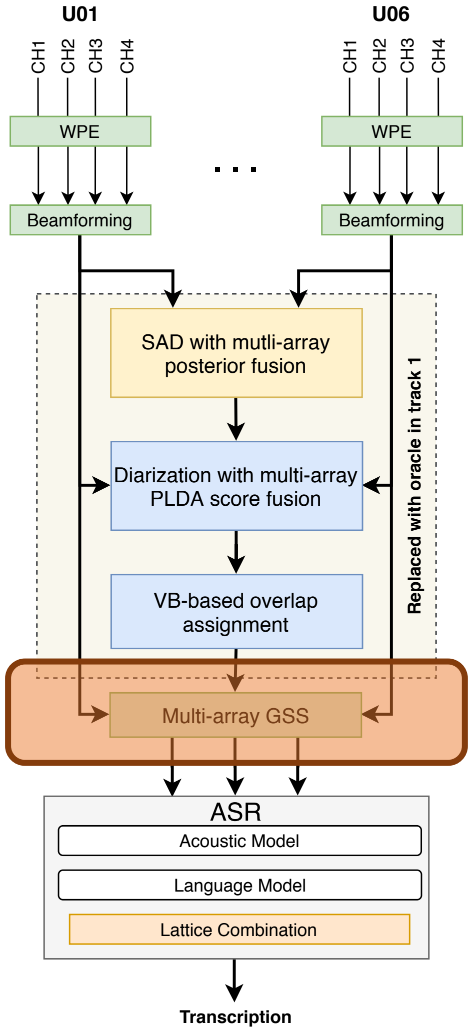


* R. Scheibler, E. Bezzam and I. Dokmanić. ”Pyroomacoustics: A Python package for audio room simulations and array processing algorithms”, *ICASSP 2018*.
+ Y. Luo and N. Mesgarani. "TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," *ICASSP 2018.*

Speech Separation

Multi-Array GSS*





- Guided source separation (GSS) was used to separate the target source using the time annotations.
- The groundtruth annotations were used for Track 1 and diarization outputs were used to obtain the time annotations for Track 2.

* Naoyuki Kanda, Christoph Boeddeker, Jens Heitkaemper, Yusuke Fujita, Shota Horiguchi, Kenji Nagamatsu, and Reinhold Haeb-Umbach. "Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party ASR." *Interspeech 2019.*

Importance of Overlap Assignment for GSS – Track 2

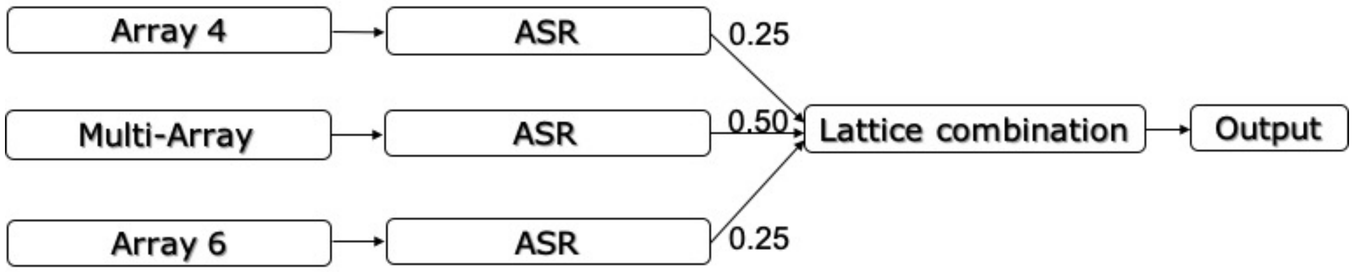
Multi-Array GSS - Sensitivity to Diarization Output

Method	Overlap Detection	Dev WER (%)	Eval WER (%)
Multi-Array GSS	N	71.0	68.8
Multi-Array GSS	Y	69.3	68.8

- GSS is very sensitive to the diarization output.
- VB-HMM based overlap detection helps GSS

Combining Early & Late Fusion – Track 2

Multi-View Decoding with GSS

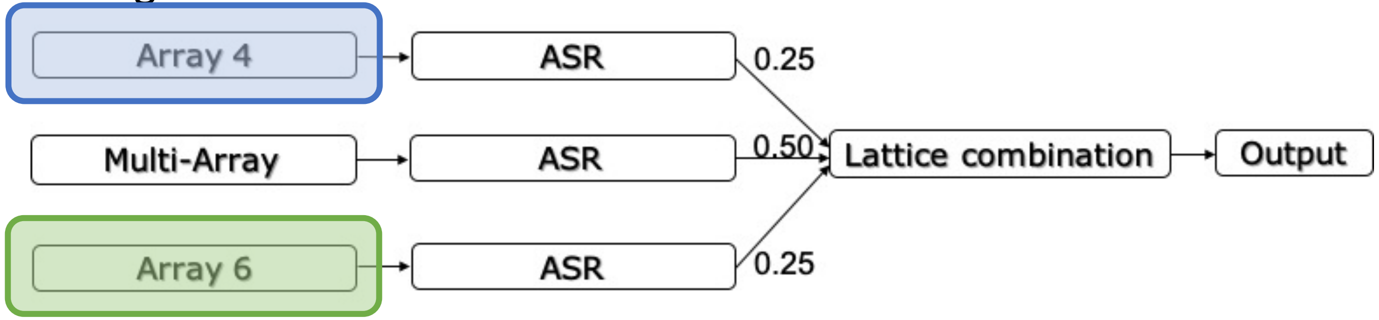


Separation Method	Early Fusion	Late Fusion	Dev WER (%)	Eval WER (%)
Multi-Array GSS	Y	N	69.3	68.8
Single-Array GSS (U06)	N	N	73.1	72.1
Single-Array GSS (U04)	N	N	72.3	74.4
Multi-Array GSS + Single-Array GSS	Y	Y	68.3	68.3

- Early fusion was performed by incorporating all arrays while beamforming.
- Late fusion was performed by lattice combination of GSS output on multi-array with individual arrays.

Combining Early & Late Fusion – Track 2

Multi-View Decoding with GSS

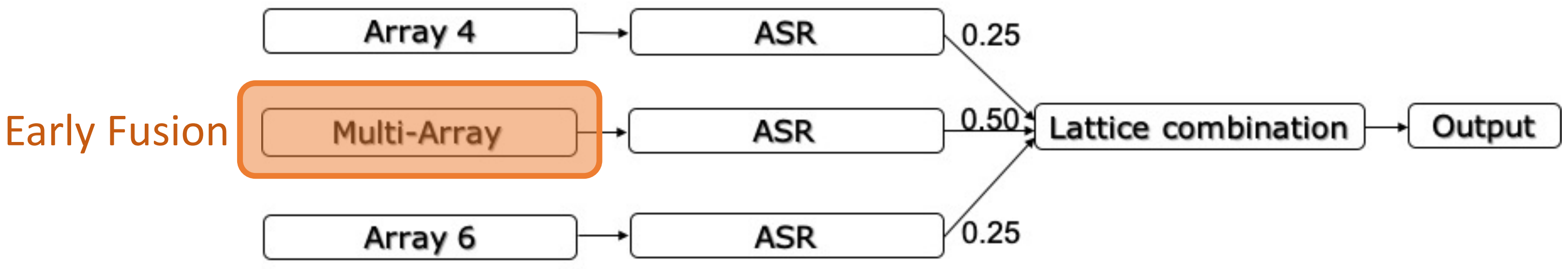


Separation Method	Early Fusion	Late Fusion	Dev WER (%)	Eval WER (%)
Multi-Array GSS	Y	N	69.3	68.8
Single-Array GSS (U06)	N	N	73.1	72.1
Single-Array GSS (U04)	N	N	72.3	74.4
Multi-Array GSS + Single-Array GSS	Y	Y	68.3	68.3

- Early fusion was performed by incorporating all arrays while beamforming.
- Late fusion was performed by lattice combination of GSS output on multi-array with individual arrays.

Combining Early & Late Fusion – Track 2

Multi-View Decoding with GSS

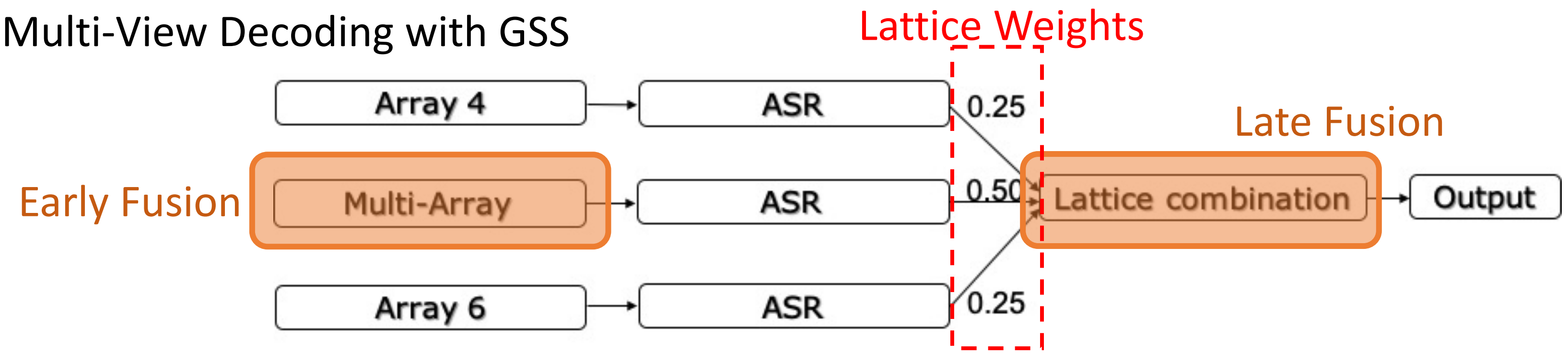


Separation Method	Early Fusion	Late Fusion	Dev WER (%)	Eval WER (%)
Multi-Array GSS	Y	N	69.3	68.8
Single-Array GSS (U06)	N	N	73.1	72.1
Single-Array GSS (U04)	N	N	72.3	74.4
Multi-Array GSS + Single-Array GSS	Y	Y	68.3	68.3

- Early fusion was performed by incorporating all arrays while beamforming.
- Late fusion was performed by lattice combination of GSS output on multi-array with individual arrays.

Combining Early & Late Fusion – Track 2

Multi-View Decoding with GSS



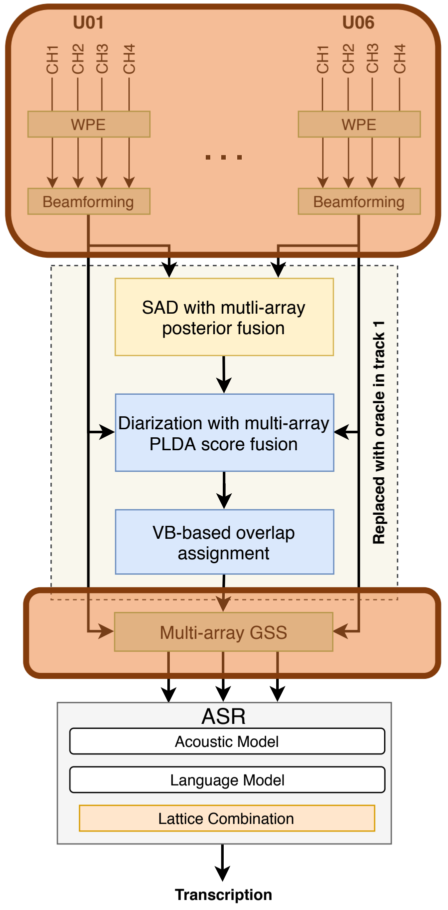
Separation Method	Early Fusion	Late Fusion	Dev WER (%)	Eval WER (%)
Multi-Array GSS	Y	N	69.3	68.8
Single-Array GSS (U06)	N	N	73.1	72.1
Single-Array GSS (U04)	N	N	72.3	74.4
Multi-Array GSS + Single-Array GSS	Y	Y	68.3	68.3

- Early fusion was performed by incorporating all arrays while beamforming.
- Late fusion was performed by lattice combination of GSS output on multi-array with individual arrays.

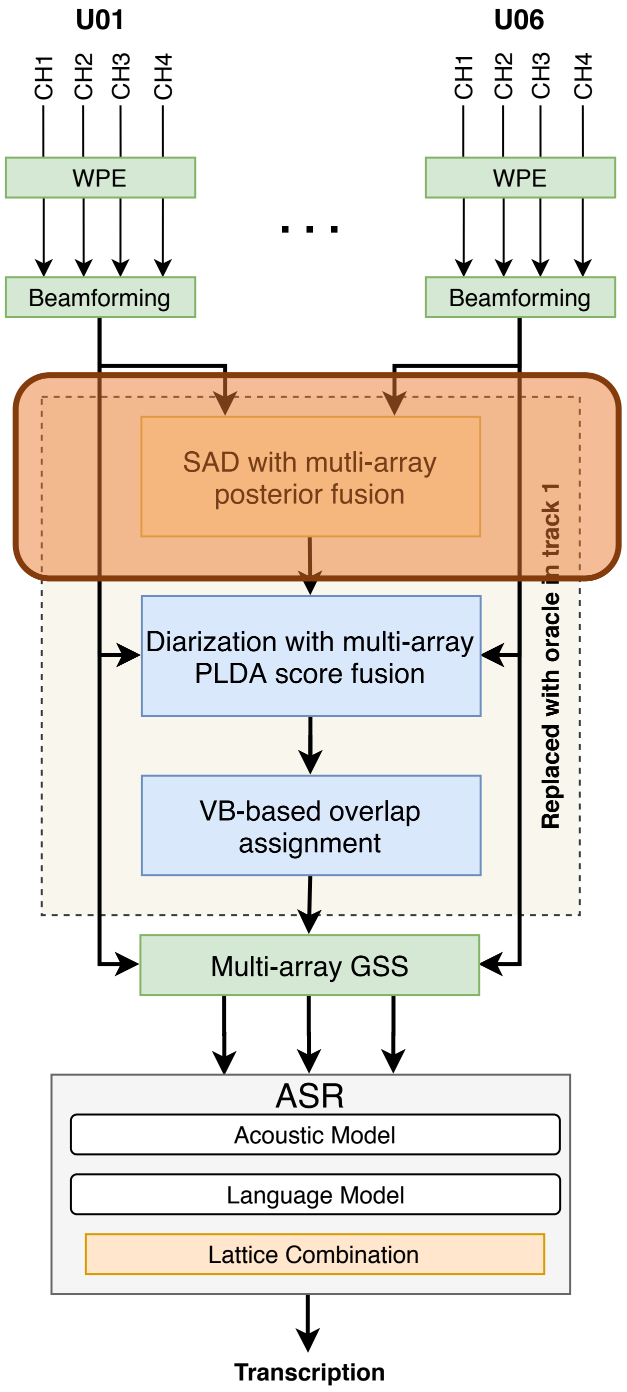
Speech Enhancement Takeaways

GSS gives good improvements and it is sensitive to the diarization output

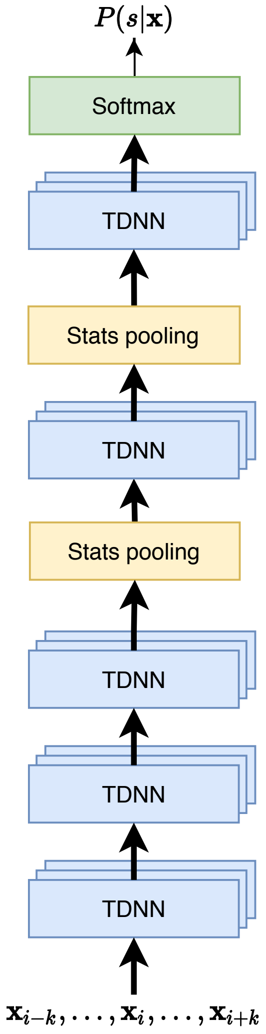
Further improvement can be obtained by combining early and late fusions



