Improved Consistency Training for Semi-Supervised Sequence-to-Sequence ASR via Speech Chain Reconstruction and Self-Transcribing

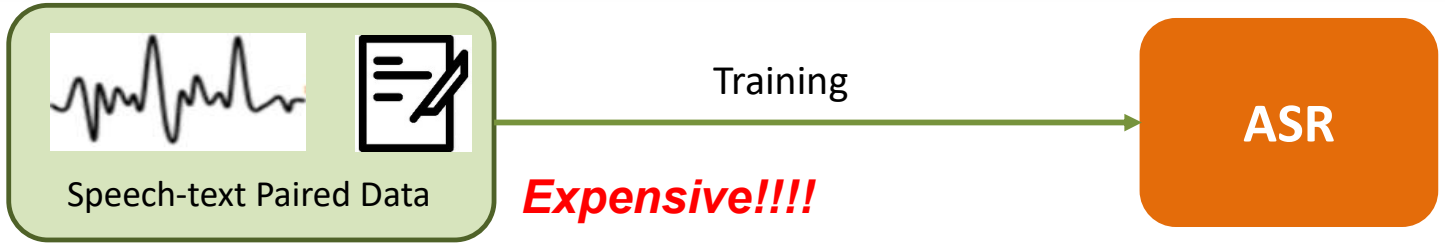
Heli Qi[1], *Sashi Novitasari*[1], *Sakriani Sakti*[2], *Satoshi Nakamura*[1]

1. Nara Institute of Science and Technology, Japan
2. Japan Advanced Institute of Science and Technology, Japan

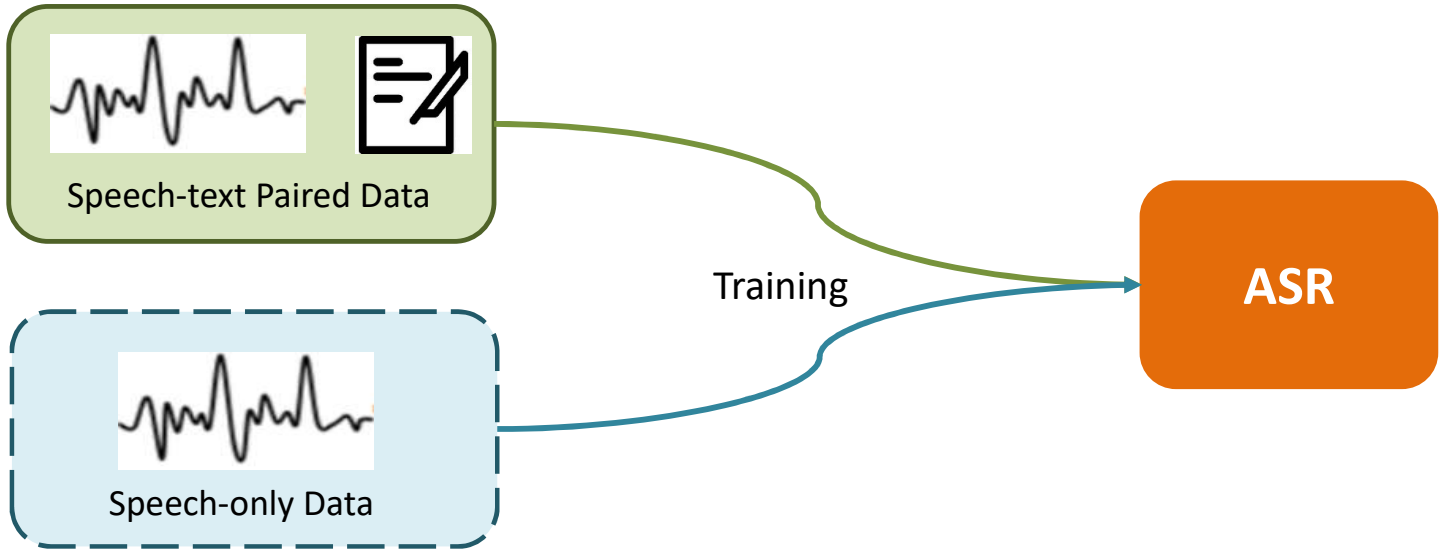
Table of Contents

Research Background—Semi-supervised ASR

Supervised ASR training with **speech-text paired data**



Semi-supervised ASR training with both **speech-text paired data** and **speech-only data**



