Combine Speech&Text SSL

Presentor: Jiatong Shi
Joint work with Ann Lee, Dongji Gao, Shinji Watanabe, Hung-yi Lee

Content

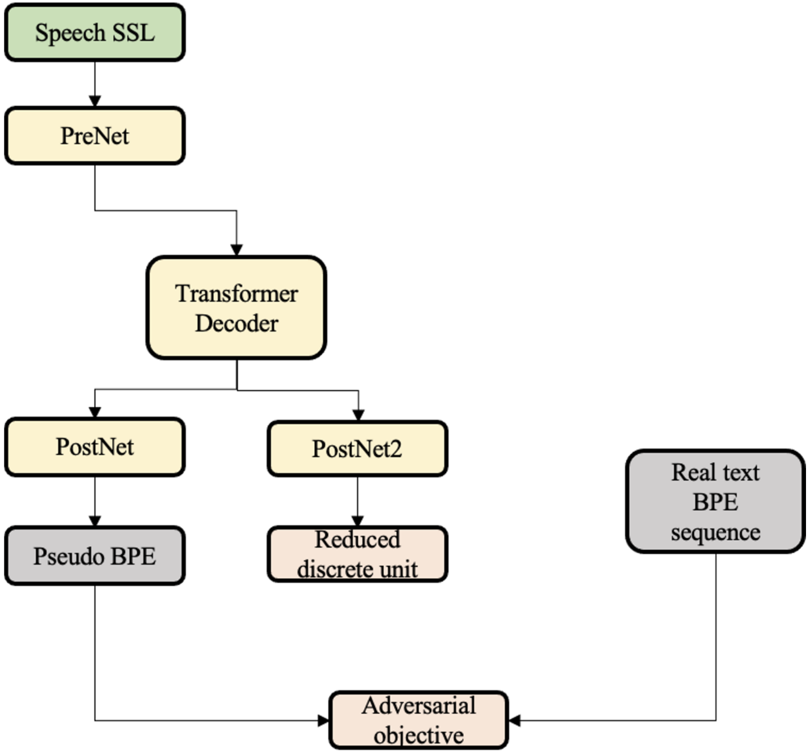
1. Recap
2. Implementation of wav2vec-u and experimental results
3. Implementation of wav2vec-u 2.0 and experimental results
4. Our proposed sequence-to-sequence model
    a. Challenges
    b. Possible solutions
5. Timeline and downstream tasks

Background Recap

- [Prior] Different modalities has different features (e.g., information, information density, length, context)
- [Target] A better framework to utilize both speech and text pre-trained models for downstream semantic tasks
- [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?

Proposal Recap

- Refine speech self-supervised features with some text flavors in unsupervised manner
    - By train with text BPE sequence
    - By use unsupervised ASR framework to learn from non-parallel data
- Introduce more flexibility by variable compression over time domain
    - Add sequence-to-sequence property in the framework to allow flexible down-sampling
- Focus on ASR task first

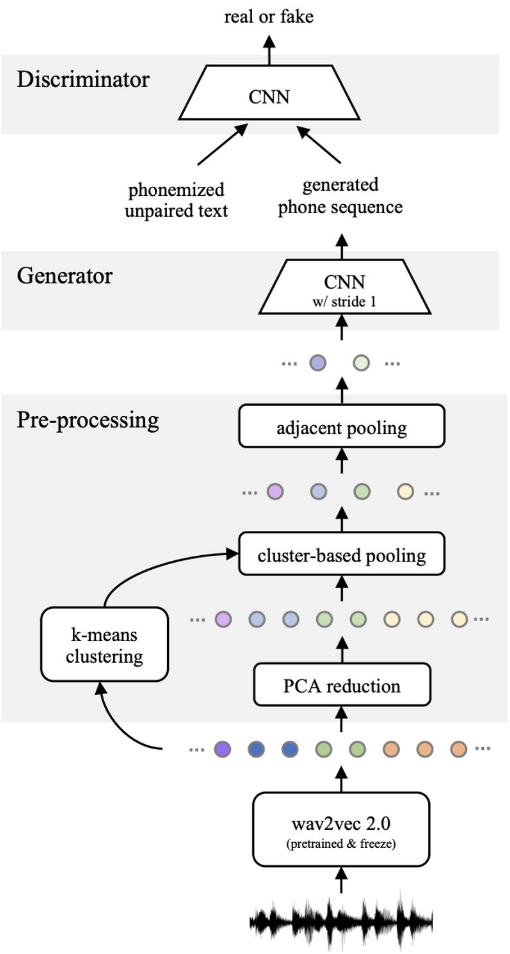


Research Plan

1. Implement & reproduce wav2vec-u/wav2vec-u 2.0
    a. Baseline of this work
    b. Today's presentation: share experiences from implementation and experimentation
2. Explore combining wav2vec 2.0 and pre-trained mBART
    a. The proposal
    b. Today's presentation: share current challenges and proposed solutions, discuss and get feedback
3. Integration to S3PRL and additional task in unsupervised ASR
    a. Pending for workshop sessions