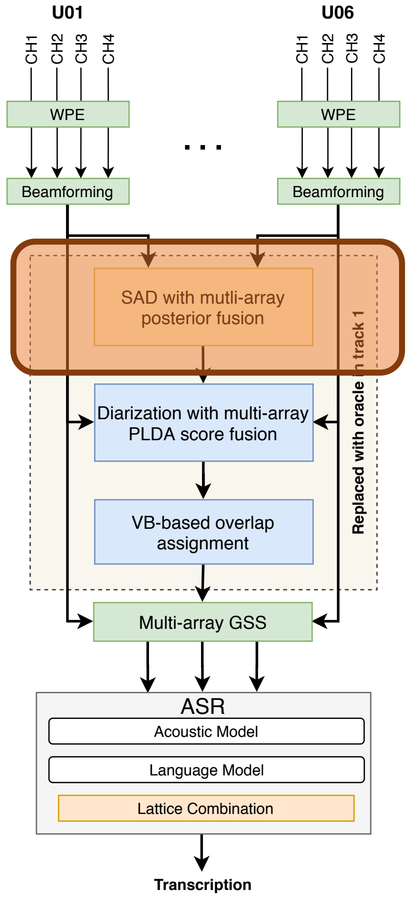
Speech Activity Detection



Speech Activity Detection

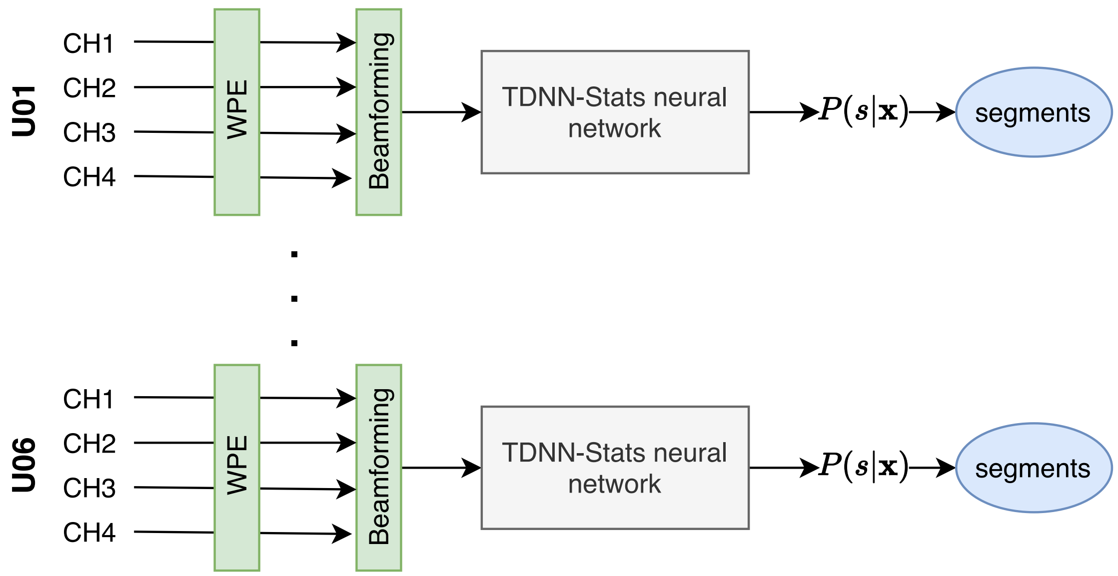


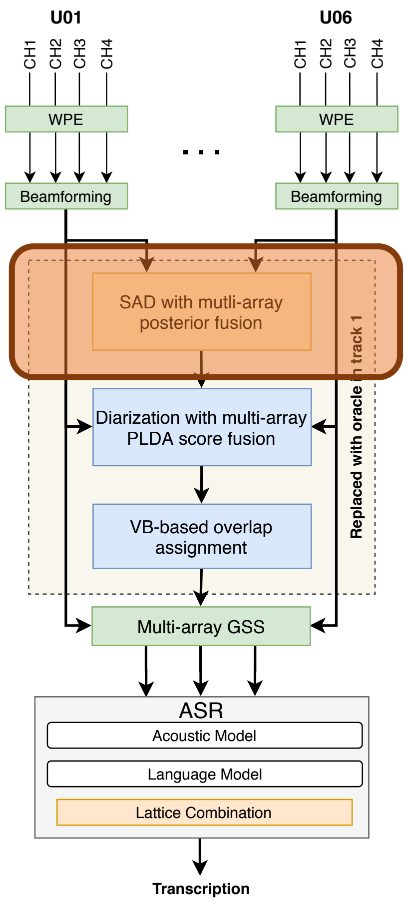
- We built a TDNN-Stats based architecture for baseline
- We use the same core model
- **NEW: Multi-array extension**