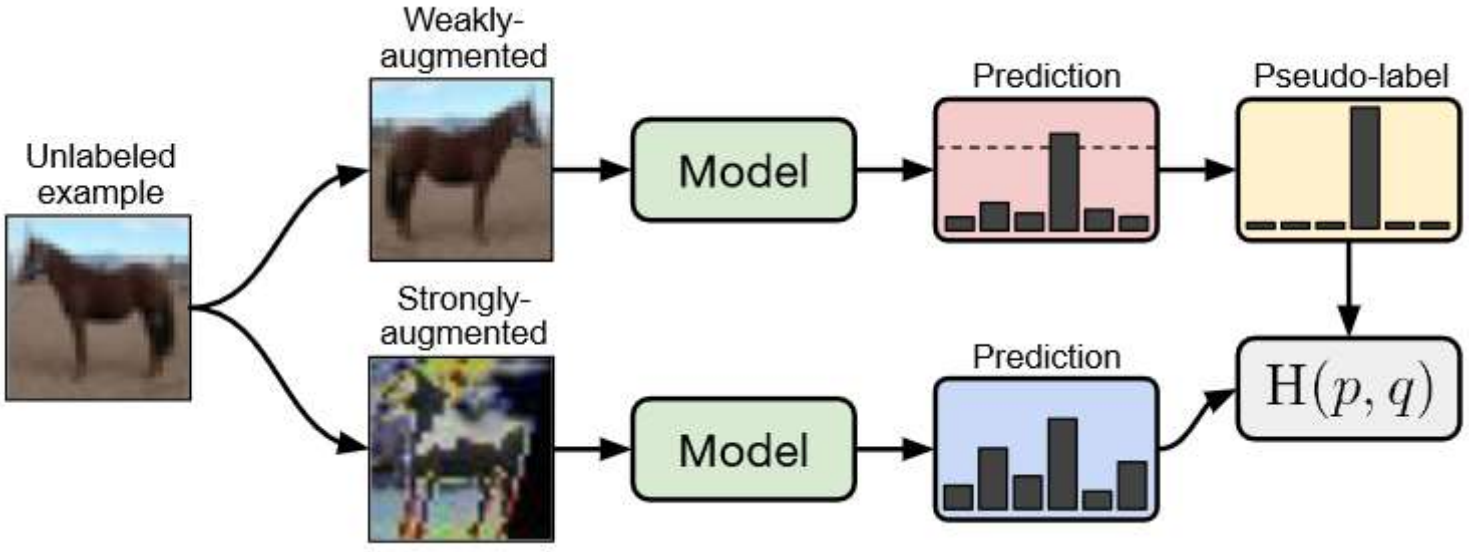
Related Work—Semi-supervised ASR Strategy

- Self-training **[Jacob et al. ICASSP2020]**
- Iterative Self-training **[Qiantong et al. Interspeech2020]**
- Noisy Student Training **[Daniel S. et al. Interspeech2020]**
- Consistency Regularization **[Felix et al. Interspeech2020]**
- ……

Jacob et al. ICASSP2020, “Self-training for end-to-end speech recognition”
Qiantong et al. Interspeech2020, “Iterative Pseudo-Labeling for Speech Recognition”
Daniel S. et al. Interspeech2020, “Improved Noisy Student Training for Automatic Speech Recognition”
Felix et al. Interspeech2020, “Semi-Supervised Learning with Data Augmentation for End-to-End ASR”

