Target Speaker ASR with Whisper

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

Brno University of Technology, Johns Hopkins University

Abstract

We fine-tuned Whisper to condition on oracle and hypothesis diarization outputs for the NOTSOFAR task.

- Two **diarization-aware** approaches to automatic speech recognition (ASR) that **repurpose Whisper** to perform **target speaker ASR**

1. Query-Key Biasing **2. Segmental State Projections**

Query-Key Biasing (QK)

- Speaker attention mask with modified positional embeddings using frame-level speaker labels

$$a_{ij} = softmax\left(\frac{(W_q q_i)^T (W_k k_j)}{\sqrt{d}}\right)$$

- Weights and keys are modified as

$$\hat{W}_{q,k} = \begin{bmatrix} W_{q,k} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}, \hat{q}_i = \begin{bmatrix} q_i \\ 1 \end{bmatrix}, \hat{k}_j = \begin{bmatrix} k_j \\ -c \end{bmatrix}$$

- For target speakers the weights are unchanged.
- For non-target speakers the weight is controlled by bias parameter c (seen below)

$$\begin{bmatrix} (W_q q_i)^T & 1 \end{bmatrix} \begin{bmatrix} W_k k_i \\ -c \end{bmatrix} = (W_q q_i)^T (W_k k_j) - c$$

- For non-target speaker frames the preceding positional encoding is repeated.
- For target-speaker frames positional encodings are correspondingly shifted

Diarization

- **Single Channel (SC)** builds off of pyannote pipeline
- EEND system trained from WavLM++ Base model
  - Weighted average of WavLM layer outputs
  - WavLM features passed to layer-norm, 4 conformer layers and a classification head

---

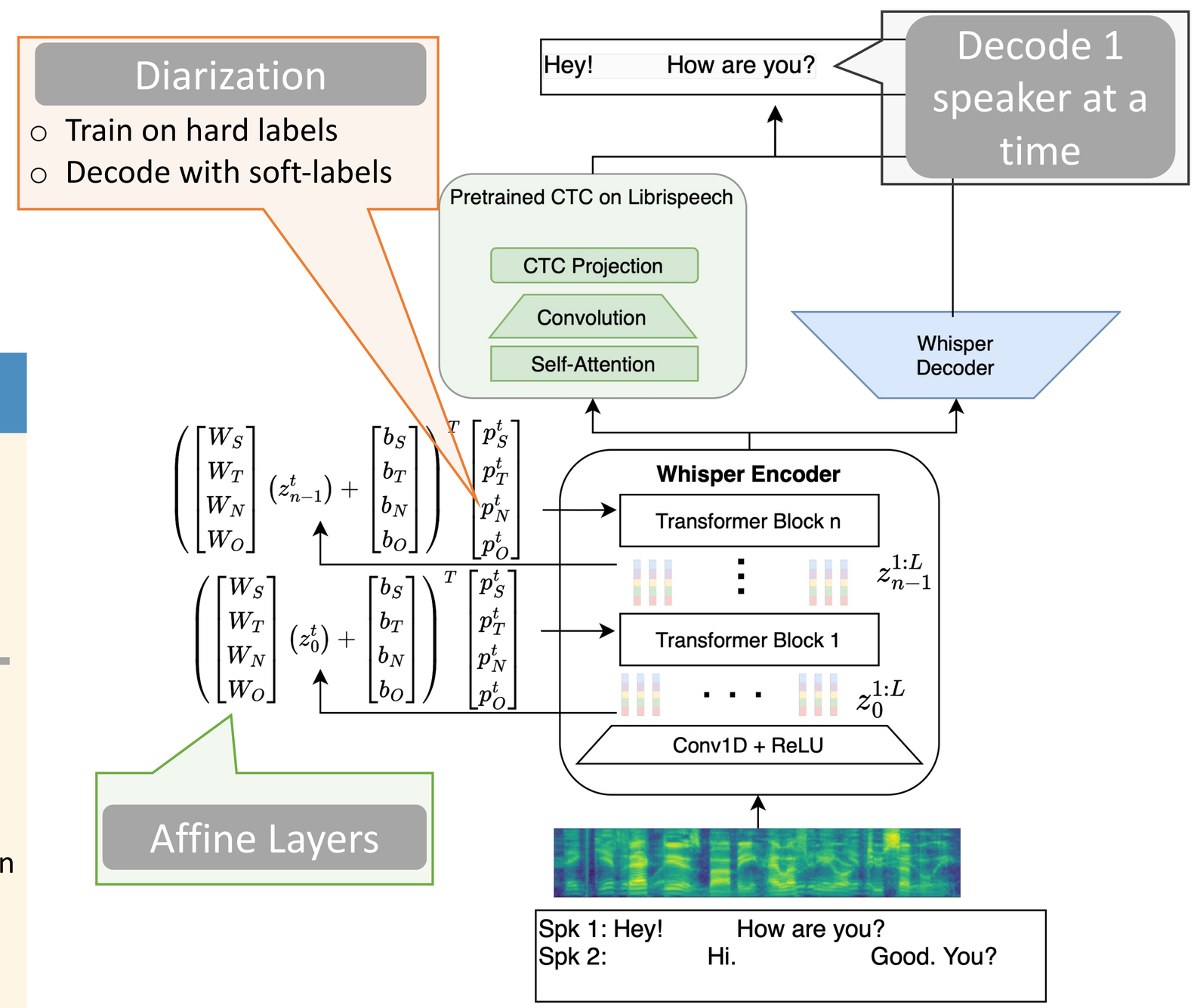
- **Multi-channel processing (MC)**
  - Run 1 WavLM model per channel in parallel
  - Model inter-channel dependencies
  - Let $\mathbf{H}_c^l$ be WavLM activations at layer c and channel l , LN be layer norm and || concatenation
  - Next layer inputs are determined b

$$\mathbf{T}^l = \frac{1}{C}\sum_{c=1}^{C} \mathbf{H}_c^l \quad \bar{\mathbf{T}}_c^l = \mathrm{LN}(\mathrm{Linear}(\mathbf{H}_c^l || \mathbf{T}^l)) \quad \hat{\mathbf{H}}_c^l = \mathbf{H}_c^l + \bar{\mathbf{T}}_c^l.$$

  - Optionally followed by GSS

Segmental State Projections (SSP)

- Conditions on the **entirety of the diarization output**
  - Encode:
    - Silence (S), Target Speech (T) Non-target Speech (N) Overlapped speech (O)
    - Each STNO class is encoded via a frame-level affine transformation at the input of transformer blocks
- Additionally fine-tune with CTC to mitigate hallucinations
  - Down-sampled outputs for CTC
- **Multi-channel decoding** by running single-channel model in parallel
  - Average 8th-layer outputs



Results

Model	tcpWER [%]	DER [%]
Baseline SC	45.8	-
Baseline MC	31.6	-
QK-Bias Med	51.3	10.9
QK-Bias Large	48.7	10.9
SSP Large - sc	36.5	10.9
SSP Large+FT - sc	35.9	10.9
SSP Small - mc	36.9	10.4
SSP Large - mc	33.2	10.4
GSS Med - mc	29.6	10.4

- QK-Biasing works slightly worse than the NOTSOFR baseline
- Segmental state projection outperforms the baseline
- Multi-channel processing improves performance slightly
- Additional GSS greatly improves performance

Conclusion / Future Work

- Whisper can be turned into a target speaker ASR system

Future Work:

- Upcoming ICASSP Submission
- Model and data scaling
- Analysis of importance of STNO transformations