Adapting Speech SSL to Text

Presentor: Jiatong Shi (jiatongs@cs.cmu.edu)

Proposed with Ann Lee, Shinji Watanabe, and Hung-yi Lee

Content

- Motivation
- Previous works
- Our proposal
- Plan

Motivation

- Different modalities has different features
  - Information variation
  - Information density variation
  - Length variation
  - Context variation

Motivation (Cont’d)

- Different modalities has different features
  - Information variation
    - Speech: phonetic info, prosody, emotion from acoustic, noise, etc.
    - Text: semantic info, syntax, morphology, etc.
  - Information density variation
  - Length variation
  - Context variation

Motivation (Cont’d)

- Different modalities has different features
  - Information variation
  - Information density variation
    - Speech: highly correlation to consecutive frames (more redundancy)
    - Text: tokens can be more informative give context
  - Length variation
  - Context variation

Motivation (Cont’d)

- Different modalities has different features
  - Information variation
  - Information density variation
  - Length variation
    - Speech: longer sequence
    - Text: shorter sequence
  - Context variation

Motivation (Cont’d)

- Different modalities has different features
  - Information variation
  - Information density variation
  - Length variation
  - Context variation
    - Speech:
      - shorter context dependency for acoustic info
      - longer context dependency for linguistic info

Motivation (Cont’d)

- [Target] A better framework to utilize both speech and text pre-trained model for downstream semantic tasks in speech processing

Motivation (Cont’d)

- [Target] A better framework to utilize both speech and text pre-trained model for downstream semantic tasks in speech processing
- [Assumption] Self-supervised features learned from different modalities are likely to be in different feature space.

Motivation (Cont’d)

- [Target] A better framework to utilize both speech and text pre-trained model for downstream semantic tasks in speech processing
- [Assumption] Self-supervised features learned from different modalities are likely to be in different feature space.
- [Research Question] How we can align the speech self-supervised feature into a similar feature space of text so as to take benefit from text pre-trained models?