Baseline SAD

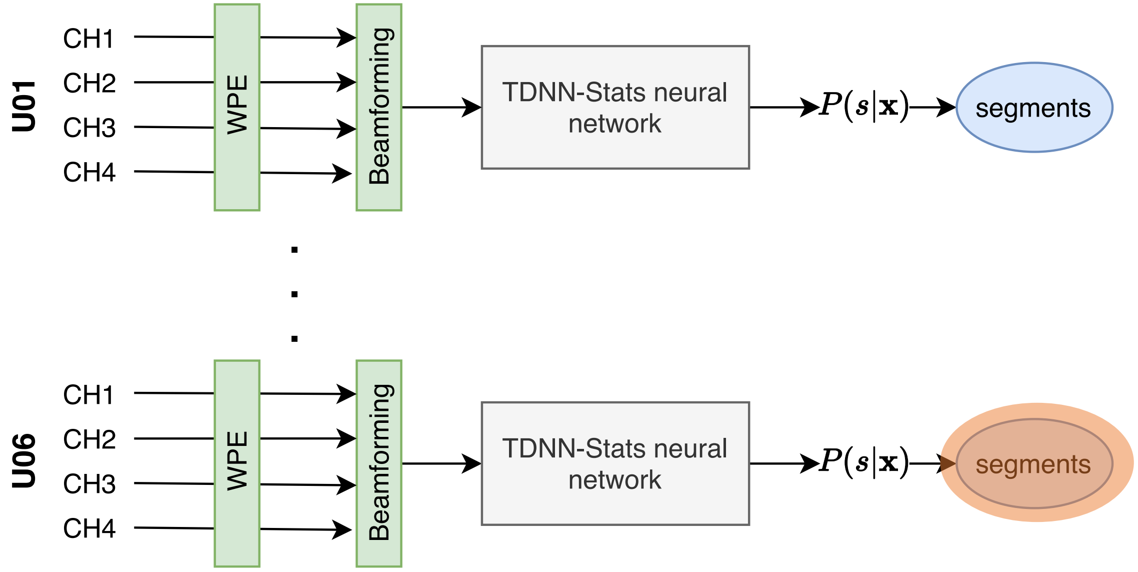
Beamforming over channels within array

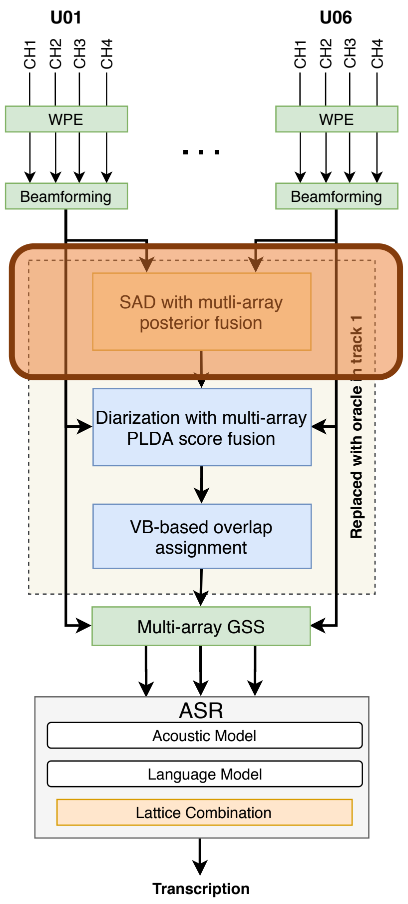




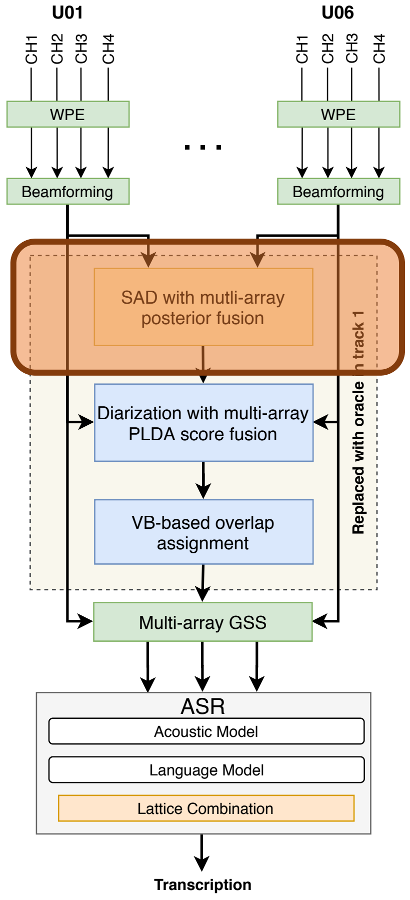
Baseline SAD

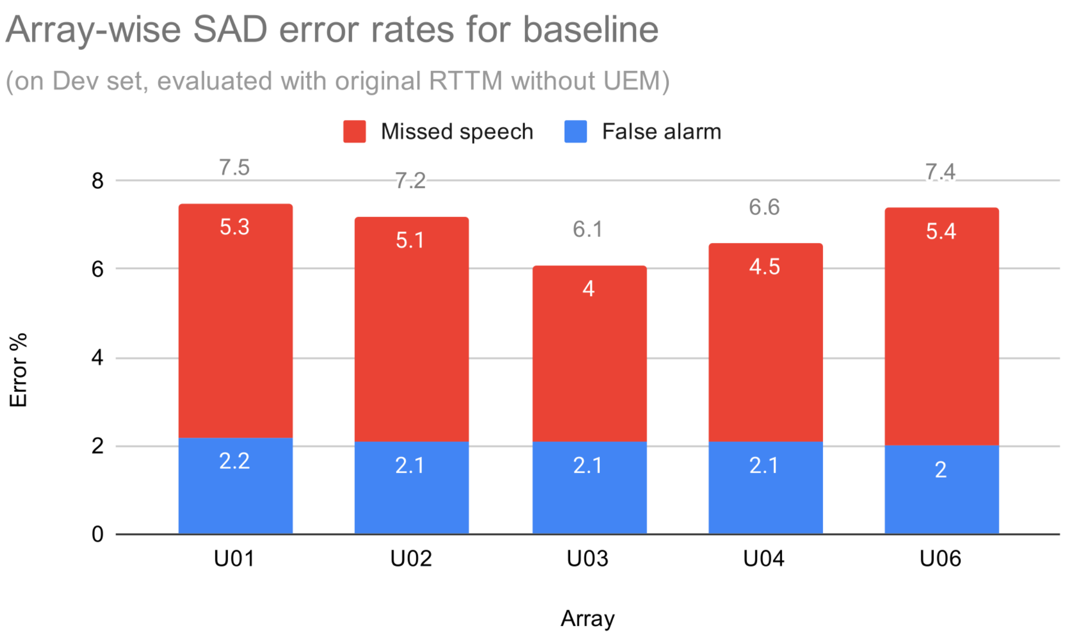
Random selection of array U06

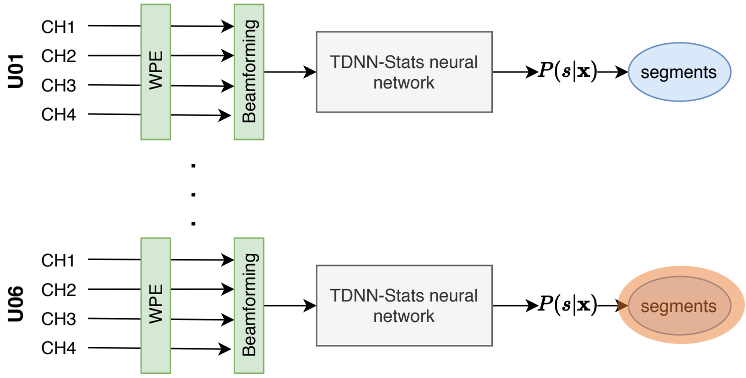




U06 is not the best selection



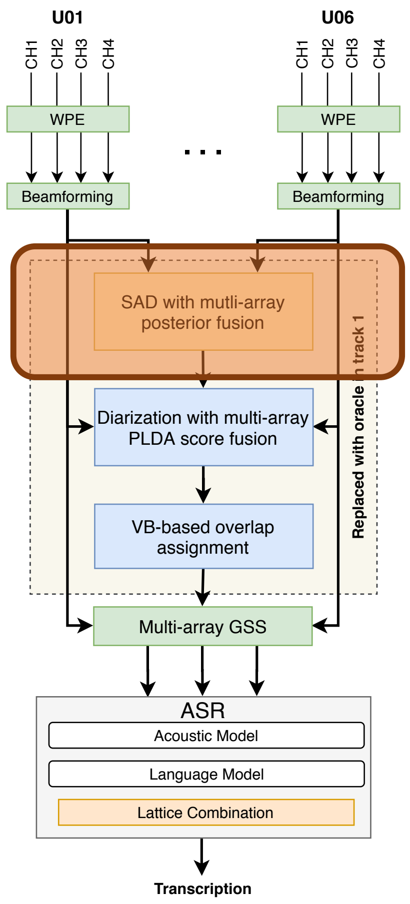


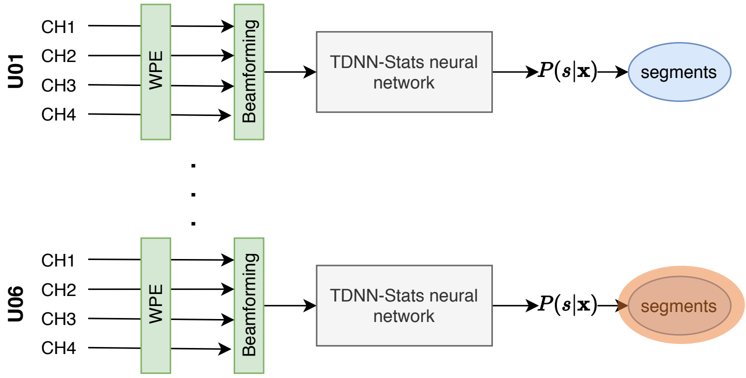


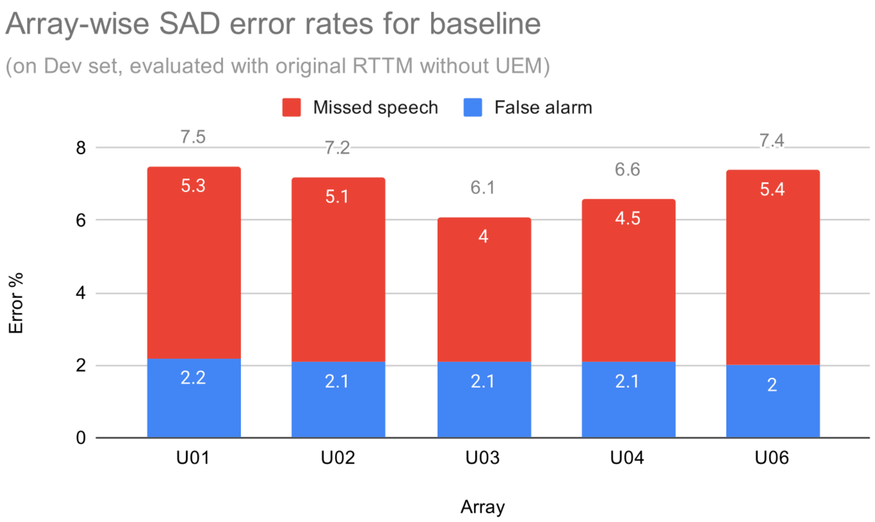
Multi-channel + multi-array

1. Can we do better than beamforming for **channel combination**?

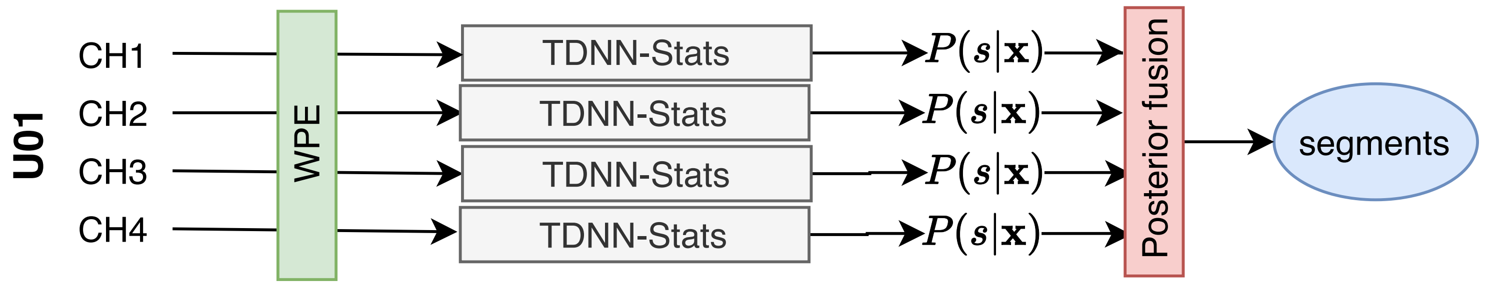
2. Can we do better than random **array selection**?

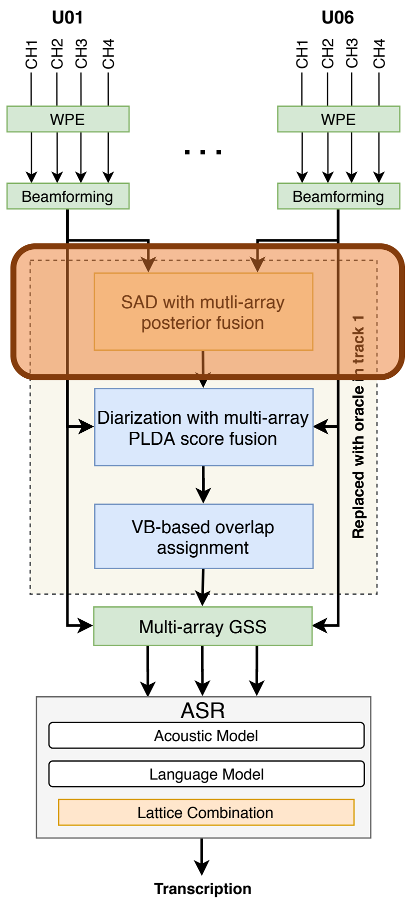






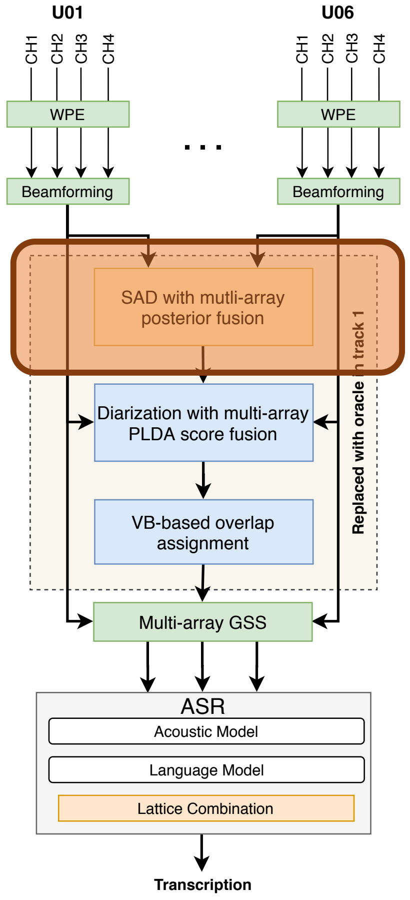
Can **channel-level posterior fusion** do better than beamforming?

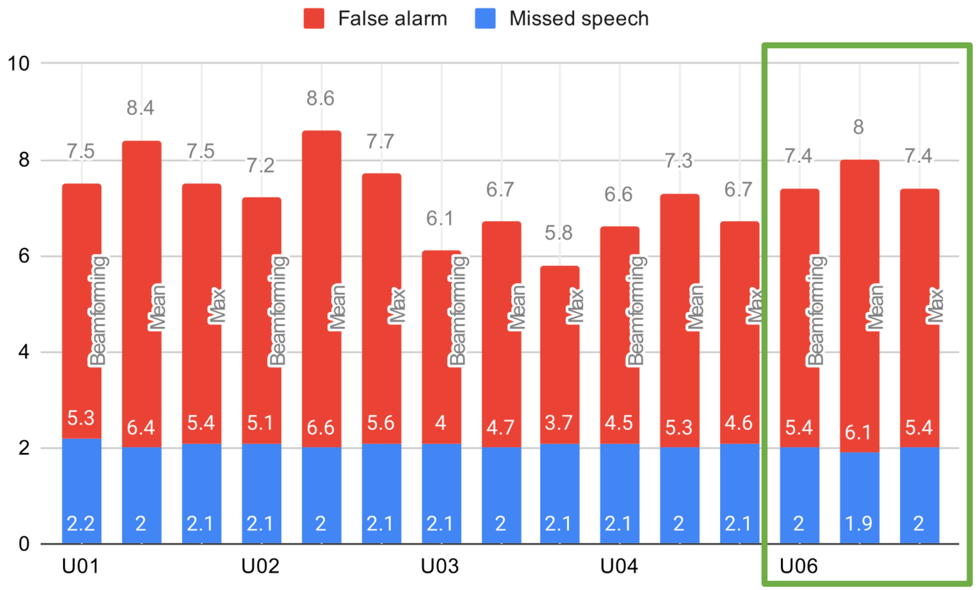


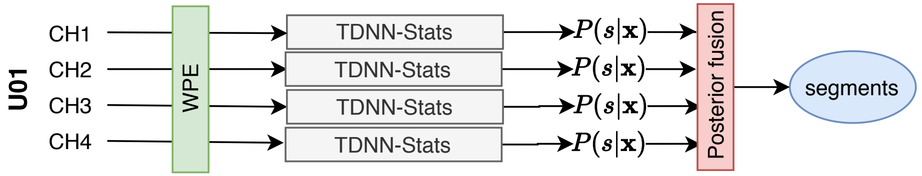


Posterior Mean < Posterior Max ≈ Beamforming

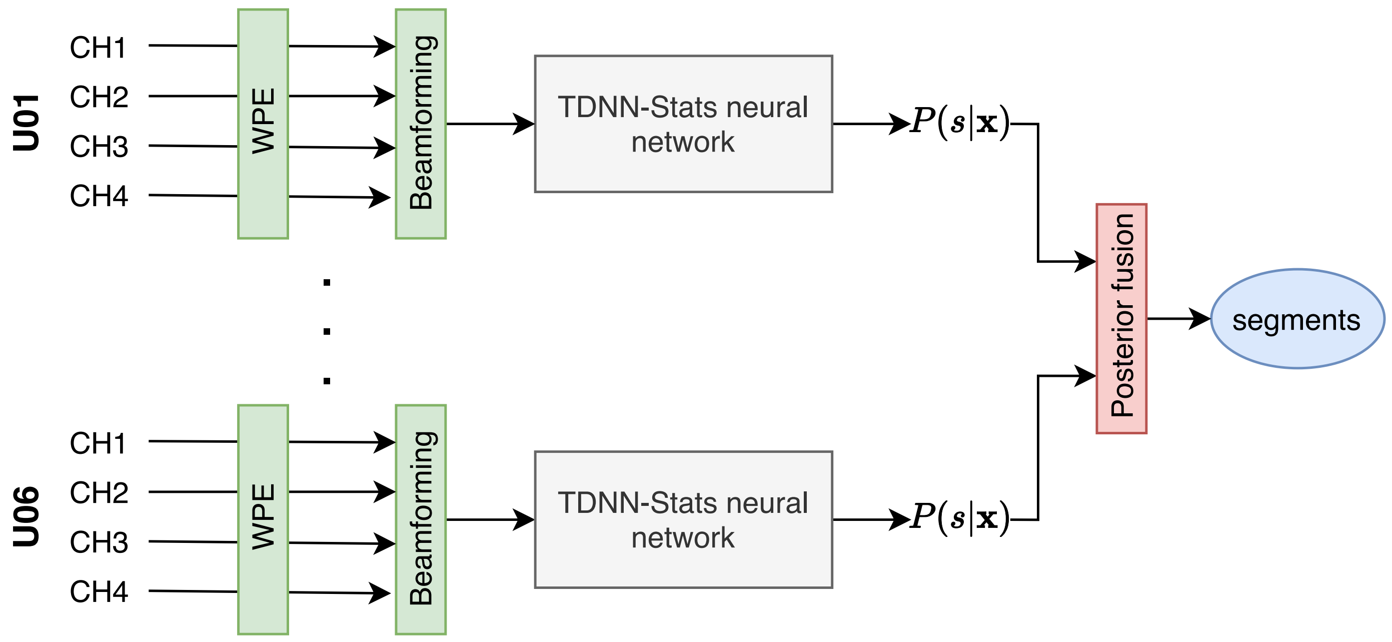
But 4x more compute -> keep Beamforming!

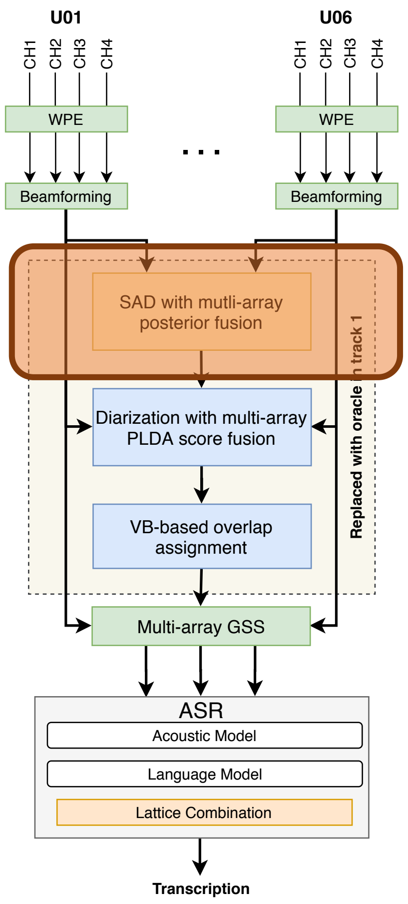




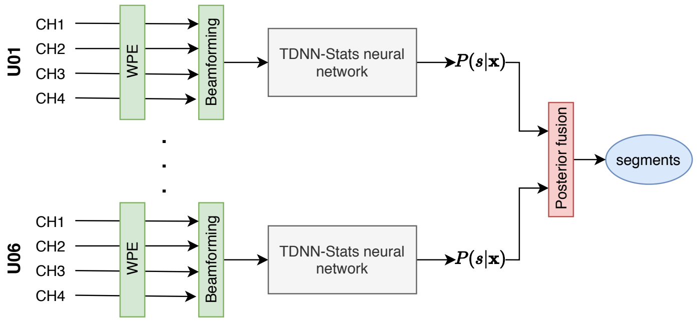


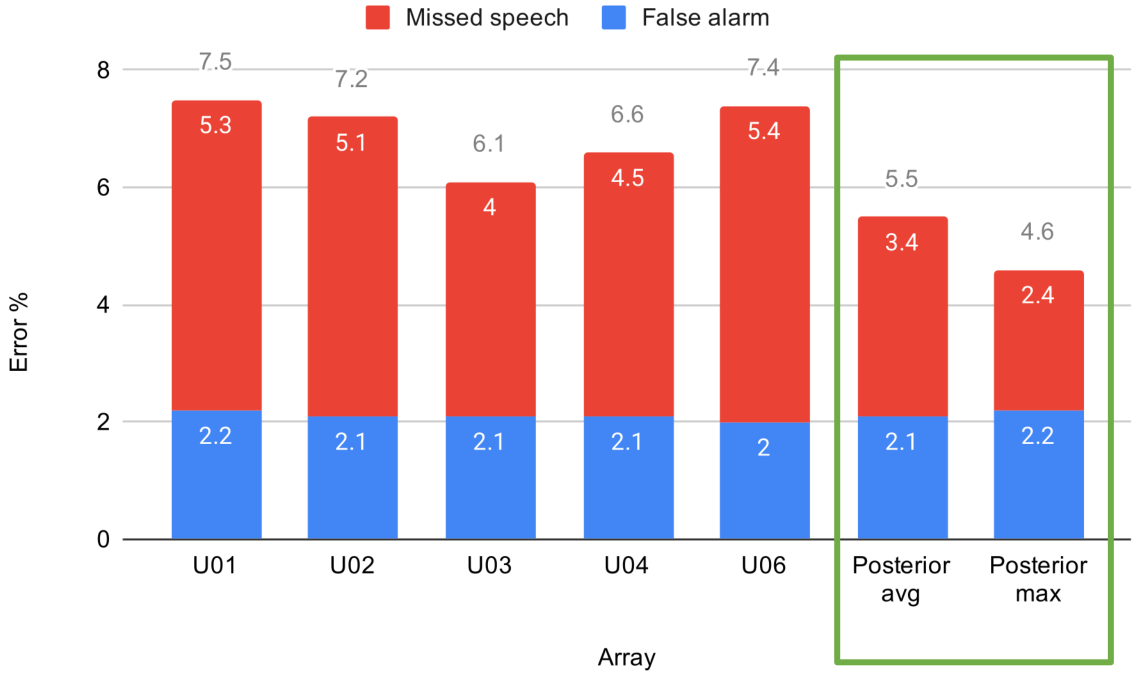
Posterior fusion for **array combination**

