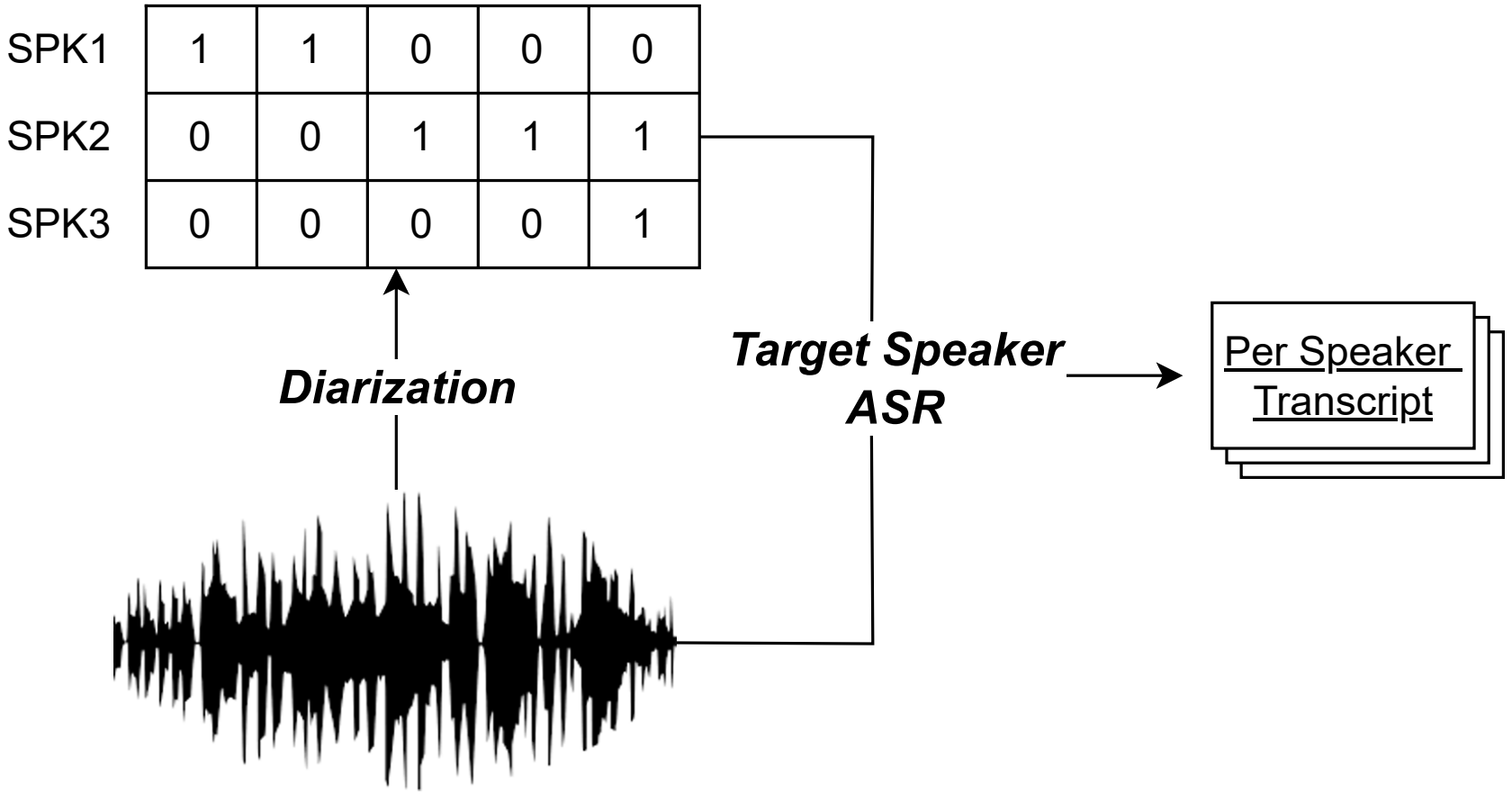
CHiME-8 BUT/JHU System Description

Alexander Polok, Dominik Klement, Jiangyu Han, Šimon Sedláček,
Bolaji Yusuf, Matthew Maciejewski, Matthew Wiesner, Lukáš Burget

General Picture



Single-channel condition

modification of local EEND in Pyannote-like system

1. Weighted average of WavLM Base+ layer outputs
2. FFN, LN, 4 Conformer layers, and a classification head
3. Trained with powerset loss