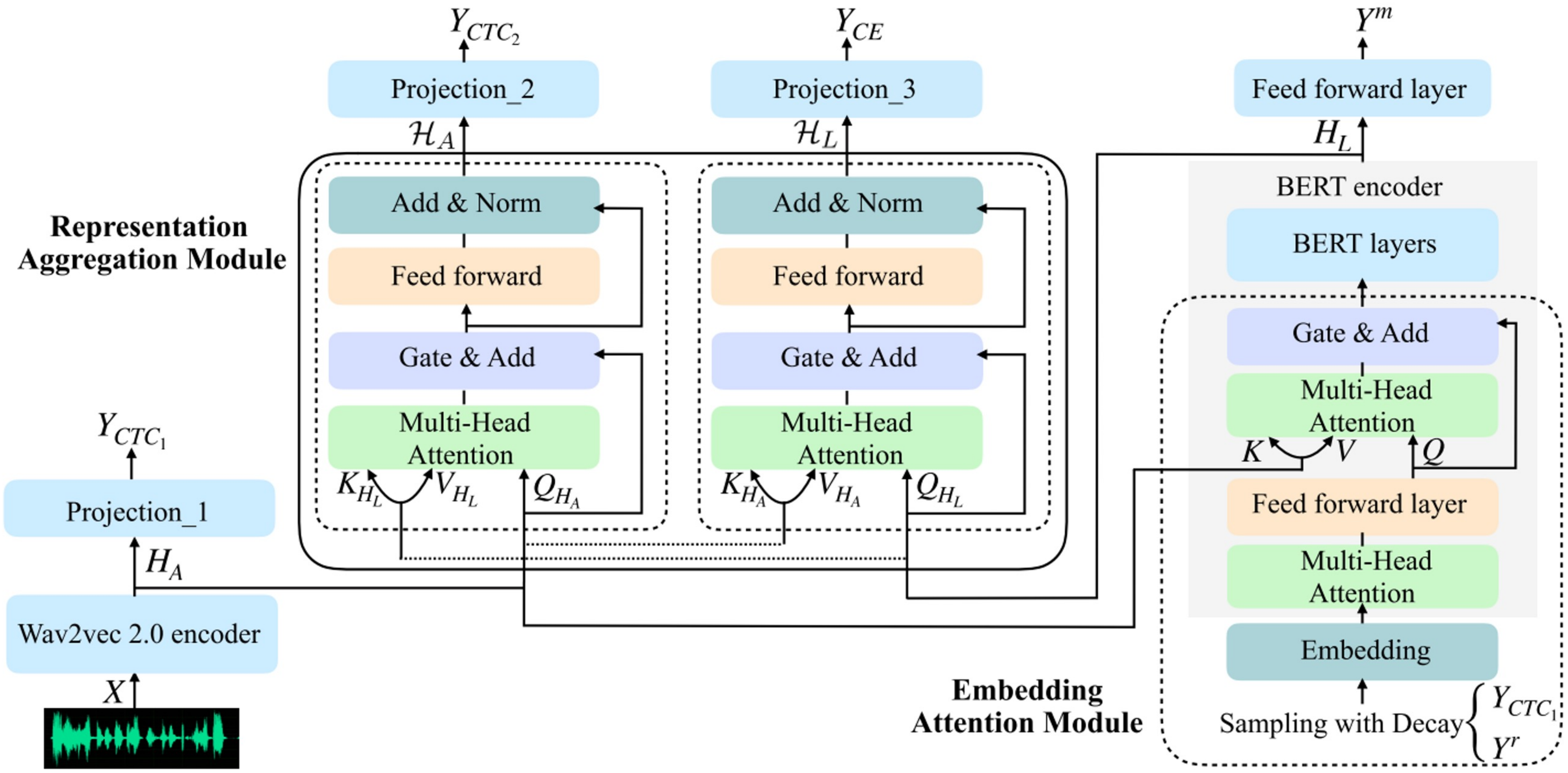
Previous works

- Cross-attention
- Pre-defined alignment
- Mixup

Cross-attention

- Cross-attention could align speech and text space.
- Pros:
  - Direct alignment
  - No required extra training
- Cons:
  - Need supervision or pre-defined hyper-params

- Related works:
  - Wav-BERT: Cooperative Acoustic and Linguistic Representation Learning for Low-Resource Speech Recognition (Zheng et al. 2021)
  - Non-autoregressive Transformer-based End-to-end ASR using BERT (Yu et al. 2021)



Wav-BERT: Cooperative Acoustic and Linguistic Representation Learning for Low-Resource Speech Recognition (Zheng et al. 2021)

