Speech Recognition Papers

List of hot directions in industrial speech recognition, i.e., **Streaming ASR** (RNA-based || RNN-T based || Attention based || unified streaming/non-streaming) / **Non-autoregressive ASR** ...

If you are interested in this repo, any pull request is welcomed.

Streaming ASR

RNA based

- Standard RNA: Recurrent Neural Aligner: An Encoder-Decoder Neural Network Model for Sequence to Sequence Mapping (Interspeech 2017)
- Extended RNA: Extending Recurrent Neural Aligner for Streaming End-to-End Speech Recognition in Mandarin (Interspeech 2018)
- Transformer equipped RNA: Self-attention Aligner: A Latency-control End-to-end Model for ASR Using Self-attention Network and Chunk-hopping (ICASSP 2019)
- CIF: CIF: Continuous Integrate-And-Fire for End-To-End Speech Recognition (ICASSP 2020)
- CIF: A Comparison of Label-Synchronous and Frame-Synchronous End-to-End Models for Speech Recognition (Interspeech 2020)

RNN-T based

- Standard RNN-T: Streaming E2E Speech Recognition For Mobile Devices (ICASSP 2019)
- Latency Controlled RNN-T: RNN-T For Latency Controlled ASR With Improved Beam Search (arXiv 2019)
- Transformer equipped RNN-T: Self-Attention Transducers for End-to-End Speech Recognition (Interspeech 2019)
- Transformer equipped RNN-T: Transformer Transducer: A Streamable Speech Recognition Model With Transformer Encoders And RNN-T Loss (ICASSP 2020)
- Transformer equipped RNN-T: A Streaming On-Device End-to-End Model Surpassing Server-Side Conventional Model Quality and Latency (ICASSP 2020)
- Tricks for RNN-T Training: Towards Fast And Accurate Streaming E2E ASR (ICASSP 2020)

- Knowledge Distillation for RNN-T: Knowledge Distillation from Offline to Streaming RNN Transducer for End-to-end Speech Recognition (Interspeech 2020)
- Transfer Learning for RNN-T: Transfer Learning Approaches for Streaming End-to-End Speech Recognition System (Interspeech 2020)
- Exploration on RNN-T: Analyzing the Quality and Stability of a Streaming End-to-End On-Device Speech Recognizer (Interspeech 2020)
- Sequence-level Emission Regularization for RNN-T: FastEmit: Low-latency Streaming ASR with Sequence-level Emission Regularization (arXiv 2020, submitted to ICASSP 2021)
- Model Distillation for RNN-T: Improving Streaming Automatic Speech Recognition With Non-Streaming Model Distillation On Unsupervised Data (arXiv 2020, submitted to ICASSP 2021)
- LM Fusion for RNN-T: Improved Neural Language Model Fusion for Streaming Recurrent Neural Network Transducer (arXiv 2020, submitted to ICASSP 2021)
- Normalized jointer network: Improving RNN transducer with normalized jointer network (arXiv 2020)
- Benchmark on RNN-T CTC LF-MMI: Benchmarking LF-MMI, CTC and RNN-T Criteria for Streaming ASR (SLT 2021)
- Alignment Restricted RNN-T: Alignment Restricted Streaming Recurrent Neural Network Transducer (SLT 2021)
- Conformer equipped RNN-T (with Cascaded Encoder and 2nd-pass beam search): A Better and Faster End-to-End Model for Streaming ASR (arXiv 2020, submitted to ICASSP 2021)
- Multi-Speaker RNN-T: Streaming end-to-end multi-talker speech recognition

Attention based

- Montonic Attention: Montonic Chunkwise Attention (ICLR 2018)
- Enhanced Montonic Attention: Enhancing Monotonic Multihead Attention for Streaming ASR (Interspeech 2020)
- Minimum Latency Training based on Montomic Attention: Minimum Latency Training Strategies For Streaming seq-to-seq ASR (ICASSP 2020)
- Triggered Attention: Triggered Attention for End-to-End Speech Recognition (ICASSP 2019)
- Triggered Attention for Transformer: Streaming Automatic Speech Recognition With The Transformer Model (ICASSP 2020)
- Block-synchronous: Streaming Transformer ASR with Blockwise Synchronous Inference (ASRU 2019)
- Block-synchronous with chunk reuse: Transformer Online CTC/Attention E2E Speech Recognition Architecture (ICASSP 2020)
- Block-synchronous with RNN-T like decoding rule: Synchronous Transformers For E2E Speech Recognition (ICASSP 2020)
- Scout-synchronous: Low Latency End-to-End Streaming Speech Recognition with a Scout Network (Interspeech 2020)
- CTC-synchronous: CTC-synchronous Training for Monotonic Attention Model (Interspeech 2020)
- Memory Augmented Attention: Streaming Transformer-based Acoustic Models Using Self-attention with Augmented Memory (Interspeech 2020)
- Memory Augmented Attention: Streaming Chunk-Aware Multihead Attention for Online End-to-End Speech Recognition (Interspeech 2020)
- Optimized Beam Search: High Performance Sequence-to-Sequence Model for Streaming Speech Recognition (Interspeech 2020)
- Memory Augmented Attention: Emformer: Efficient Memory Transformer Based Acoustic Model For Low Latency Streaming Speech Recognition (arXiv 2020, submitted to ICASSP 2021)