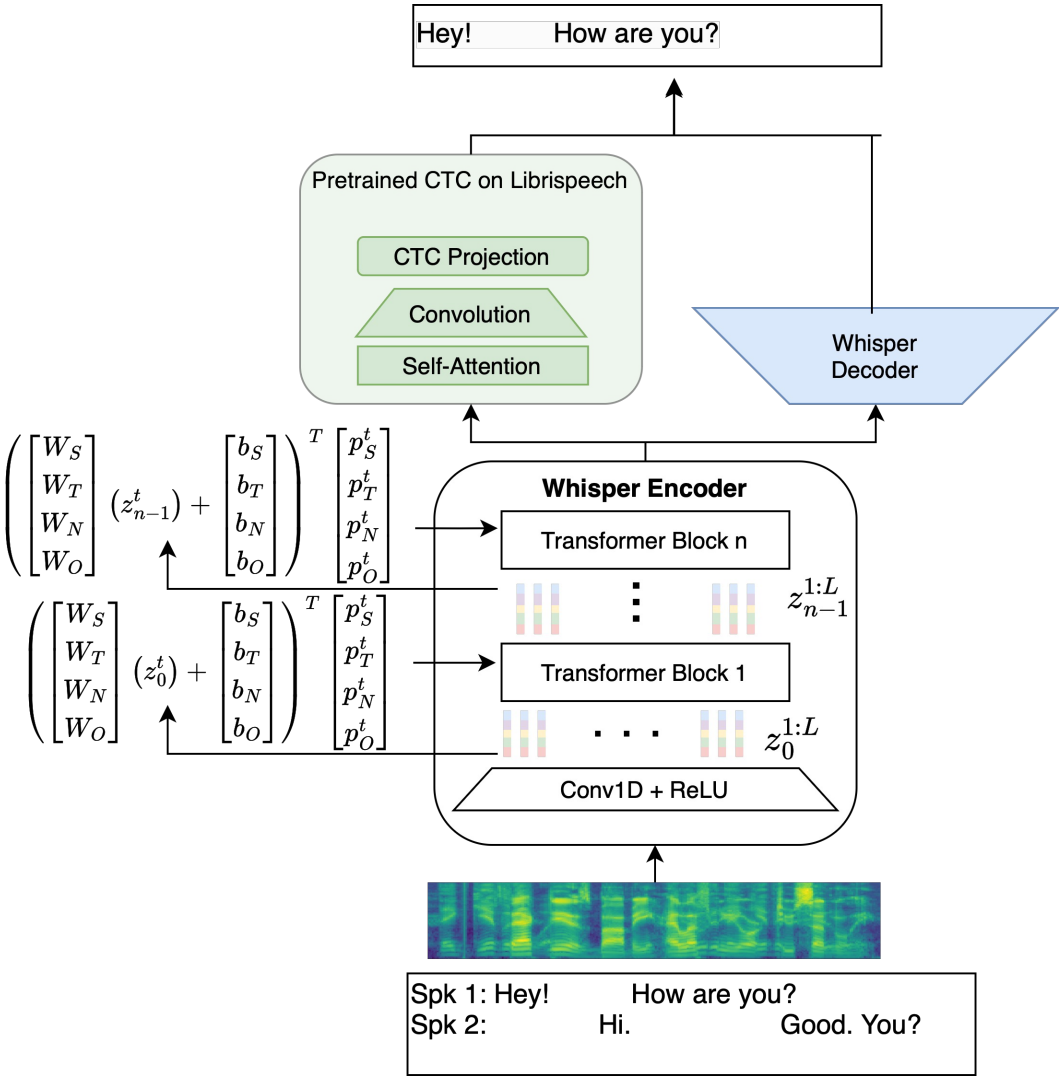
Multi-channel condition

parallel processing by first 4 layers of WavLM

$$T^l = \frac{1}{C}\sum_{c=1}^{C} H_c^l$$

$$\widehat{H}_c^l = H_c^l + \mathrm{LN}(\mathrm{Linear}(H_c^l || T^l)$$

1. **Query-Key Biasing**
2. **Segmental State Projections**
   1. STNO mask (Silence, Target speaker, Non-target speaker, Overlap)
   2. Affine transformations to modify per layer inputs W_S, W_T, W_N, W_O
   3. CTC head



https://github.com/BUTSpeechFIT/CHIME2024

Team Name	System Tag	tcpWER (%) (eval)	tcpWER (%) (dev)
BUTJHU	sys2	40.1 %	35.7 %

Table 5. Effect of enlarging and extending Whisper with a CTC head.

Model	No CTC	From Random CTC	+ Preheating	+ Amplification
small.en	33.82	32.54	30.59	30.04
medium.en	31.15	29.42	27.03	–
large-v3	–	–	–	23.79 (25.19)

Table 3. Analysis of different Segmental State Projection initialization methods and the impact of non-disturbing initilialization.

	rand. init	non-disturbing-init	disparagement
bias only	28.44	28.02	28.05
diagonal & bias	118.97	27.34	**26.73**
full	129.00	46.13	44.60

Table 4. Effect of reducing information for Segmental State Projection.

	STNO	STO	ST	T
diagonal & bias 12 l.	27.14	**26.62**	59.04	68.15