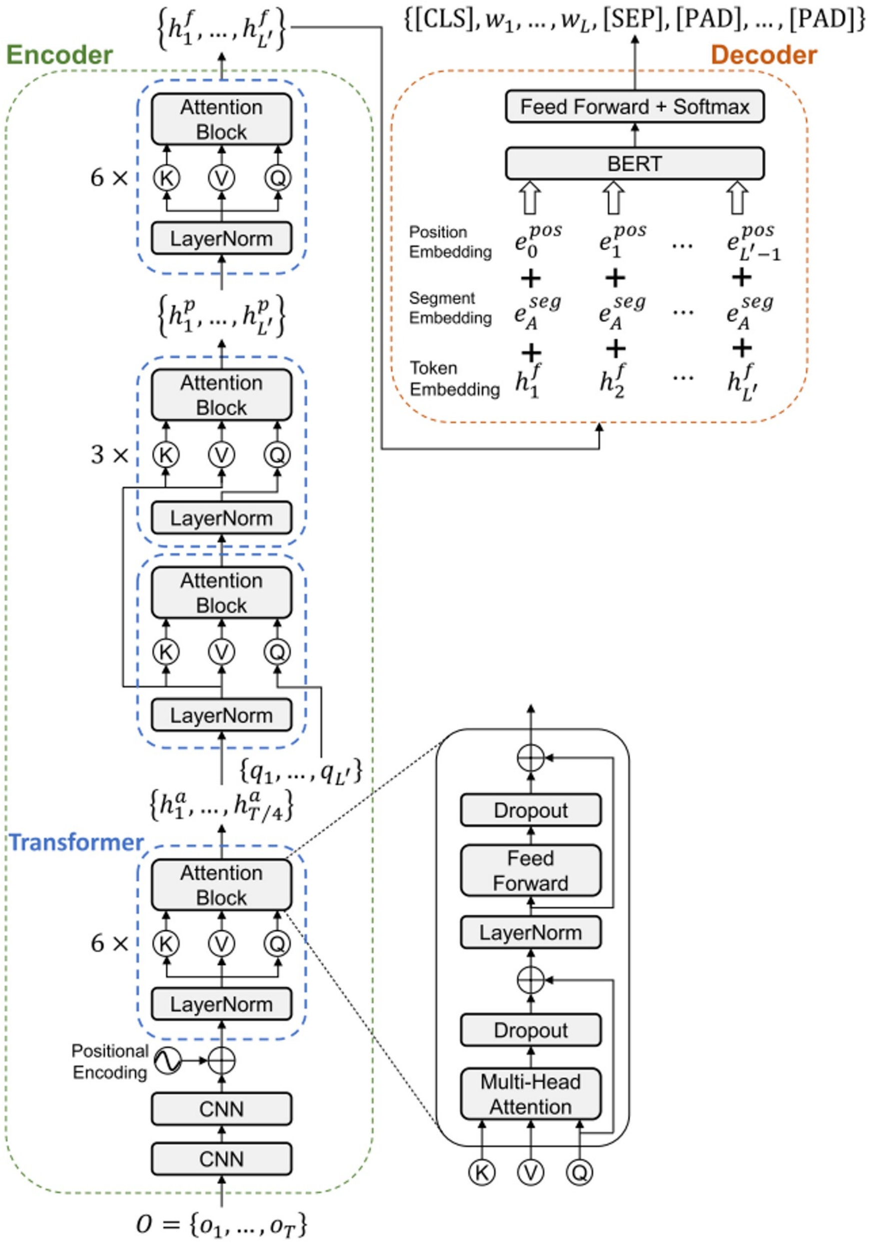
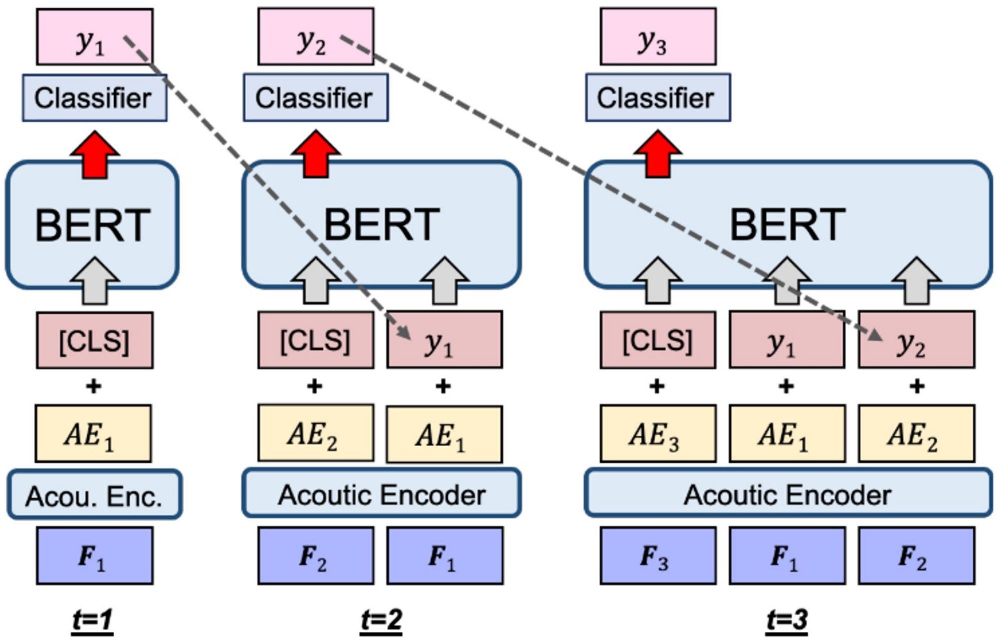