Related Work—Consistency Regularization

FixMatch algorithm **[K. Sohn et al. NIPS2020]**



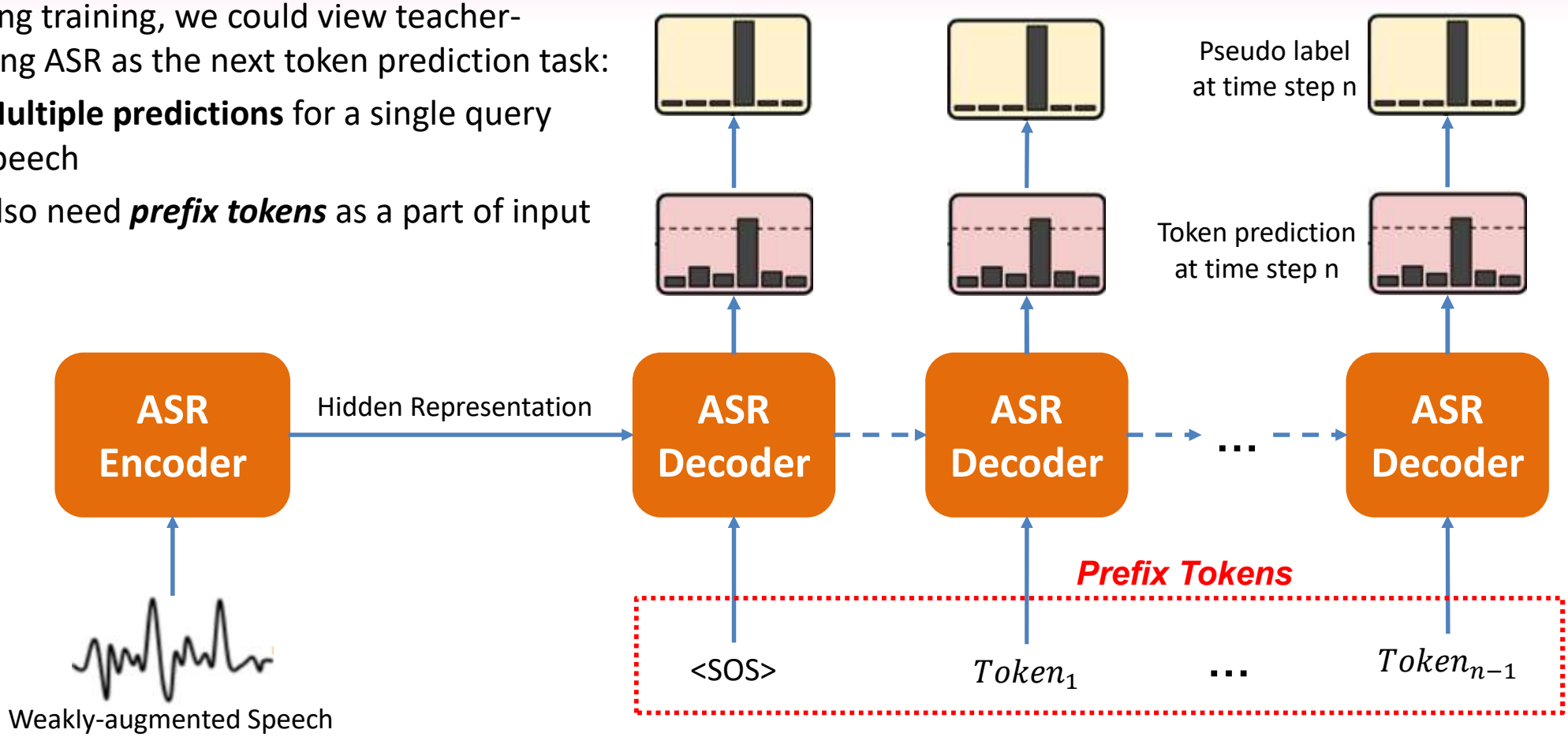
- Designed for semi-supervised image classification (IC)
- Make pseudo labels by the weakly-augmented image
- Train the model by the strongly-augmented image and the acquired pseudo label.

K. Sohn et al. NIPS2020, “Fixmatch: Simplifying semi-supervised learning with consistency and confidence”

Handle the difference between IC and S2S ASR

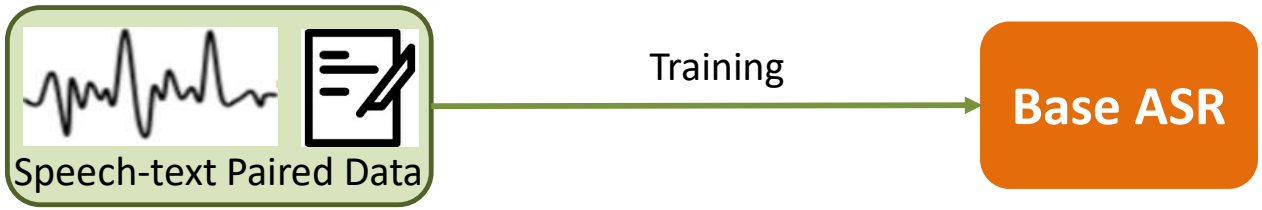
During training, we could view teacher-forcing ASR as the next token prediction task:

- **Multiple predictions** for a single query speech
- Also need ***prefix tokens*** as a part of input

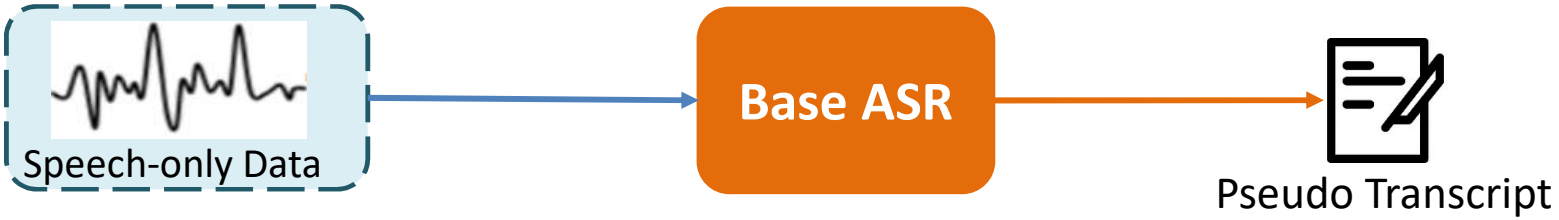


Semi-supervised ASR Training Paradigm

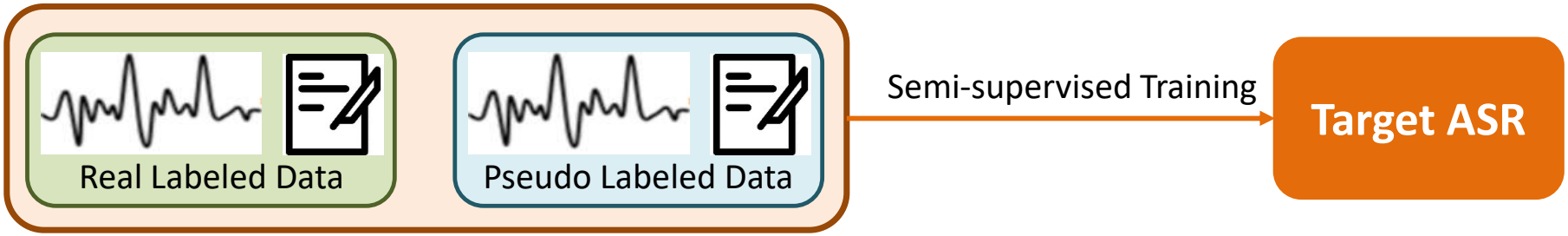
1. Base ASR Training on labeled data



2. Pseudo Transcript Generation

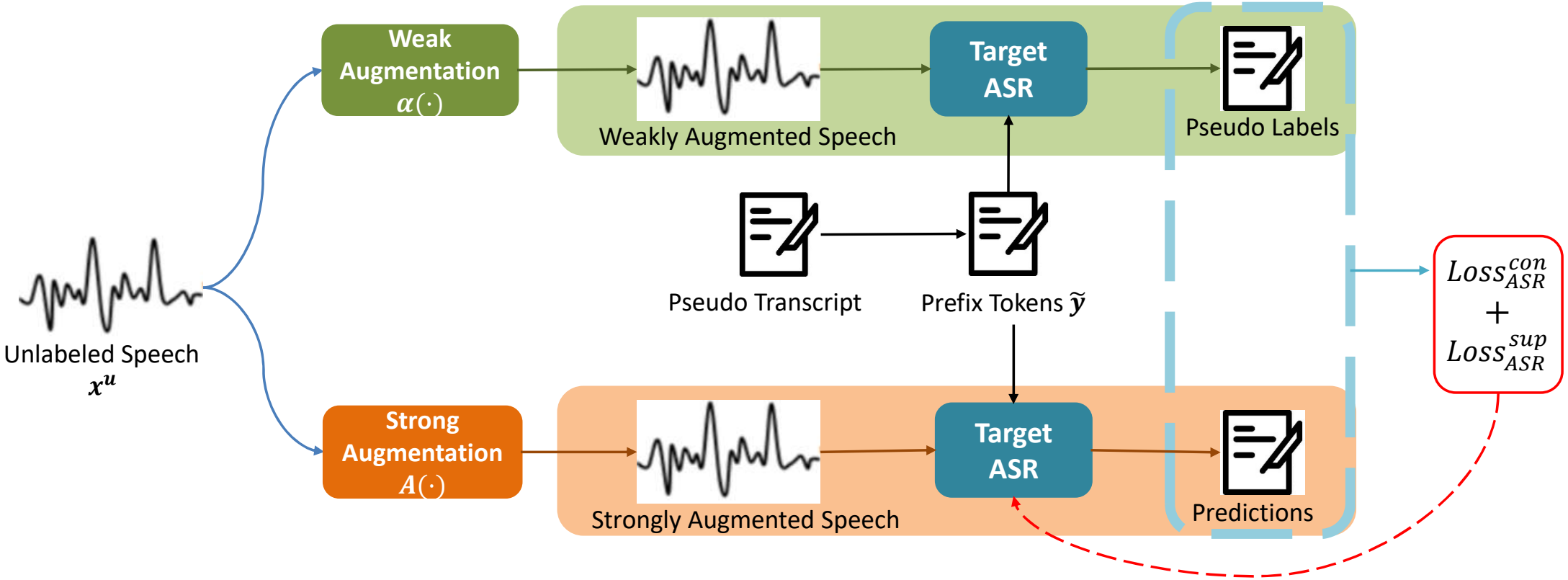


3. Semi-supervised Training on the enlarged dataset