Unified Streaming/Non-streaming models

- Transformer Transducer: One Model Unifying Streaming And Non-Streaming Speech Recognition (arXiv 2020)
- Universal ASR: Unify And Improve Streaming ASR With Full-Context Modeling (ICLR 2021 under double-blind review)
- Cascaded encoders for unifying streaming and non-streaming ASR (arXiv 2020)
- Asynchronous Revision for non-streaming ASR: Dynamic latency speech recognition with asynchronous revision (arXiv 2020, submitted to ICASSP 2021)
- 2-pass unifying (1st Streaming CTC, 2nd Attention Rescore): Unified Streaming and Non-streaming Two-pass End-to-end Model for Speech Recognition
- 2-pass unifying (1st Streaming CTC, 2nd Attention Rescore): One In A Hundred: Select The Best Predicted Sequence from Numerous Candidates for Streaming Speech Recognition (arXiv 2020)

Non-autoregressive (NAR) ASR

- MASK-Predict: Listen and Fill in the Missing Letters: Non-Autoregressive Transformer for Speech Recognition (arXiv 2019)
- Imputer: Imputer: Sequence modelling via imputation and dynamic programming (arXiv 2020)
- Insertion-based: Insertion-Based Modeling for End-to-End Automatic Speech Recognition (arXiv 2020)
- MASK-CTC: Mask CTC: Non-Autoregressive End-to-End ASR with CTC and Mask Predict (Interspeech 2020)
- Spike Triggered: Spike-Triggered Non-Autoregressive Transformer for End-to-End Speech Recognition (Interspeech 2020)
- Similar to MASK-Predict: Listen Attentively, and Spell Once: Whole Sentence Generation via a Non-Autoregressive Architecture for Low-Latency Speech Recognition (Interspeech 2020)
- Improved MASK-CTC: Improved Mask-CTC for Non-Autoregressive End-to-End ASR (arXiv 2020, submitted to ICASSP 2021)
- Refine CTC Alignments over Latent Space: Align-Refine: Non-Autoregressive Speech Recognition via Iterative Realignment (arXiv 2020)
- Also Refine CTC Alignments over Latent Space: CASS-NAT: CTC Alignment-based Single Step Non-autoregressive Transformer for Speech Recognition (arXiv 2020, submitted to ICASSP 2021)
- Refine CTC Alignments over Output Space: Non-Autoregressive Transformer ASR with CTC-Enhanced Decoder Input (arXiv 2020, submitted to ICASSP 2021)

ASR Rescoring / Spelling Correction (2-pass decoding)

- Review: Automatic Speech Recognition Errors Detection and Correction: A Review (N/A)
- LAS based: A Spelling Correction Model For E2E Speech Recognition (ICASSP 2019)
- Transformer based: An Empirical Study Of Efficient ASR Rescoring With Transformers (arXiv 2019)
- Transformer based: Automatic Spelling Correction with Transformer for CTC-based End-to-End Speech Recognition (Interspeech 2019)
- Transformer based: Correction of Automatic Speech Recognition with Transformer Sequence-To-Sequence Model (ICASSP 2020)
- BERT based: Effective Sentence Scoring Method Using BERT for Speech Recognition (ACML 2019)

- BERT based: Spelling Error Correction with Soft-Masked BERT (ACL 2020)
- Parallel Rescoring: Parallel Rescoring with Transformer for Streaming On-Device Speech Recognition (Interspeech 2020)

On-device ASR

- Review: A review of on-device fully neural end-to-end automatic speech recognition algorithms (arXiv 2020)
- Lightweight Low-Rank transformer: Lightweight and Efficient End-to-End Speech Recognition Using Low-Rank Transformer (ICASSP 2020)
- Attention replacement: How Much Self-Attention Do We Need ƒ Trading Attention for Feed-Forward Layers (ICASSP 2020)
- Lightweight transducer with WFST based decoding: Tiny Transducer: A Highly-efficient Speech Recognition Model on Edge Devices (ICASSP 2021)
- Cascade transducer: Cascade RNN-Transducer: Syllable Based Streaming On-device Mandarin Speech Recognition with a Syllable-to-Character Converter (SLT 2021)

Noisy Student Training(Self Training)

- Self training with filtering and ensembles: Self-training for end-to-end speech recognition (ICASSP 2020)
- Improved Noisy Student Training by gradational filtering: Improved Noisy Student Training for Automatic Speech Recognition (Interspeech 2020)

Self Supervised Learning(SSL)

APC(Autoregressive Predictive Coding)

- An Unsupervised Autoregressive Model for Speech Representation Learning (Interspeech 2019)
- Generative Pre-Training for Speech with Autoregressive Predictive Coding (ICASSP 2020)

CPC(Contrastive Predictive Coding)

- wav2vec: Unsupervised Pre-training for Speech Recognition (Schneider et al., 2019)
- vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations (Baevski et al., 2019)
- wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations (Baevski et al., 2020)