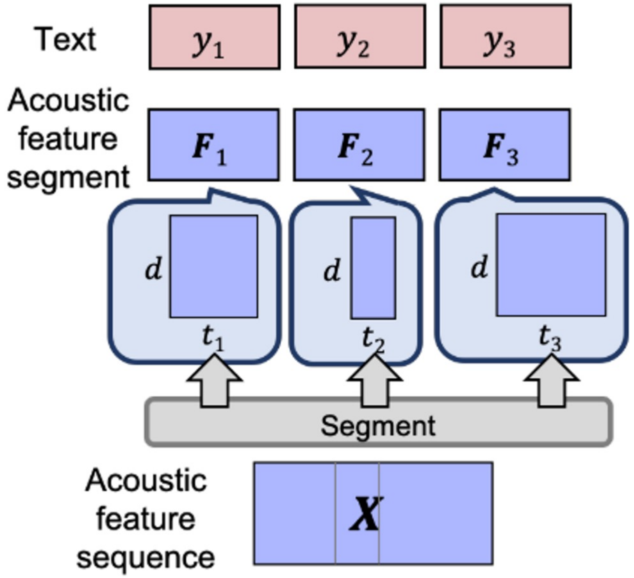
Figure 1: *The architecture of the proposed non-autoregressive transformer-based end-to-end ASR.*

- Apply L’ as a positional vector to perform cross-attention
- L’ = 60 when apply the method on AISHELL-1

Non-autoregressive Transformer-based End-to-end ASR using BERT (Yu et al. 2021)

Pre-defined alignment





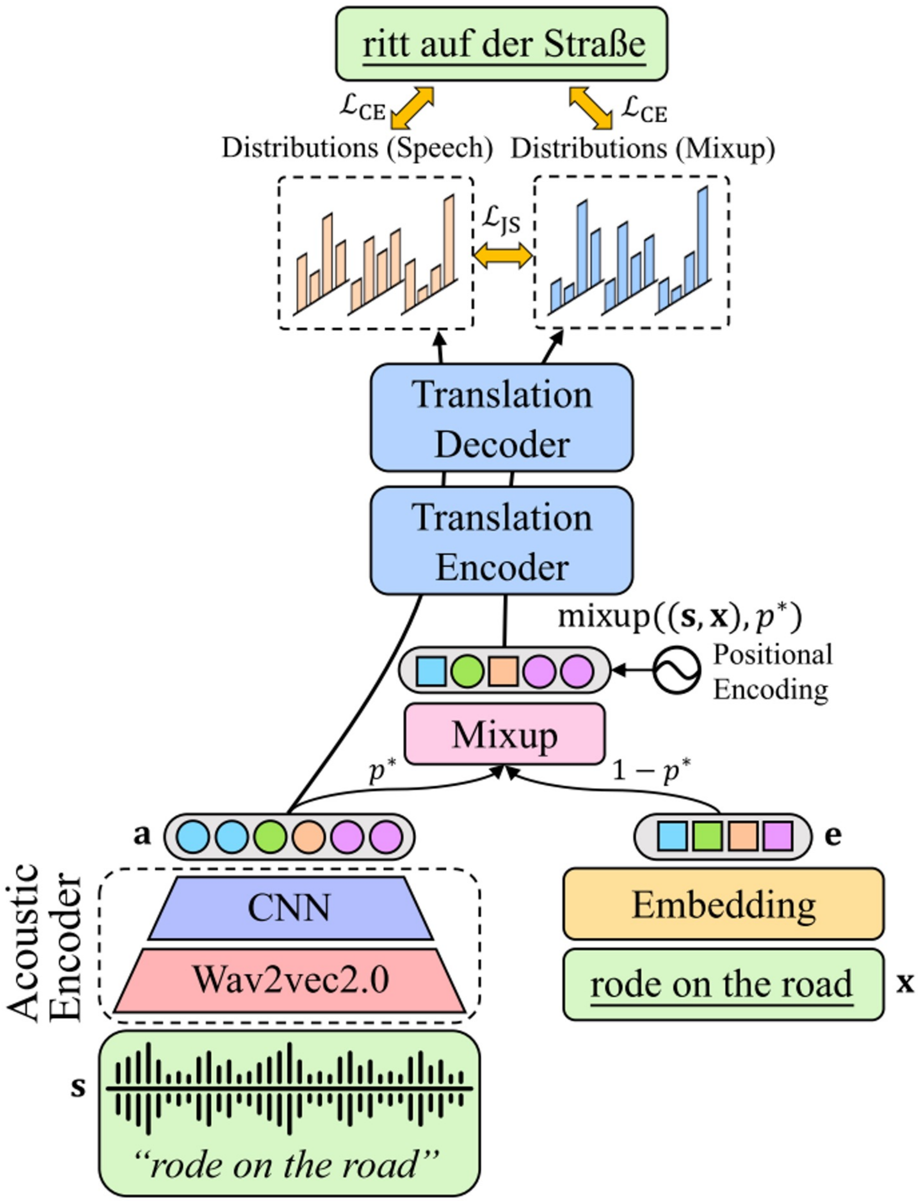
Alignment is either from oracle alignment or by pre-defined segment length (estimate by average length in train set)