Fixmatch-based Semi-supervised ASR Training

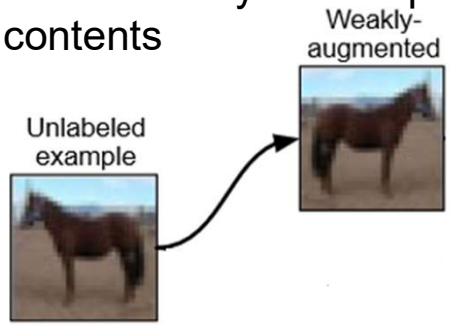
- Pseudo transcript generated by the base ASR model acts as prefix tokens at each time step
- SpecAugment [Daniel S. et al. Interspeech2019] is adopted as both weak augmentation and strong augmentation



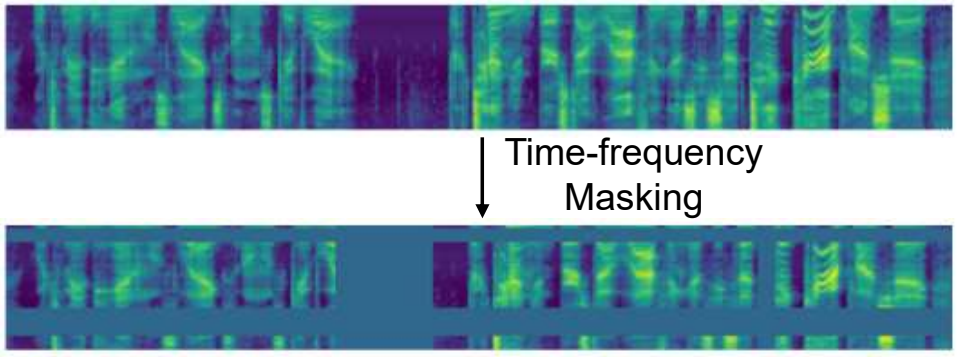
Daniel S. et al. Interspeech2019, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition"