Wav2vec-U

- Use wav2vec2 as feature extractor
- Apply preprocessing (i.e., PCA, Kmeans-pooling, adjacent pooling) to downsample features
- Use single layer CNN for the generator



Wav2vec-u Experiments (with Librispeech-100)

Key findings:

Some factors are crucial to good convergence:

1. Layer for feature extraction -> 7, 14 are the best, layer combination cannot always converge
2. Network simplicity -> For example, adding two layer CNN would harm the results (+10-20 PER or not converge); Layer combination sometimes also hurt results (+10 PER)

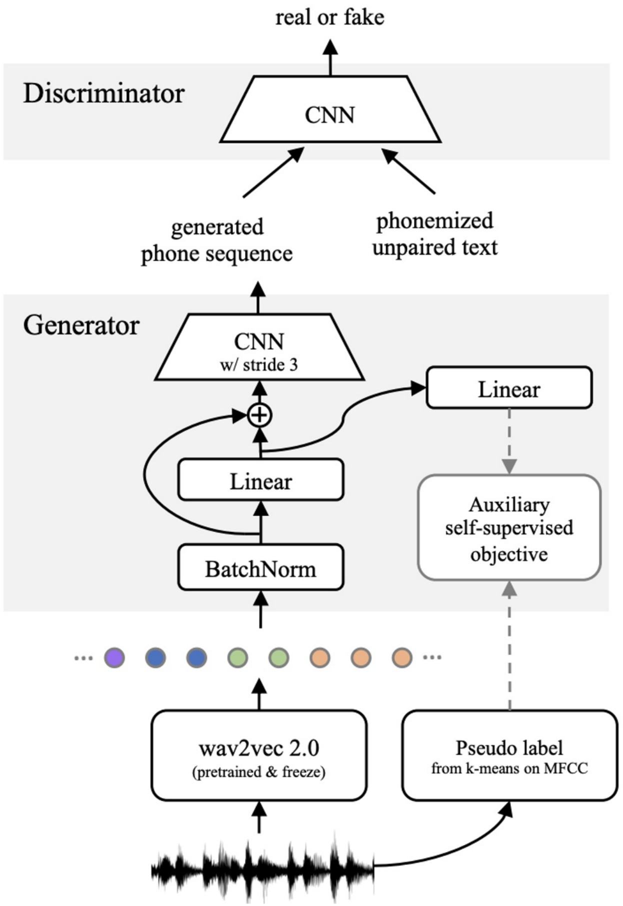
Some factors are good to tune

1. Preprocessing parameters: cluster number fo Kmean pooling (K=128, K=256, K=64) and adjacent pooling
2. Training parameters: learning rates, weights for losses (for gradient penalty, phoneme diversity, and others)

Our best PER results with wav2vec-u after tuning on Librispeech-100 is 24.1%

Wav2vec-U 2.0

- Use an Batchnorm to replace the preprocessing
- Add K-means cluster objectives to stable the results
- Use CNN with stride to conduct downsampling



Wav2vec-U 2.0 Experiments (Librispeech100)

Key factor for convergence:

- Batchnorm with scaling factor + large batch size
  - Standard scaling factor 1.0 does not suitable for wav2vec2 feature (might different for other ssls?) -> get 20+PER or non-converge
  - Large batch size is necessary to get reasonable performances -> get non-converge results with small batch size like 10
- Network simplicity
  - Similar to wav2vec-u 1.0, cannot hold very large network -> e.g., even additional layer of CNN
  - But can be mitigate / even get improvements by adding auxiliary losses (e.g., K-means clustering as prediction target)
    - If add additional CNN with auxiliary loss to regularize, the results can be maintained or even a bit better
- Layer selection
  - Layer combination not get better results but more un-stability