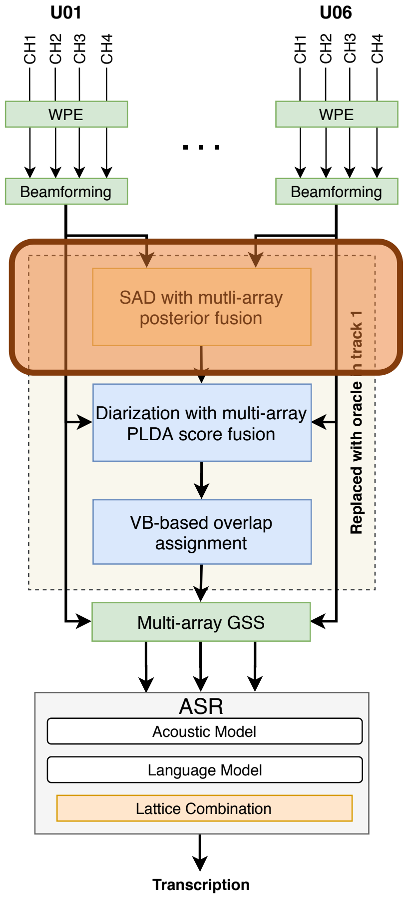




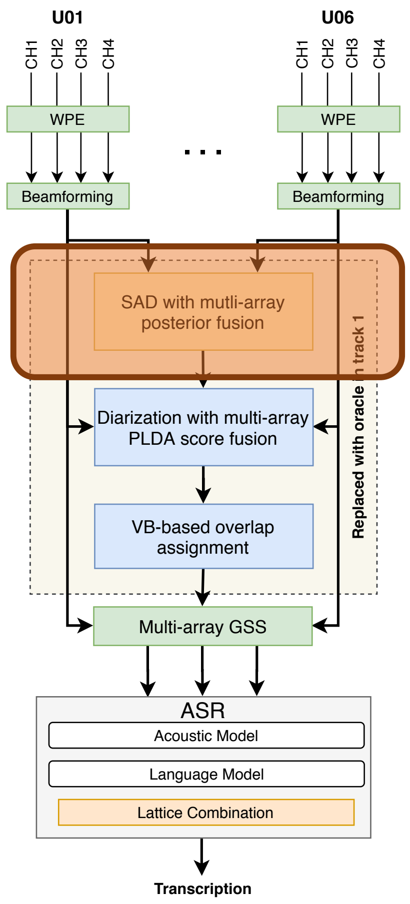
Posterior fusion helps array combination



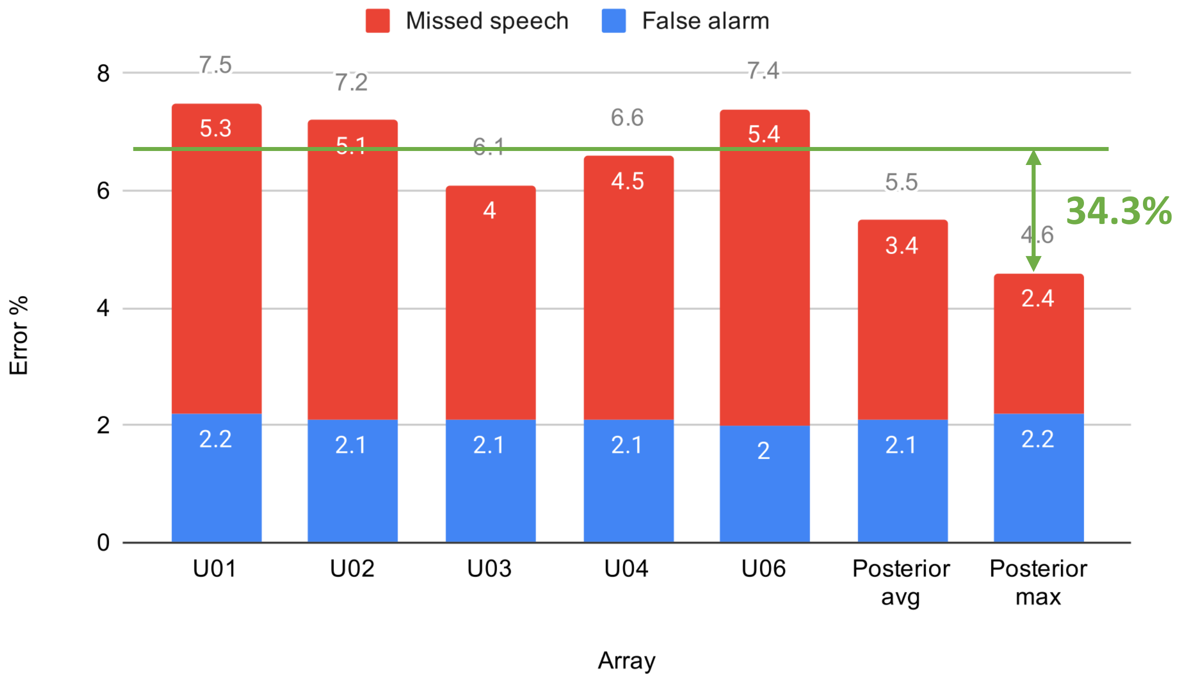


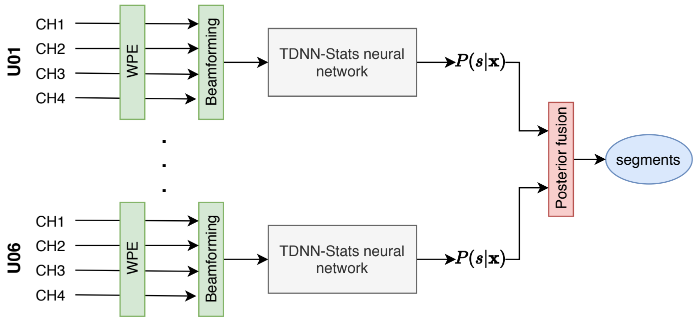


Posterior fusion helps array combination **significantly**

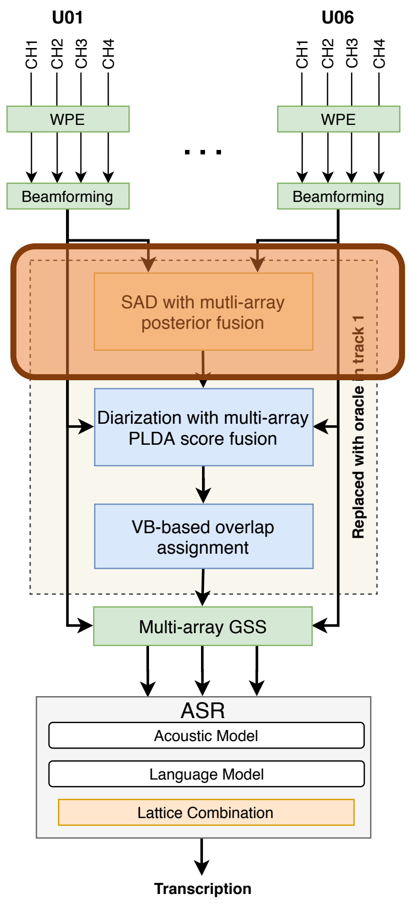


- Huge improvement in missed speech
- Directly affects downstream DER and WER





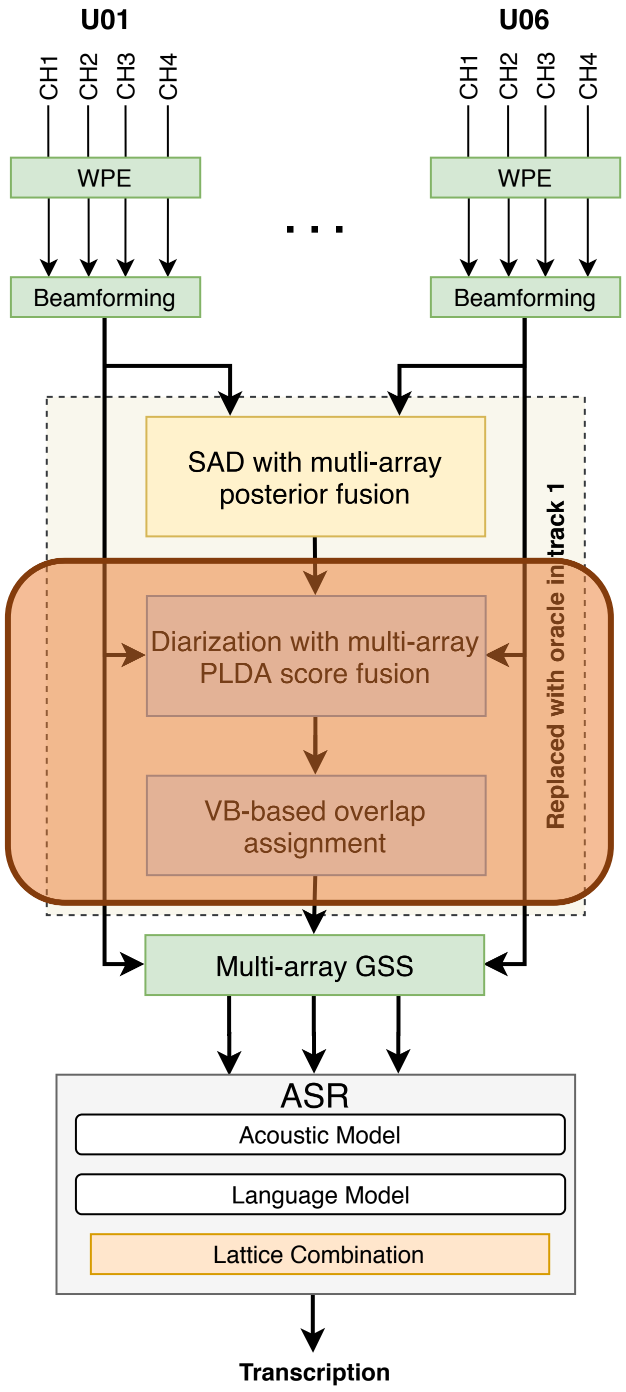
SAD Takeaways



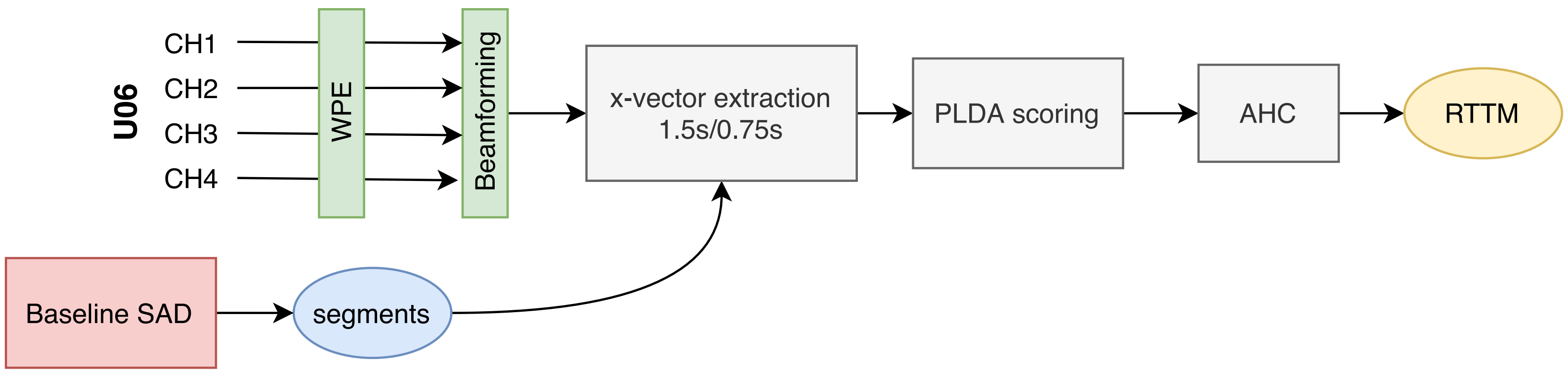
Beamforming is better than channel-level posterior fusion.

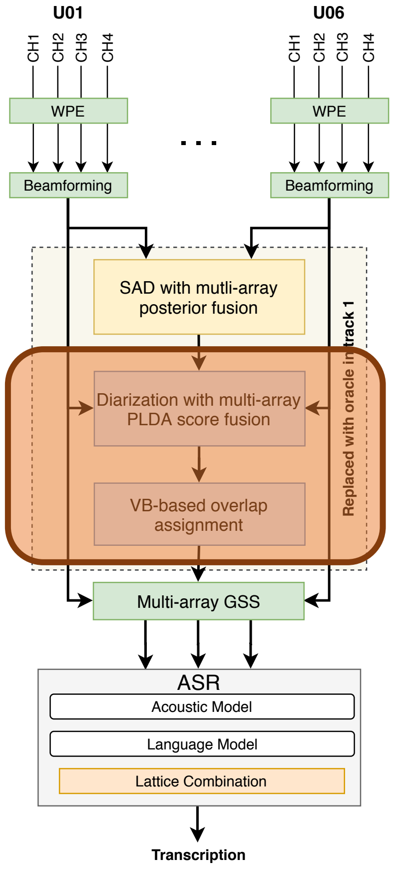
Array-level posterior fusion gives 34% improvement.