Existing Problems to Be Solved

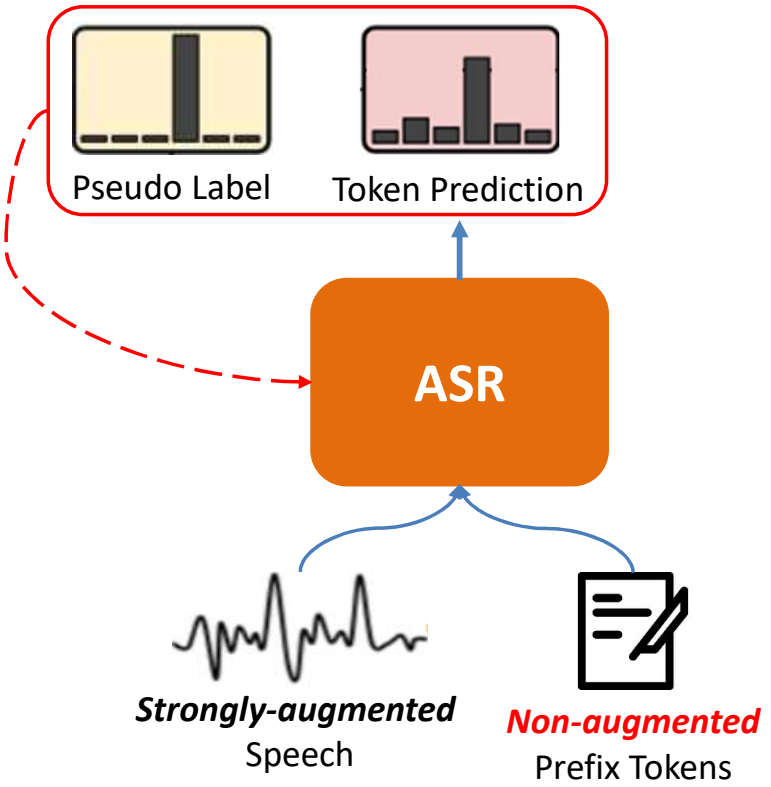
- Ideal weak augmentation for high quality pseudo labels
  - Modify the acoustic style while preserving the linguistic contents



- Hard to decide the masking width
  - Too narrow → Meaningless augmentation
  - Too wide → Hurt the linguistic contents



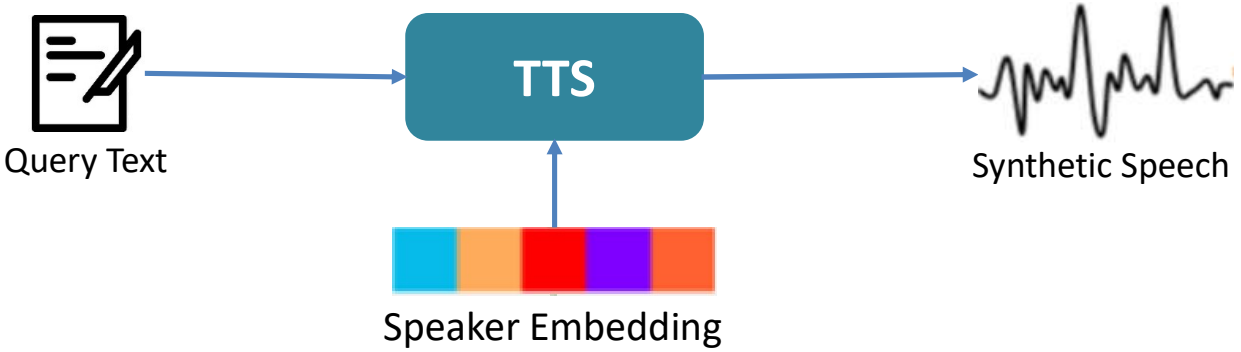
ASR Training by Teacher-Forcing:



As a part of input, the augmentation on prefix tokens also need to be considered.

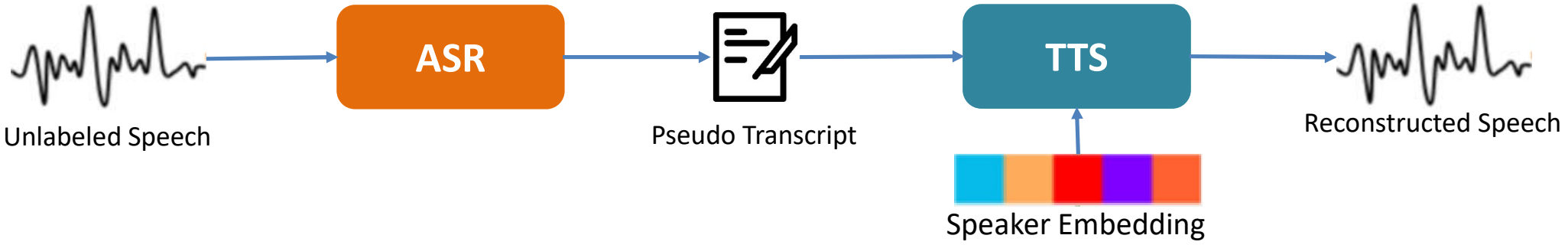
Speech Chain Reconstruction

Text-to-Speech Synthesis (TTS)