SPEECH RECOGNITION BY SIMPLY FINE-TUNING BERT (Huang et al. 2021)

Mixup

- Mixup the speech encoder states with text embeddings
- Pros:
  - Not only align the shape but also likely to align feature space
- Cons:
  - Need alignment
  - Need supervision

- Related work
  - STEMM: Self-learning with Speech-text Manifold Mixup for Speech Translation (Fang et al. 2022)



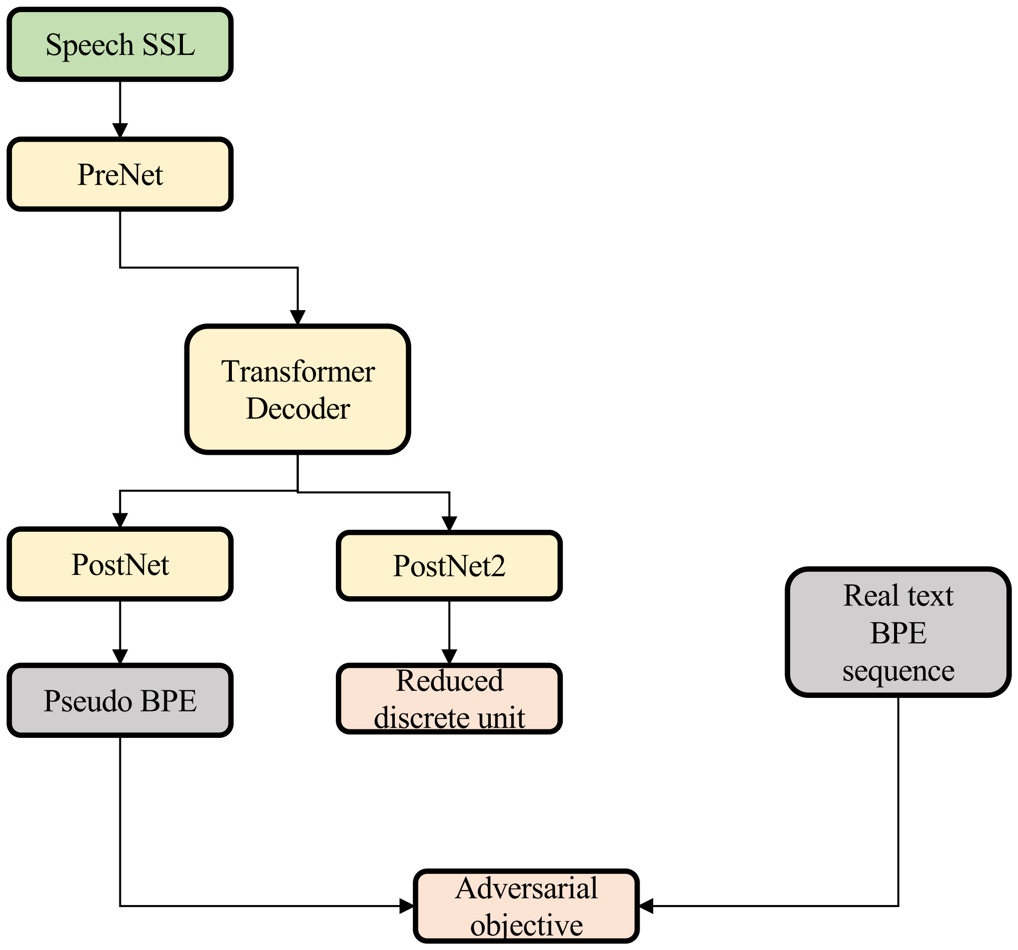
STEMM: Self-learning with Speech-text Manifold Mixup for Speech Translation (Fang et al. 2022)