Speaker Diarization

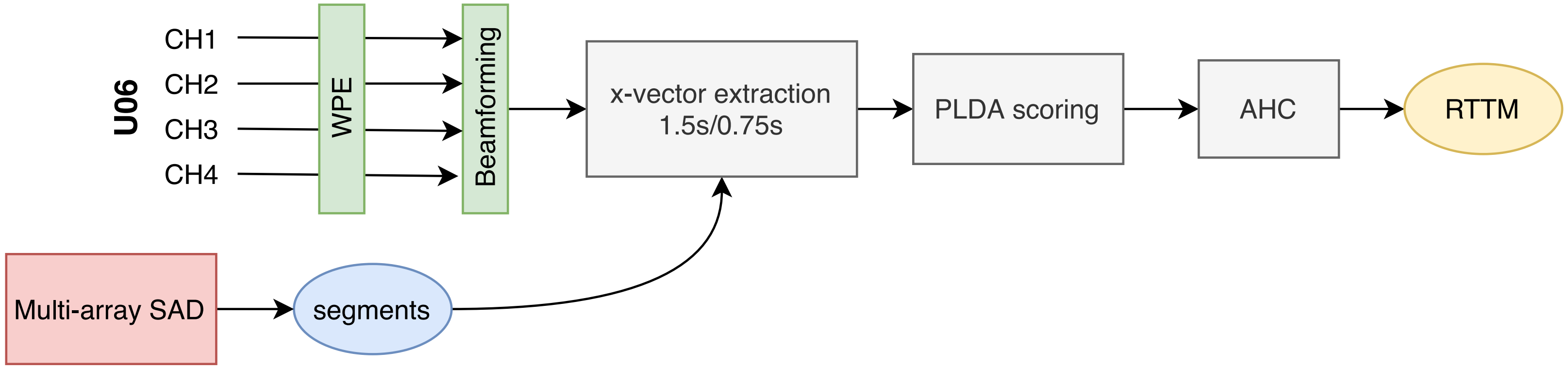


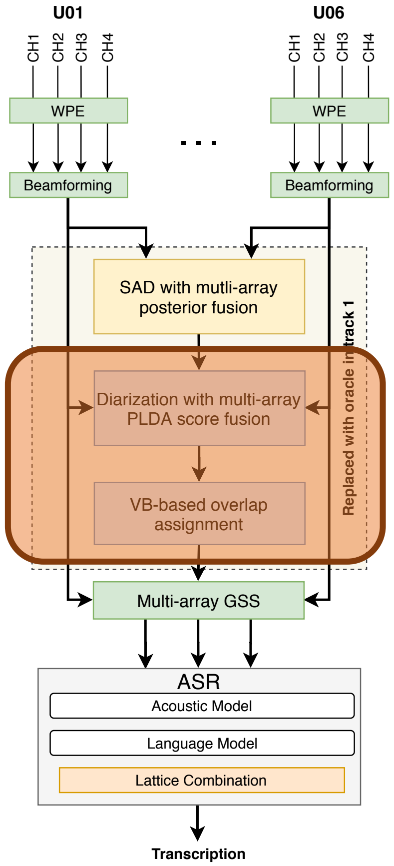
Baseline system





Better SAD improves diarization





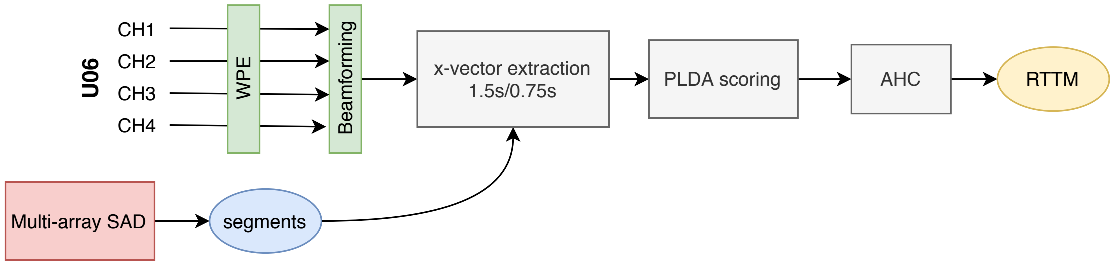
DER on dev for U06 *(evaluated using original reference without UEM)*

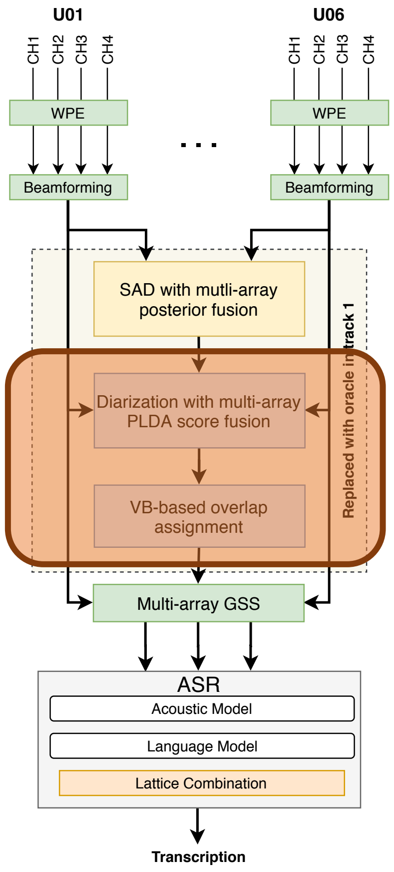
63.49% -> 58.95%

(59.82% mean DER of all arrays)

All further results using multi-array SAD

Two Questions

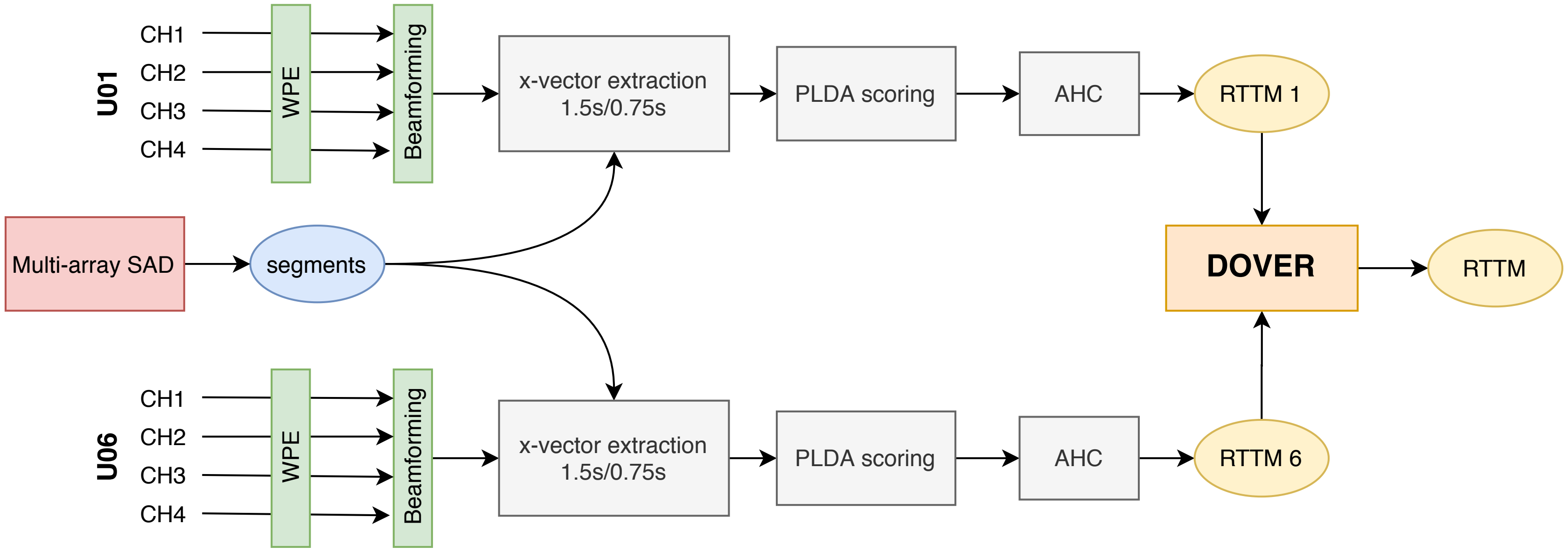


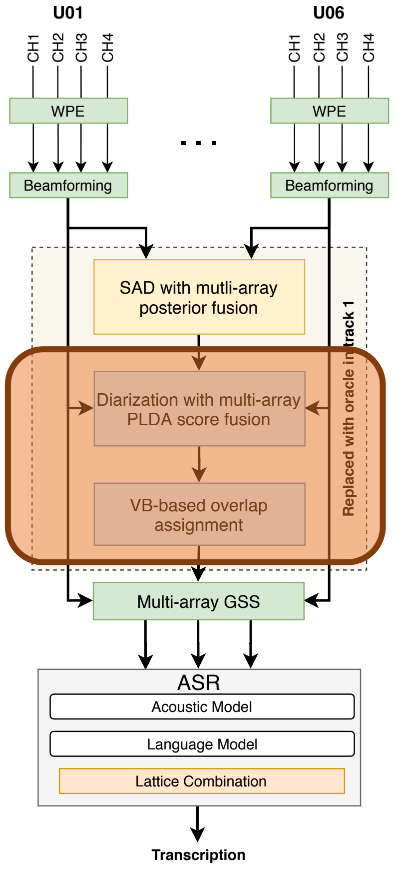


1. Only 1 array is used -> how to use **multi-array** information?

2. Overlaps are ignored -> how to handle **overlapping speakers**?

DOVER to combine array outputs

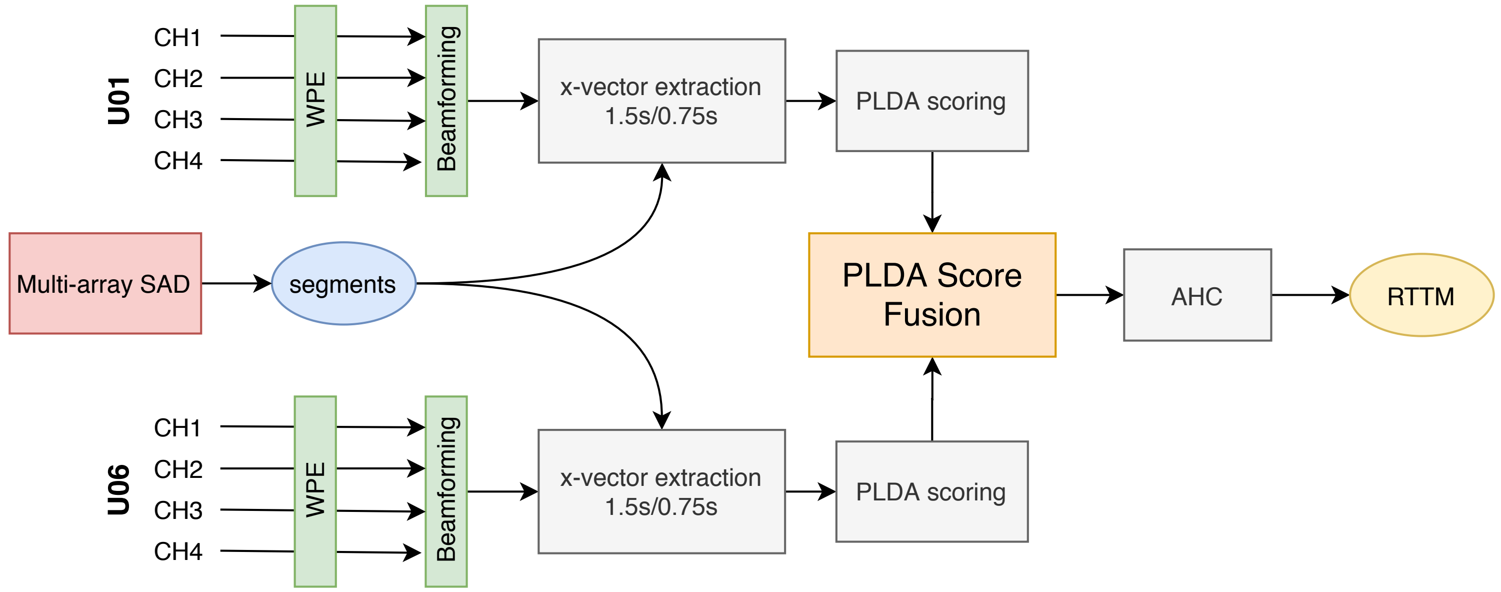


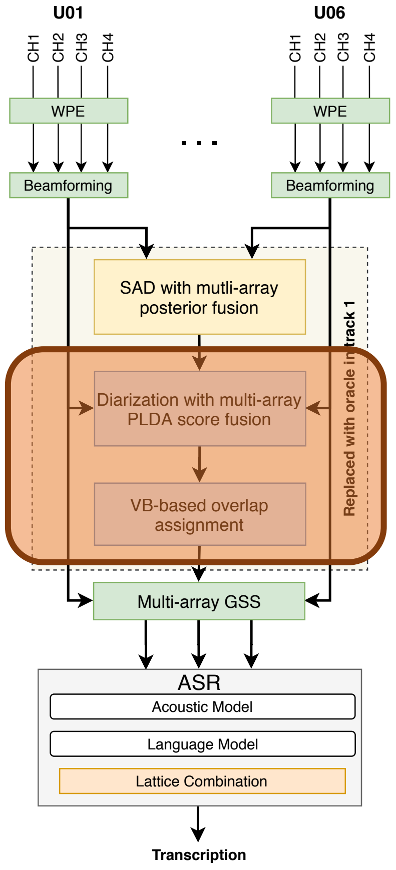


System	Dev DER (%)
Baseline (mean)	**59.8**
DOVER	61.1

Stolcke, Andreas and Takuya Yoshioka. "DOVER: A Method for Combining Diarization Outputs." *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2019): 757-763.

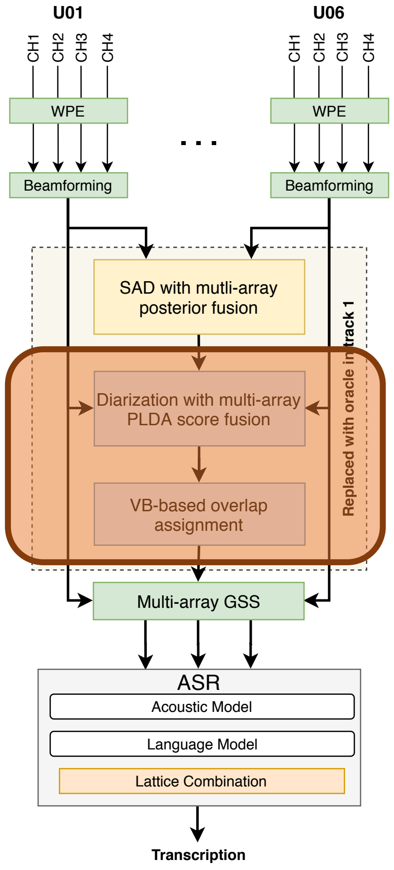
PLDA score fusion helps slightly



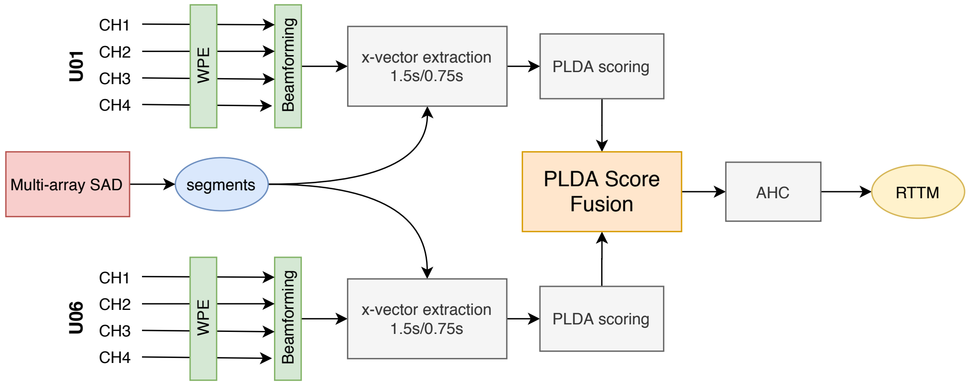


System	Dev DER (%)
Baseline (mean)	59.8
PLDA score MEAN	59.9
PLDA score MAX	**59.0**

Can we do better?