Results with same method (without auxiliary loss): 23.3 PER

Results with same methods in the paper: 22.6 PER

Best results for now: add additional CNN layer with auxiliary loss 22.3 PER

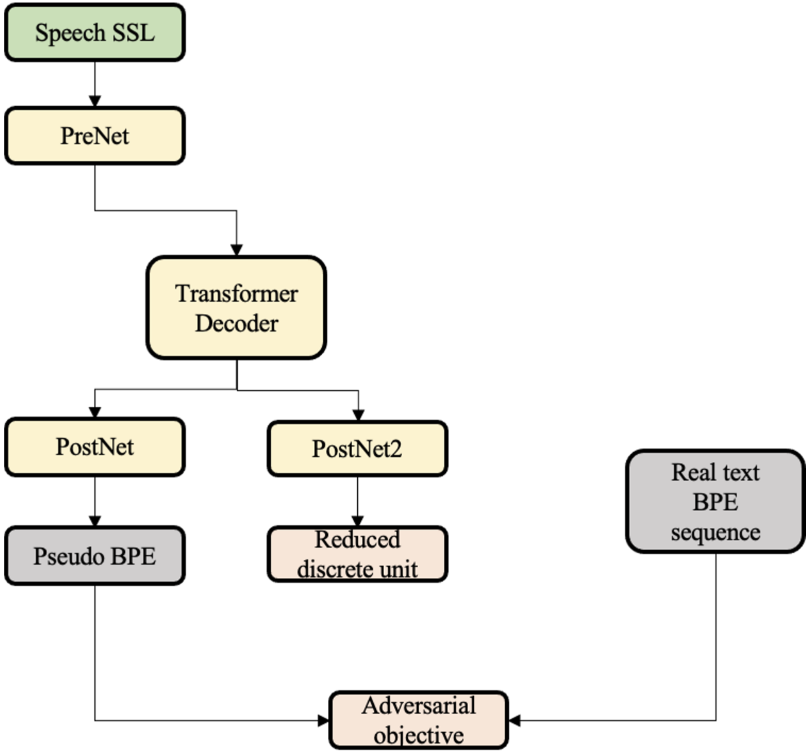
Next Steps for wav2vec-U

- Continue iterating on wav2vec-U 2.0 (more variations with other networks, e.g., transformer layers)
- S3prl upstream/downstream
- SUPERB
  - A speech SSL model with extra (unsupervised) information from text!
- Any other applications?

Combining Speech & Text SSL – Current Proposal

Given the previous experiments with wav2vec-u 1.0/2.0, we need to revisit our current proposal with following factors:

- Prediction of textual BPE
- Use a seq2seq model guided with reduced discrete unit



Combining Speech & Text SSL – Challenges & Solutions?

- Will textual BPE work for the framework?
  - We currently tried a 500-BPE model as prediction target for librispeech-100, and cannot get convergence
  - Various reasons
    - BPE size can be tuned -> experiments with more bpe size
    - Fixed downsample in wav2vec-u 2.0 might not work for BPE -> use our proposed framework might help
  - For current solutions, we will still focus on phoneme-based task for start, leaving it for future investigation

Combining Speech & Text SSL – Challenges & Solutions?

- Can a more complex model get good convergence than the simple CNN?
  - Findings:
    - Seq2seq models are difficult to train especially in unsupervised setting
    - It may also get trival predictions for adversarial objectives
  - Possible solution:
    - Adding reduced unit sequence
    - Add auxiliary loss over other layers (e.g., prenet with MFCC kmeans)
    - Use a pre-trained seq2seq model
      - Option1: BART with input as features, output as discrete units
      - Option2: Self-training with input as features, output as BPE discrete units
      - Noted the discrete units here can be clusters from other ssl or features (e.g., Hubert, MFCC, etc.)

Next Steps for Combining Speech & Text SSL

1. Further investigation on wav2vec-u 2.0
2. Prepare wav2vec-u 2.0 as a SSL model in s3prl for semantic tasks (1.0 version is difficult to deploy with too much preprocessing procedures)
    a. Speech translation
    b. Spoken language understanding
    c. Unsupervised TTS (maybe not in s3prl)
3. Refine, iteration on our proposed method for unsupervised ASR
    a. Study of auxiliary loss for our proposed method: different ssl models as objectives
    b. Study of pre-trained seq2seq model in unsupervised ASR
    c. Study of possibility of using BPE in unsupervised ASR
4. Experiments on large scale data with Librispeech-960

Timeline

By Jun 27: publish our pre-trained wav2vec-u 2.0 feature extractor (on Librispeech360) to https://github.com/JSALT-2022-SSL/s3prl_unsupervised_combine_ssl (with corresponding training code)

By Jul. 14: Benchmarking the semantic tasks performances with the feature extractor

At Jul. 14-28 depends on the progress: Publish wav2vec-u 2/0 feature extractor (Librispeech960) and our proposed method extractor (Librispeech360)

By Aug. 5: Benchmarking semantic tasks performances with both feature extractors