Our proposal

- Summary of the issues
  - Supervision is necessary for previous method
  - Lack of flexibility because of adopting fixed compression rate over time domain

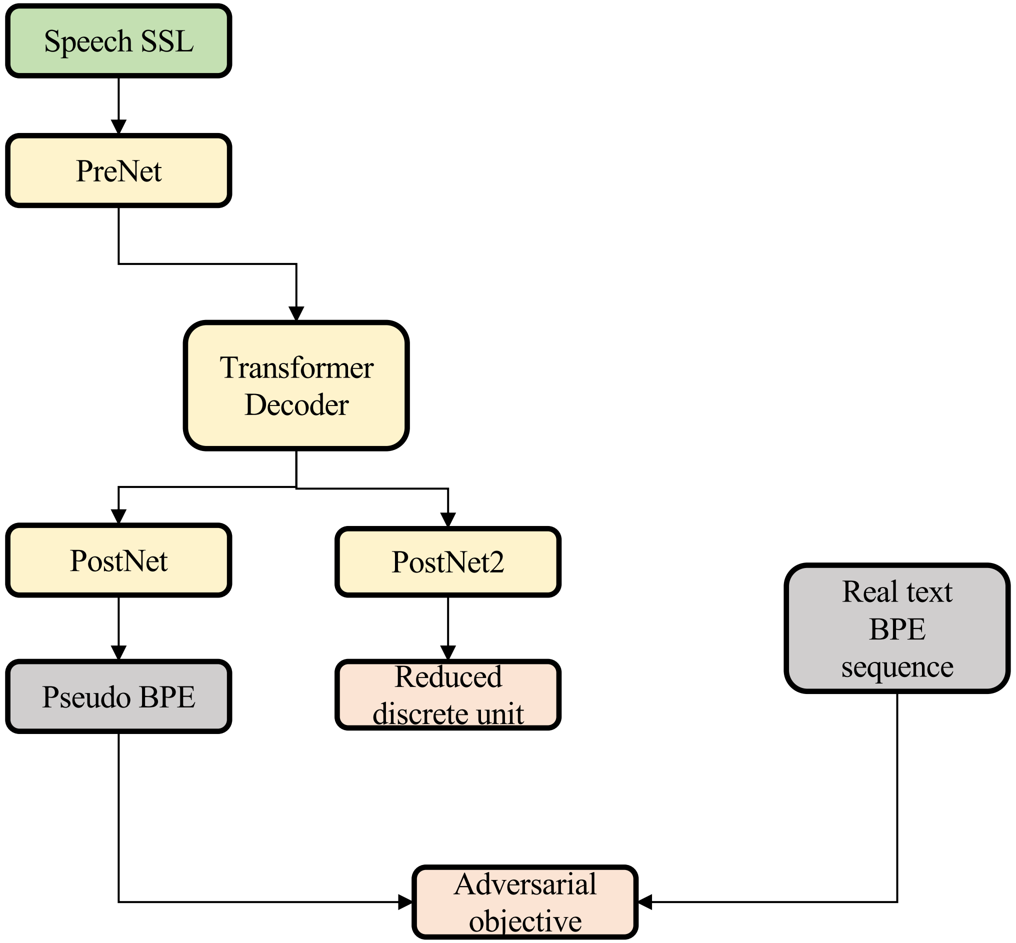
- Our proposal
  - Refine speech self-supervised features with some text flavors in unsupervised mannor
  - Introduce more flexibility by variable compression over time domain

Our proposal (Cont'd)