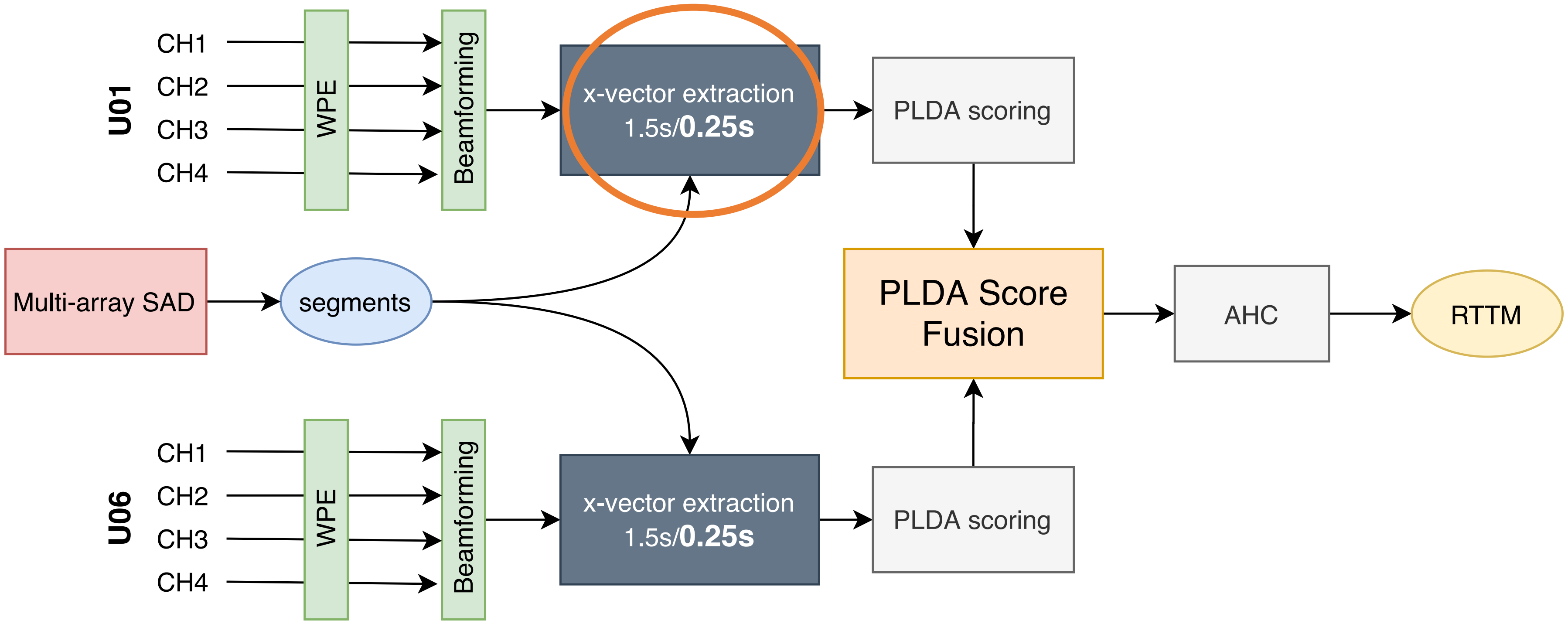


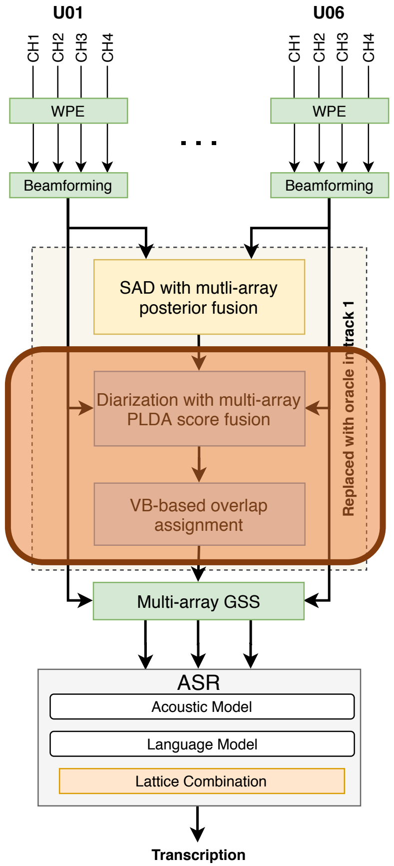
- Small improvement in DER, but underwhelming!
- Maybe need to break segments into **more pieces**?



System	Dev DER (%)
Baseline (mean)	59.8
PLDA score MEAN	59.9
PLDA score MAX	**59.0**

Use **0.25s shift** for x-vector extraction

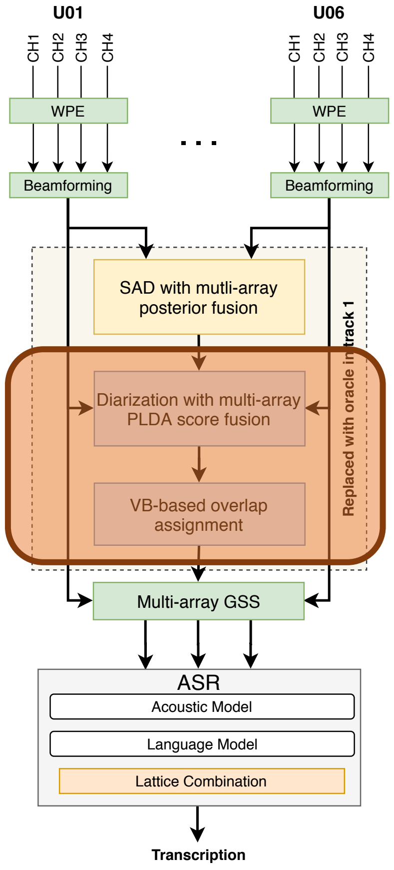


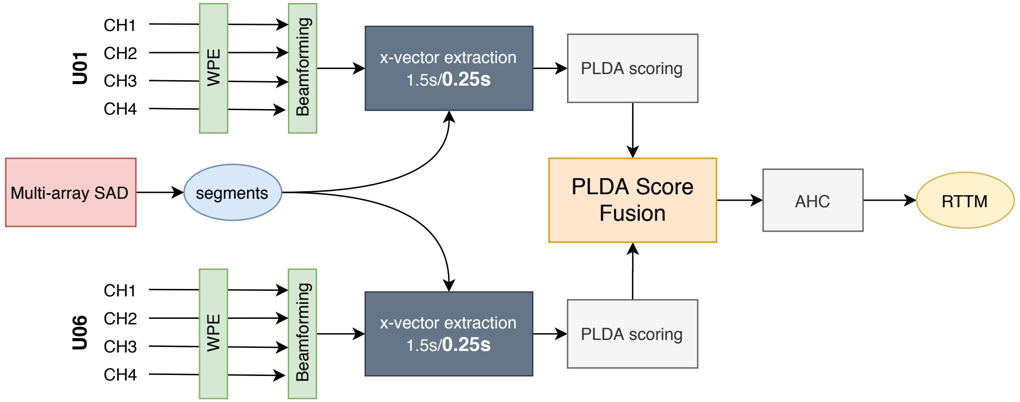


Landini, Federico et al. "BUT System for the Second DIHARD Speech Diarization Challenge." *arXiv: Audio and Speech Processing* (2020): n. pag.

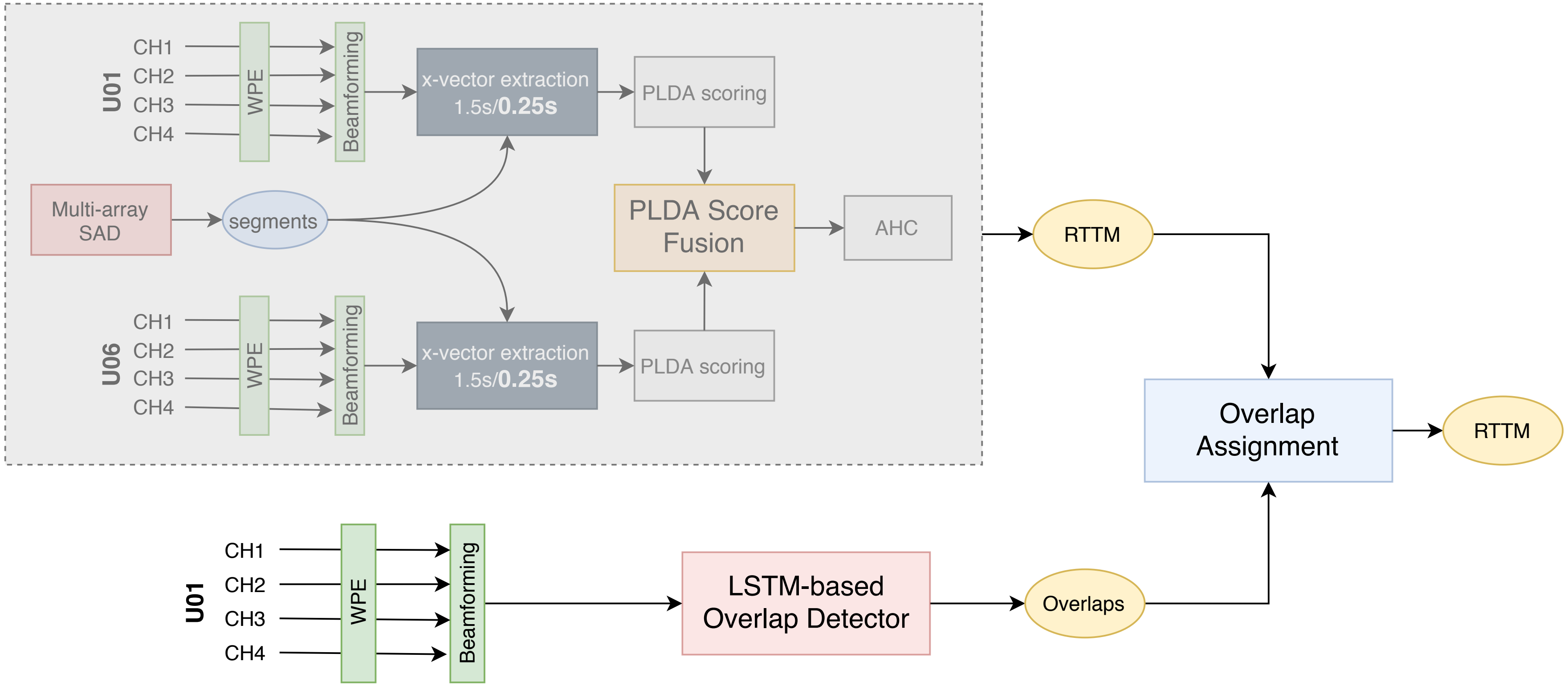
Better DER with smaller segments

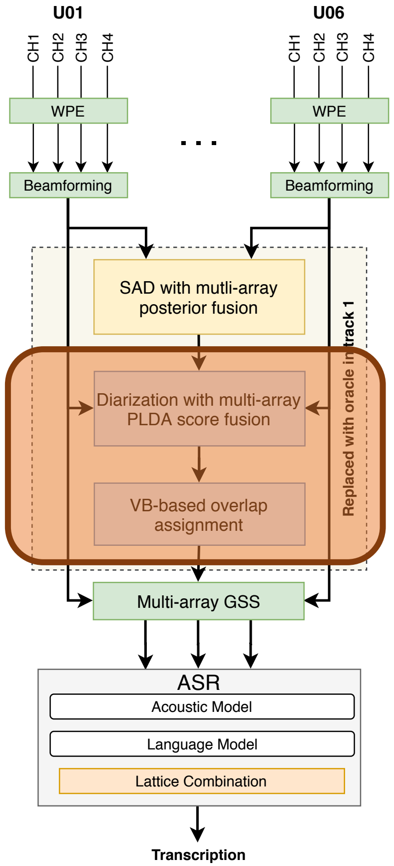
System	Dev DER (%)
Baseline (mean)	59.8
PLDA score MAX	59.0
+ 0.25s shift	**57.9**





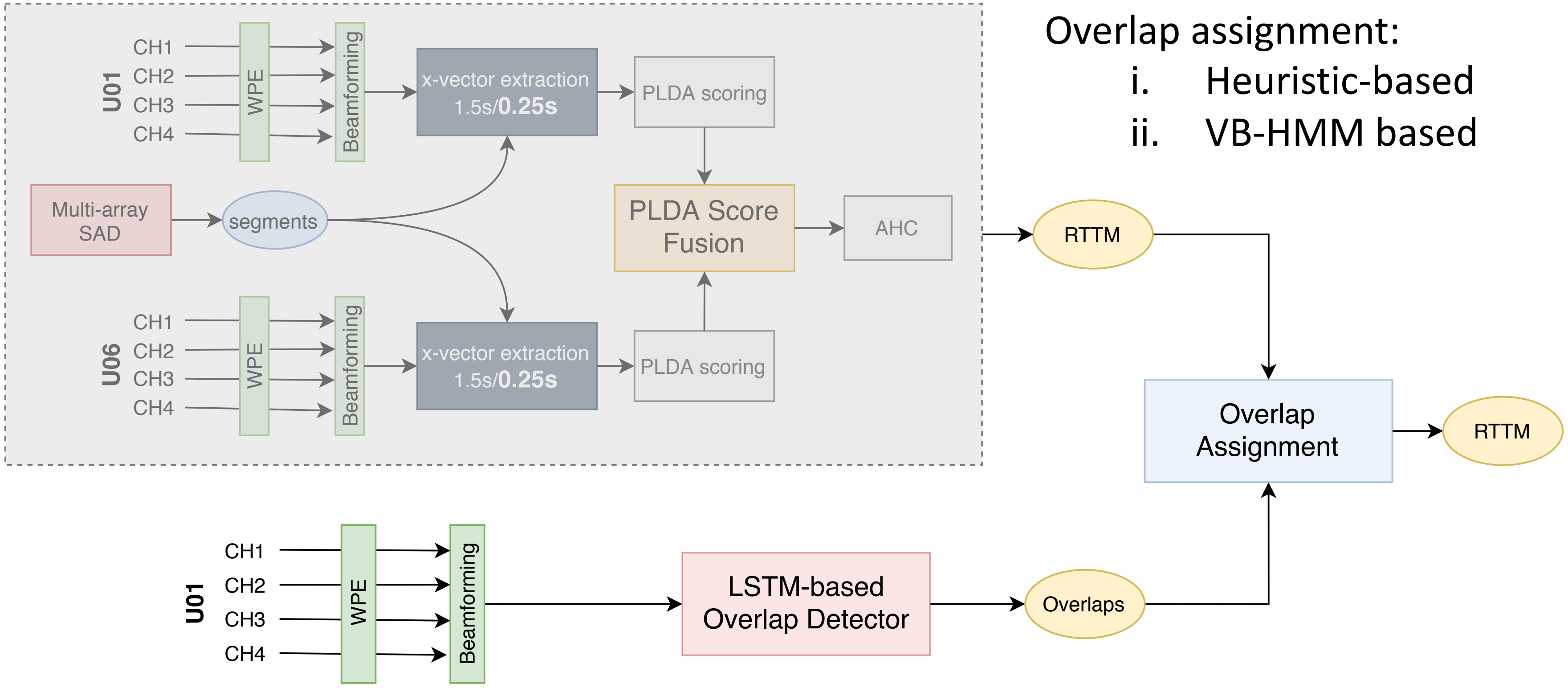
Overlap Handling





Bullock, Latané et al. “Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection.” *ArXiv* abs/1910.11646 (2019): n. pag.