Speech Chain Reconstruction [A. Tjandra et al. ASRU2017]

Complete Linguistic Content



A. Tjandra et al. ASRU2017, "Listening while speaking: Speech chain by deep learning"

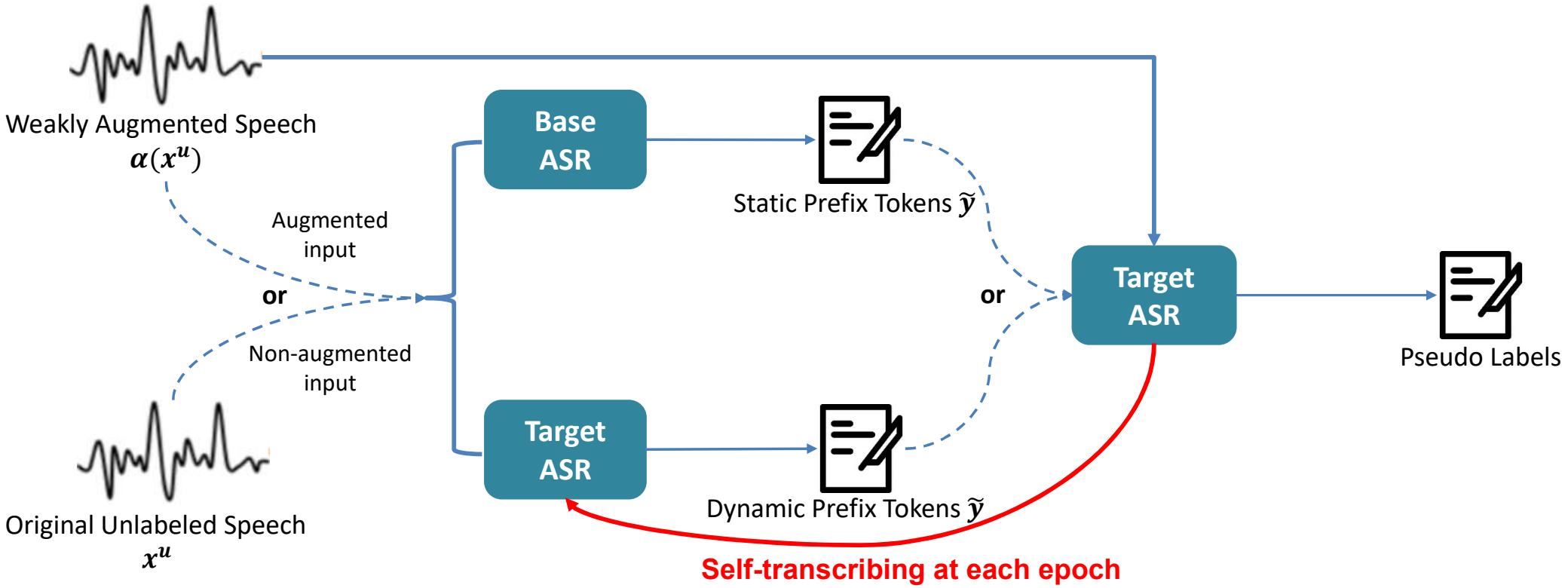
Different Prefix Token Production Strategies

Static prefix tokens $\tilde{\boldsymbol{y}}$ produced by original speech $\boldsymbol{x^u}$

Static prefix tokens $\tilde{\boldsymbol{y}}$ produced by weakly-augmented speech $\boldsymbol{\alpha(x^u)}$

Dynamic prefix tokens $\tilde{\boldsymbol{y}}$ produced by original speech $\boldsymbol{x^u}$

Dynamic prefix tokens $\tilde{\boldsymbol{y}}$ produced by weakly-augmented speech $\boldsymbol{\alpha(x^u)}$