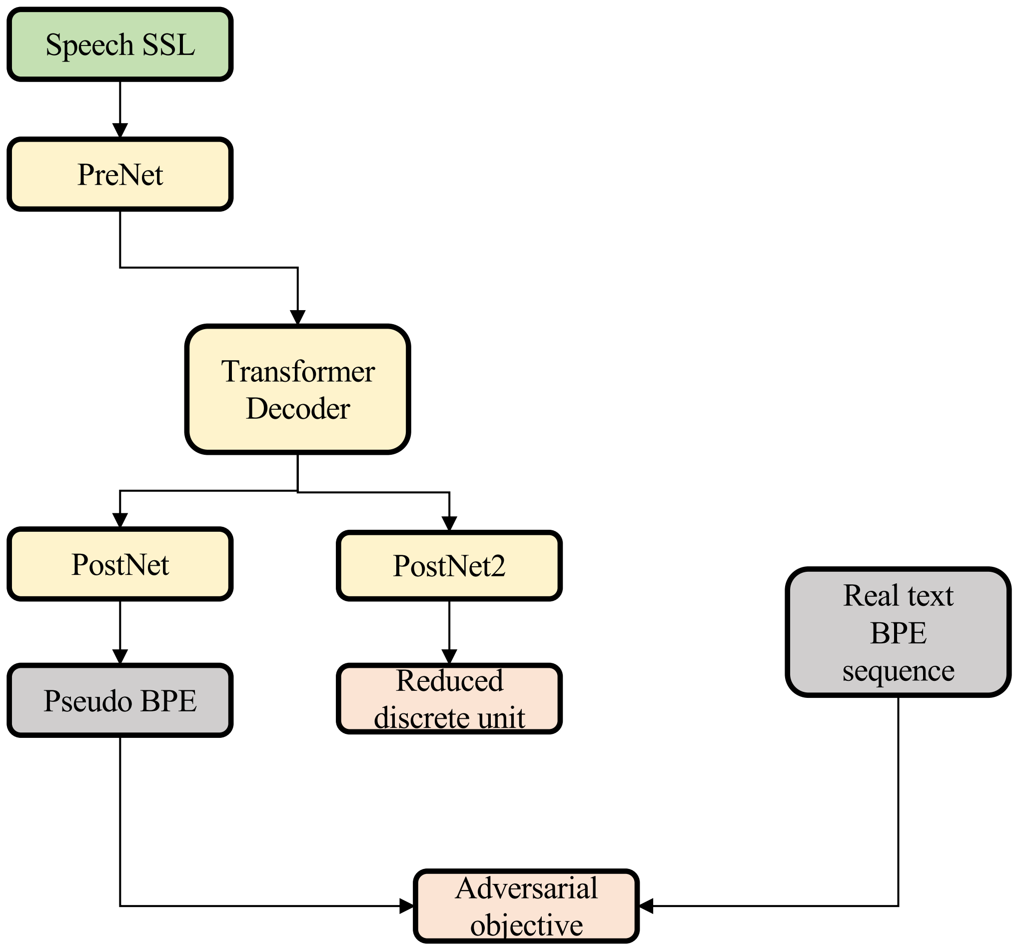
- Adapted from wav2vec-U 2.0 (https://arxiv.org/pdf/2204.02492.pdf)
  - Use text bpe instead phoneme to better align with text
  - Use source-target attention to allow flexible alignment

Our proposal (Cont’d)



- Yellow blocks: training network
  - PreNet: transformer/cnn layers
  - TransformerDecoder: either BART-like pre-trained or random initialed decoder
  - PostNet: CNN subsampling layers
  - PostNet2: transformer/cnn layers

Our proposal (Cont'd)



- Orange blocks: objectives
  - Adversarial objective (for unsupervised ASR training)
  - Reduced discrete unit (for time-domain compression):
    - First use K-means cluster (uniqued)
    - Iteratively update discrete unit (by reclustering hidden states from PostNet) to optimize the alignment

Application for the proposed framework

- Unsupervised ASR
  - As the training is adopted from wav2vec-U 2.0, the framework can be directly use for unsupervised ASR training
- Downstream task
  - The framework can be applied as an adapter function to compress speech SSL features into textual space, which could be used for semantic downstream cases)
  - We prepared to first focus on speech-to-text translation in SUPERB benchmark ⍰ then for speech-to-speech translation if possible

Plan

- Baseline (wav2vec-U and wav2vec-U 2.0) in mid-May
  - Add unsupervised ASR to superb benchmark
- Base Framework for the proposal in mid-June
- Intensive experiments during JSALT
  - Verify the results with larger corpora
  - Explore combination with efforts from other directions (e.g., model compression, sequence compression, multilingual)
- Prepare the work as a submission to confs