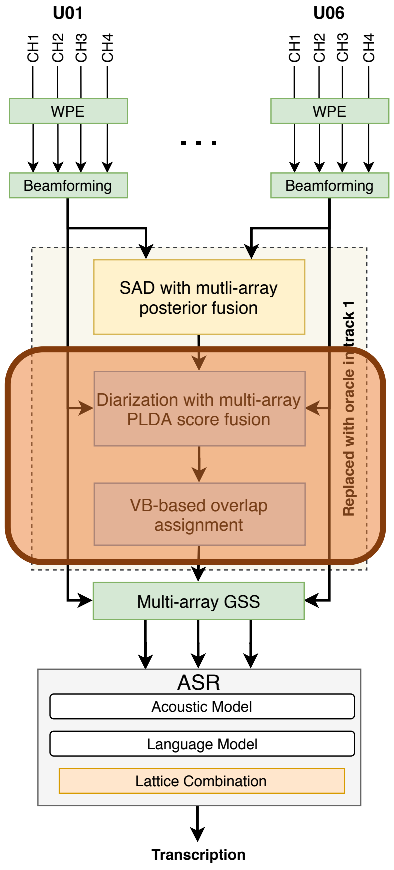
Overlap Handling



Overlap assignment:

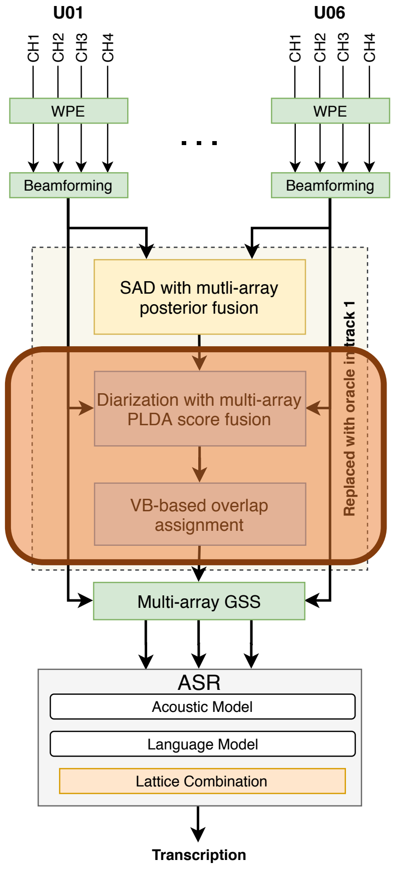
i. Heuristic-based

ii. VB-HMM based

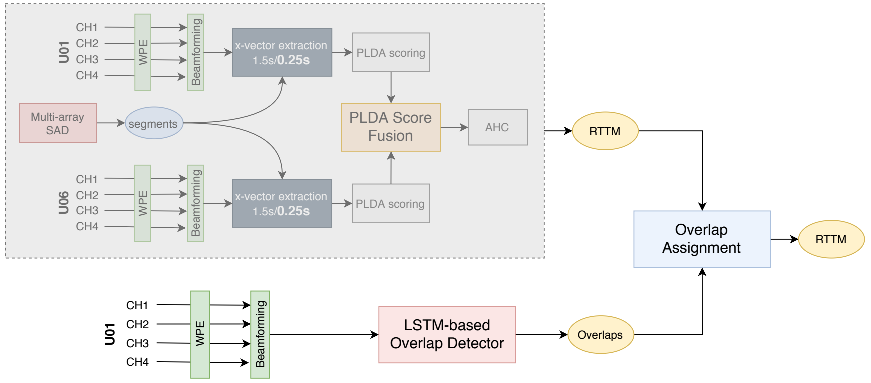


Bullock, Latané et al. “Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection.” *ArXiv* abs/1910.11646 (2019): n. pag.

Simple heuristic gives **small** DER improvement

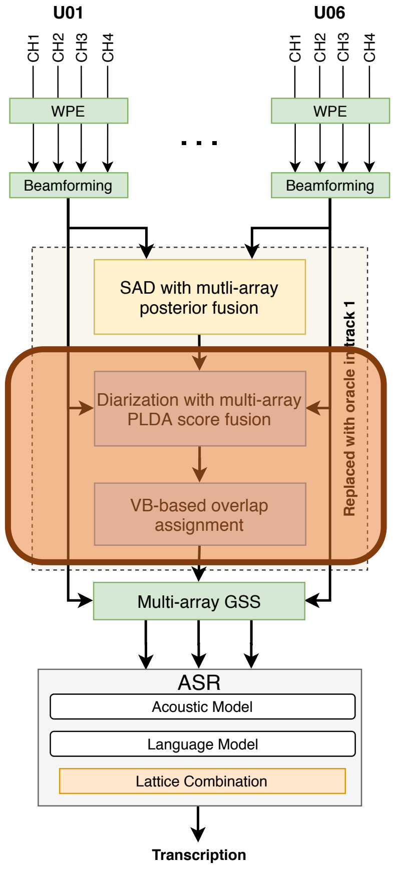


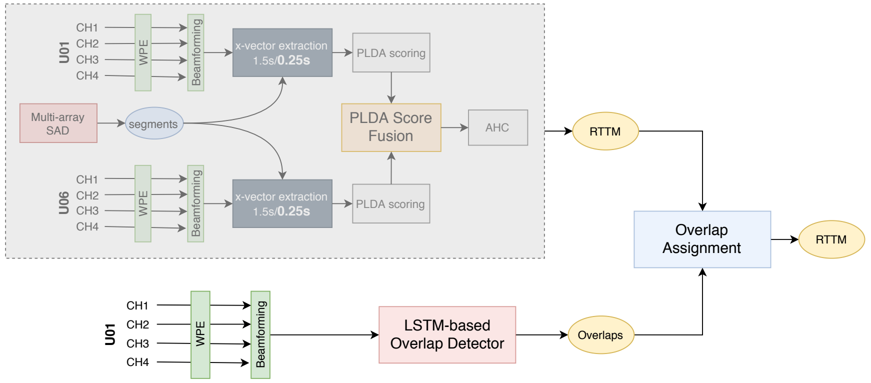
System	Dev DER (%)
Baseline (mean)	59.8
PLDA score MAX	59.0
+ 0.25s shift	57.9
Overlap (heuristic)	**56.2**



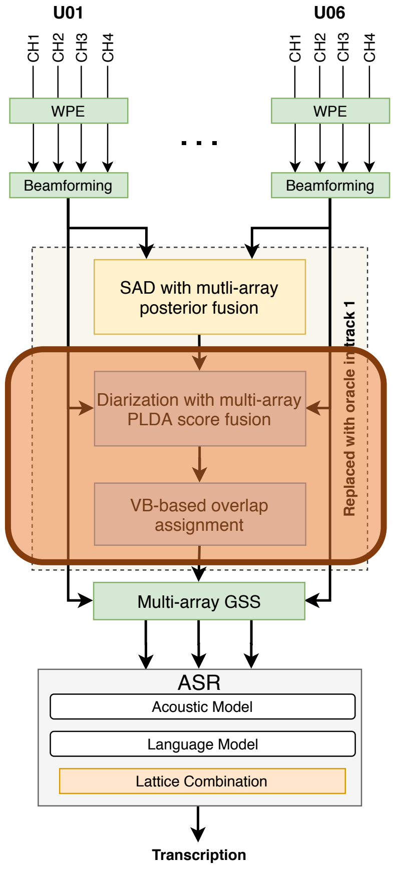
VB-HMM gives **large** DER improvement

System	Dev DER (%)
Baseline (mean)	59.8
PLDA score MAX	59.0
+ 0.25s shift	57.9
Overlap (heuristic)	56.2
Overlap (VB-HMM)	**50.4**





Diarization Takeaways

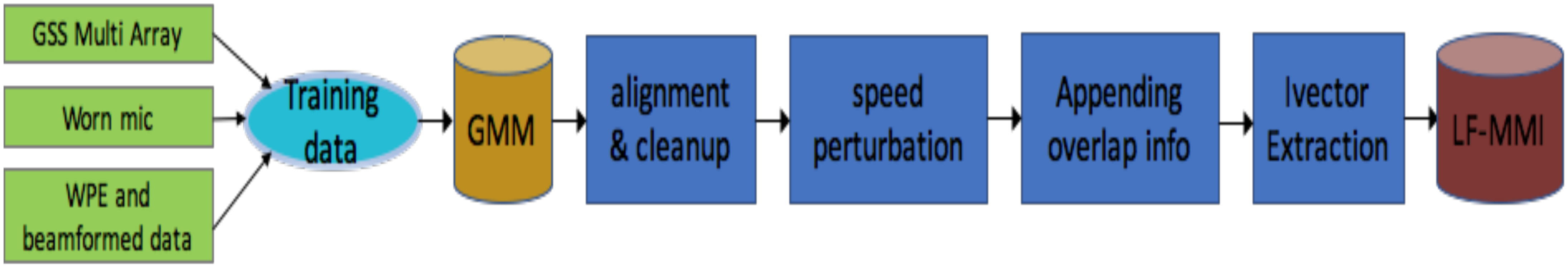


Multi-array PLDA fusion helps with smaller segments.

VB-HMM overlap assignment helps significantly!

Acoustic Model Training Pipeline

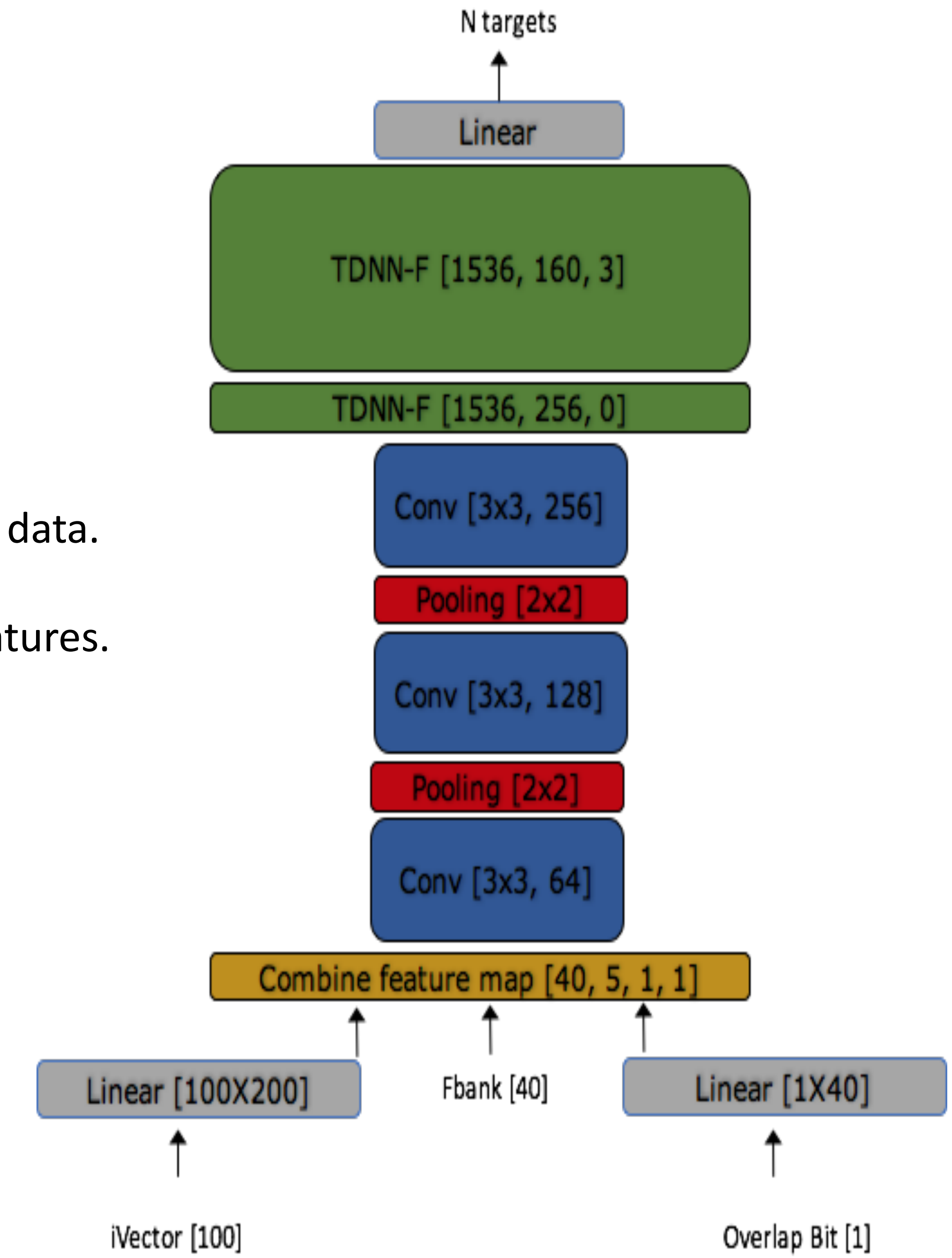
- HMM-GMM/DNN with LFMMI Objective
  - Training data: combination of worn mic utterances (80h), beamformed array data (160h), and multi-array GSS enhanced data (40h)
  - Pronunciation and Silence modeling is used in HMM-GMM training stage.
  - Speed perturbation is used in LFMMI training stage.



Optimizing Model Architecture

- CNN-TDNN-F model outperforms other models on CHiME 6 data.
- Oracle overlap information bit is appended to the Fbank features.