Experiment Setups

Datasets:

1. LJSpeech (Single Speaker):
   - **Labeled data:** 6300 utterances
   - **Unlabeled data:** 6300 utterances
2. LibriSpeech-100h (Multiple speakers):
   - **Labeled data:** 8750 utterances (75 speakers)
   - **Unlabeled data:** 19968 utterances (176 speakers)

Mode Input & Output:

1. Acoustic Features:
   - 16,000 sampling rate
   - 50ms frame length & 12.5ms frame shift
   - 80d log Mel-spectrogram
2. Tokenization:
   - Character-based models
   - 26 English letters (a~z) + 3 special tokens (apostrophes, space, and “sos/eos”)

ASR model:

1. Encoder:
   - Single-Speaker Setting: 3 Bi-LSTM layers (2 * 256 dim)
   - Multi-Speaker Setting: 5 Bi-LSTM layers (2 * 256 dim)
2. Decoder:
   - 1 LSTM layer (512 dim)
   - Additive Attention

TTS model:

- Same structure as Tacotron2 **[J. Shen et al. ICASSP2018]**
- X-vector **[D. Snyder et al. ICASSP2018]** for speaker embedding

J. Shen et al. ICASSP2018, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions”
D. Snyder et al. ICASSP2018, “X-vectors: Robust dnn embeddings for speaker recognition”

Experiment Results

Contrast Experiments:

1. **Weak Augmentation $\alpha(\cdot)$:**
   - SpecAugment with narrow masking width (Weak SpecAugment)
   - Speech Chain Reconstruction
2. **Strong Augmentation $A(\cdot)$:**
   SpecAugment with wide masking width
3. **Different Pseudo Labeling Threshold τ:**
   The larger τ is, the less pseudo labels used for training

Experiment Conclusions:

- Speech Chain Reconstruction outperforms Weak SpecAugment in all scenarios.
- Static prefix tokens generate by $\alpha(x^u)$ benefit semi-supervised ASR training a lot
- No large improvement of dynamic prefix tokens has been observed

CER Results

(Red boxes represent the best performance in each scenario)

$\alpha(\cdot)$	LJSpeech τ=0.5	τ=0.6	τ=0.7	τ=0.8	τ=0.9	LibriSpeech τ=0.5	τ=0.7	τ=0.9
Supervised Baseline								
–	8.2	8.2	8.2	8.2	8.2	28.0	28.0	28.0
Static $\tilde{y}$ produced by x^u (the existing paradigm [5])								
Weak SpecAugment	8.3	7.8	7.7	7.5	7.7	18.3	19.6	20.8
Speech Chain Reconstruction	7.9	7.6	7.4	7.5	7.8	18.2	19.8	18.5
Static $\tilde{y}$ produced by $\alpha(x^u)$								
Weak SpecAugment	7.8	7.7	7.7	7.8	7.6	18.8	19.6	20.3
Speech Chain Reconstruction	7.9	7.7	7.2	7.2	7.6	17.2	18.4	18.3
Dynamic $\tilde{y}$ produced by x^u								
Weak SpecAugment	7.9	7.9	7.6	7.4	7.6	19.1	19.1	19.9
Speech Chain Reconstruction	8.2	7.2	7.5	7.4	7.6	18.1	18.4	18.4
Dynamic $\tilde{y}$ produced by $\alpha(x^u)$								
Weak SpecAugment	7.5	7.4	8.1	7.6	8.0	19.8	20.5	18.9
Speech Chain Reconstruction	7.7	7.7	7.6	7.4	7.2	20.0	19.1	18.5

Conclusion & Future Work

Conclusion:

1. Speech Chain Reconstruction protects the linguistic information of the speech after augmentation.
2. As a part of ASR input, prefix tokens also need augmentation for the application of consistency regularization.
3. Updating prefix tokens during training need more smart designs to better evaluate its effectiveness, such as updating interval.

Future work:

1. Move from RNN-based ASR to Transformer-based ASR models.
2. Explore other semi-supervised ASR training strategies.
3. Conduct experiments on more challenging datasets, e.g. large-scale speech datasets, noisy speech dataset, and so on.

Thank you very much for listening!

Really appreciate your patience so far!