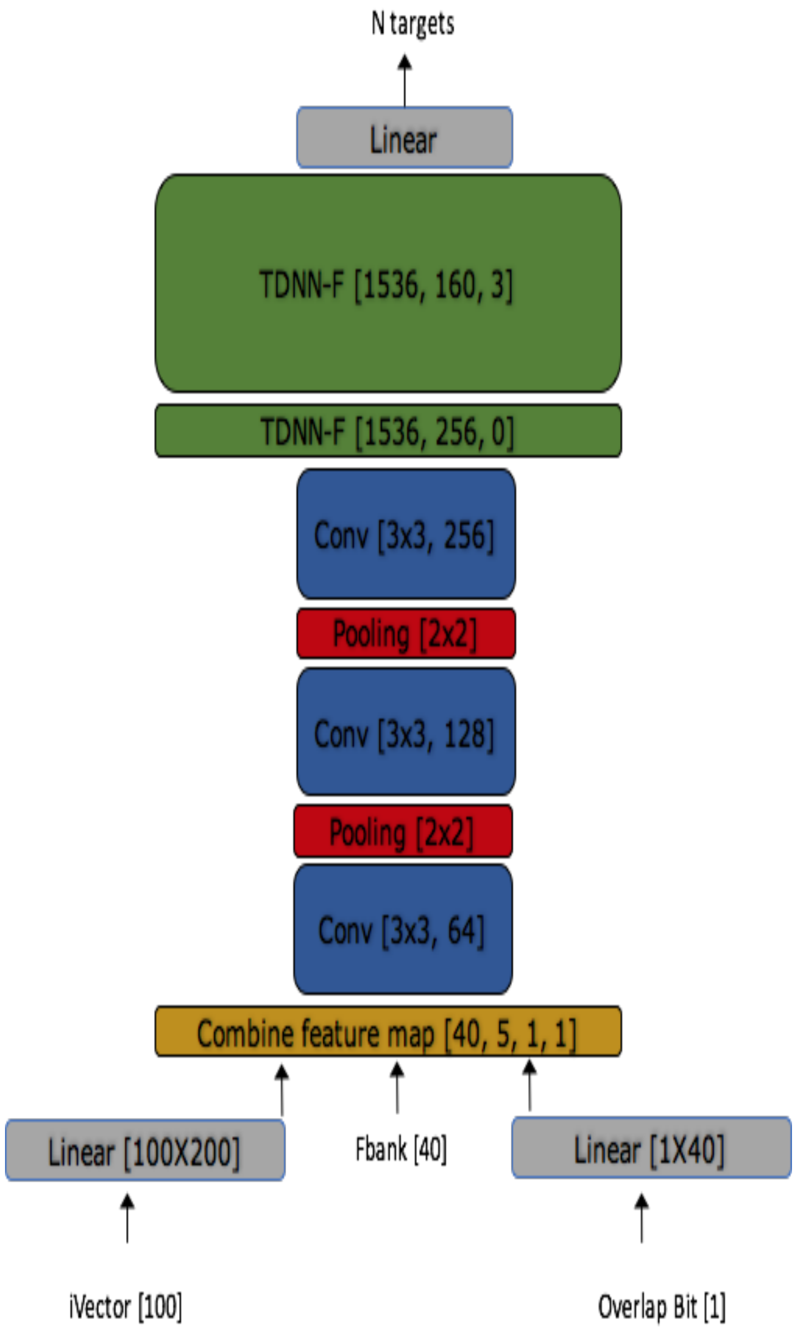
Model Architectures	Dev WER (%)	Eval WER(%)
TDNN-F	51.8	51.3
CNN-TDNN-LSTM	50.1	49.8
SA-CNN-TDNN-F	49.9	49.4
CNN-TDNN-F	48.3	48.5



Training Data Selection

- Enhanced far-field data outperforms raw far-field and simulated data.
- Data selection speeds up experiments.
- Speed perturbation in all cases.

Training data					Dev WER (%)	Eval WER(%)
Worn Mic (80h)	**Worn Mic + Aug (320h)**	**Array (200h)**	**Array +Beamformit (160h)**	**Array + GSS (40h)**		
yes	yes	yes			49.6	49.4
yes	yes	yes		yes	44.6	45.4
yes			yes	yes	44.5	44.9

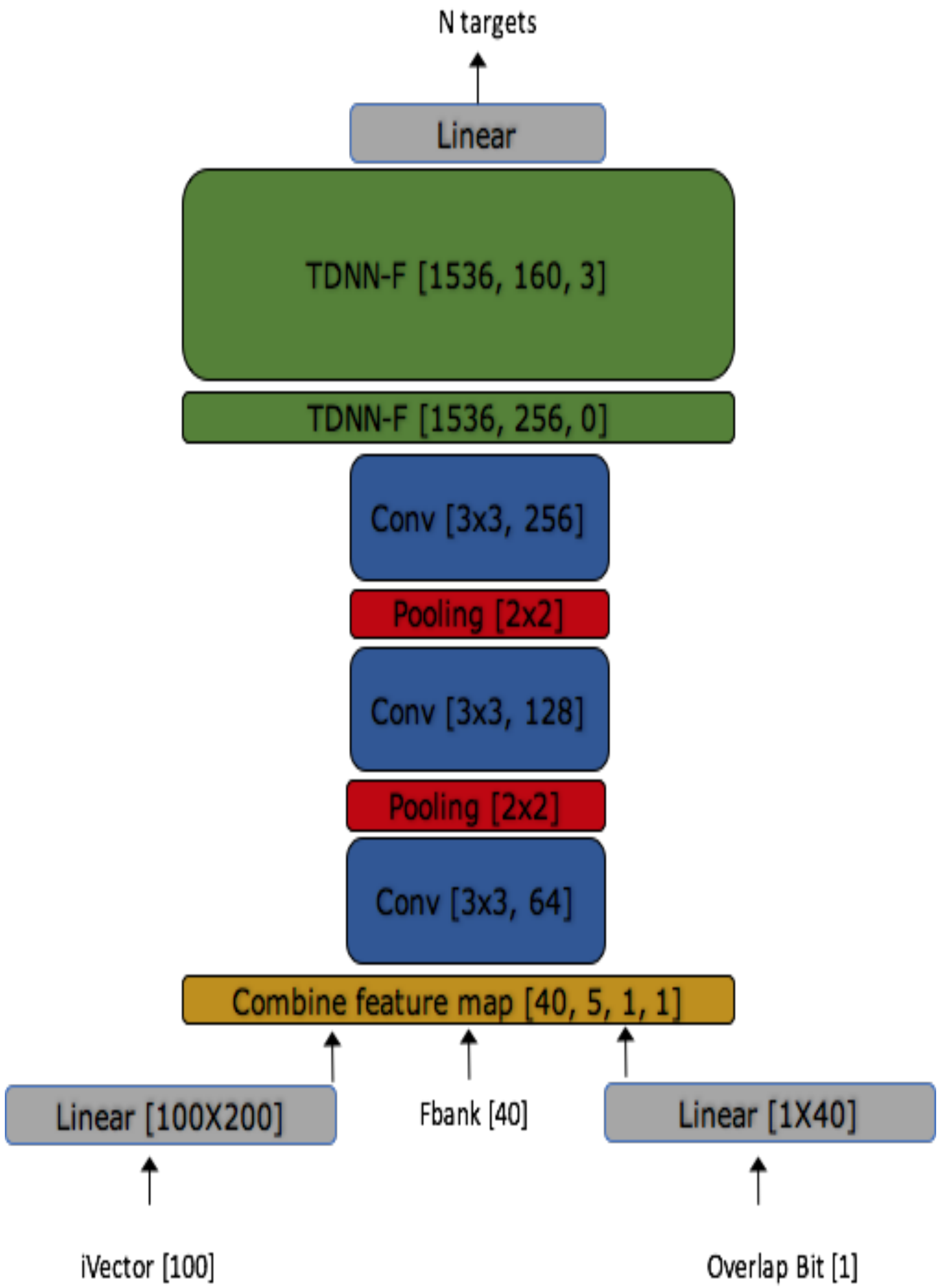


Zorila, Catalin, et al. "An Investigation into the Effectiveness of Enhancement in ASR Training and Test for CHiME-5 Dinner Party Transcription." *arXiv preprint arXiv:1909.12208* (2019).

Appending Oracle Overlap Information (Track 1)

- Using overlap information as auxiliary input for AM training obtains slight WER improvements.
- Track 2 experiments in progress.

Overlap info in ivector	Overlap info in nnet	Dev WER	Eval WER
no	no	44.5	44.9
yes	no	44.9	45.2
yes	yes	44.3	44.4



Neural LM and Rescoring

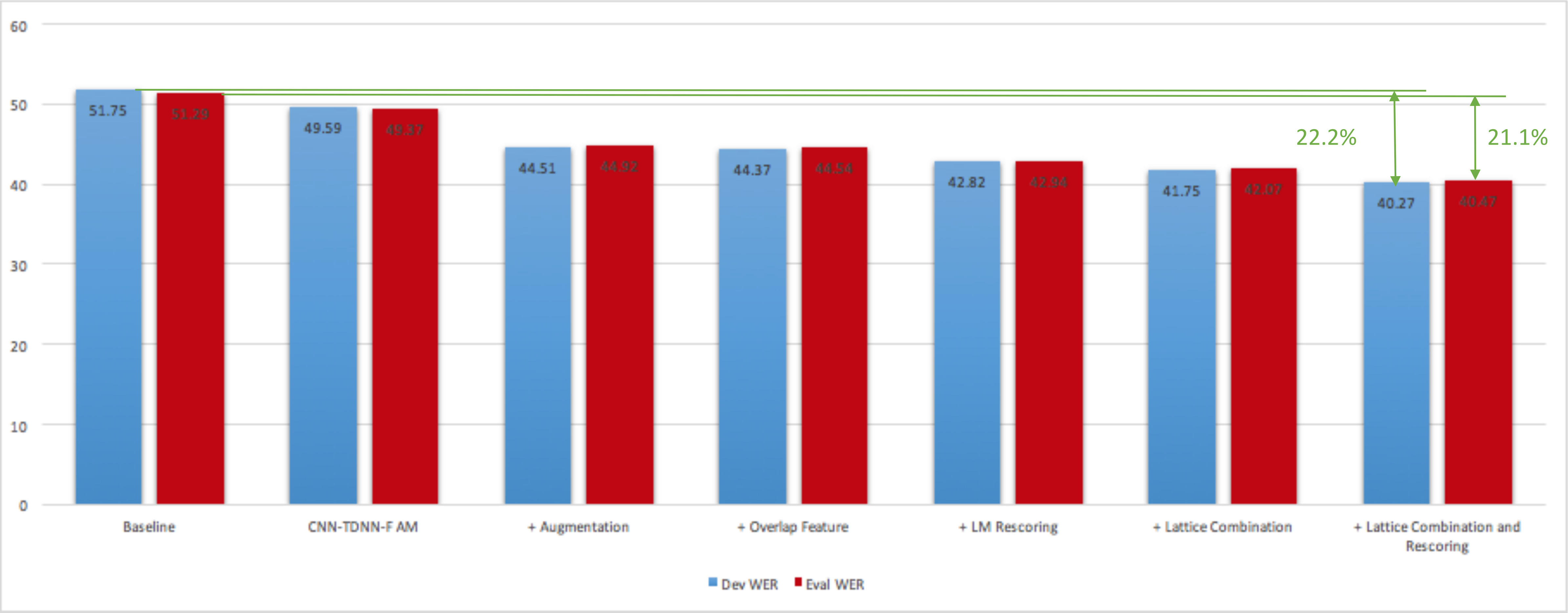
- A forward and a backward LSTM
  - Each is a 2-layer projected LSTM
  - Hidden dim = 512, projection dim = 128
  - Backward LSTM is trained on transcription reversed on sentence level

- 2-stage pruned lattice rescoring
  - Stage 1: Forward LSTM
  - Stage 2: Backward LSTM

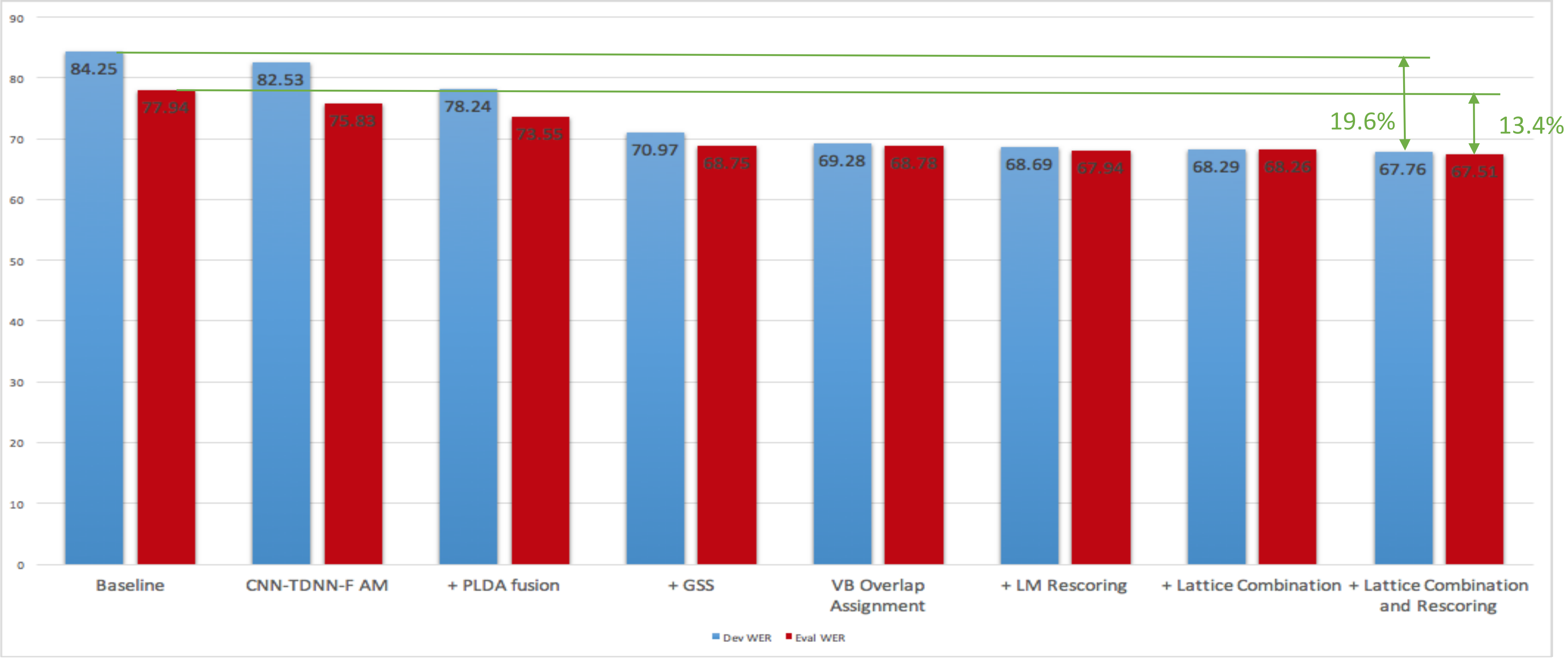
- Kaldi for neural LM training and rescoring

Results and Discussion

Step-by-Step Improvements for Track 1



Step-by-Step Improvements for Track 2



Summary

Frontend

- GSS performance improved with improvement in DER.

SAD

- Multi-array posterior fusion improves error rate by 34% relative

Diarization

- Multi-array PLDA score fusion shows small improvement in DER
- VB-HMM based overlap assignment shows large DER gain and some WER improvement

Acoustic Model

- Deep CNN-TDNN-F model is an effective architecture.
- Enhanced far-field data outperforms raw far-field and simulated data as data augmentation.

Language model

- Neural LM rescoring obtains modest WER reductions.