











# STT preliminary research

Overview of the STT field

SOTA STT Models

Aa Name	 Paper	 Year	 Affiliation	 Country	 # Citations	 Datasets	 Metrics	 Models	 Loss Func	 Type
<a href="#">Comparative study on Transformer vs RNN in speech applications</a>	<a href="https://arxiv.org/pdf/1909.06317.pdf">https://arxiv.org/pdf/1909.06317.pdf</a>	2019	University Colab	<span>Japan</span>	287	Multiple (Multi-language), most notable - LibriSpeech, CSJ, TED-LIUM, CHiME	<b>LibriSpeech</b> - RNN: 3.3 / 10.8 - Trans: 2.6 / 5.7 <b>CHiME-5</b> - RNN: 88.1 - Trans: 87.1	Transformer vs RNN	CE, CTC	<span>SOTA STT Models</span>
<a href="#">Conformer: Convolution-augmented Transformer for Speech Recognition</a>	<a href="https://arxiv.org/pdf/2005.08100.pdf">https://arxiv.org/pdf/2005.08100.pdf</a>	2020 (May)	Google		278	LibriSpeech	WER: 2.1% (clean) / 4.3% WER (Language Model): 1.9% (clean) / 3.9%	Conformer (Transformer + CNN)	Unknown	<span>SOTA STT Models</span>
<a href="#">Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition</a>	<a href="https://arxiv.org/pdf/2010.10504.pdf">https://arxiv.org/pdf/2010.10504.pdf</a>	2020 (Oct)	Google		64	LibriSpeech & LibriLight (Used for pre-training)	WER: 1.4% / 2.6%	Conformer SSL (Gen3)	Constrastive Loss (Pre-training)	<span>SOTA STT Models</span>
<a href="#">RWTH ASR Systems for LibriSpeech: Hybrid vs Attention - w/o Data Augmentation</a>	<a href="https://arxiv.org/pdf/1905.03072.pdf">https://arxiv.org/pdf/1905.03072.pdf</a>	2019	RWTH Aachen University	<span>Germany</span>	154	LibriSpeech	WER: 3.3% / 8.8%	LSTM + Transformer	CE	<span>SOTA STT Models</span>
<a href="#">A better and faster end-to-end model for streaming ASR</a>	<a href="https://arxiv.org/pdf/2011.10798.pdf">https://arxiv.org/pdf/2011.10798.pdf</a>	2021	Google	<span>USA</span>	21	Voice Search	WER, Latency	Conformer RNN-T with Cascaded Encoder	Unknown	<span>SOTA STT Models</span>
<a href="#">Multi-task self-supervised learning for robust speech recognition</a>	<a href="https://arxiv.org/pdf/2001.09239.pdf">https://arxiv.org/pdf/2001.09239.pdf</a>	2020	University Colab	<span>Multiple</span>	83	TIMIT, DIRHA and CHiME-5				<span>SOTA STT Models</span>
<a href="#">Exploring speech enhancement with generative adversarial networks for robust speech recognition</a>	<a href="https://arxiv.org/pdf/1711.05747.pdf">https://arxiv.org/pdf/1711.05747.pdf</a>	2018	UC San Diego Department of Music and Google		117	Wall Street Journal (WSJ) corpus and additional data from Youtube	WER	GAN	L1	<span>Speech Enhancement</span>
<a href="#">Time-domain speech enhancement using generative adversarial networks</a>	<a href="https://www.semanticscholar.org/paper/Time-domain-speech-enhancement-using-generative-Pascual-Serr%C3%A0/b0f67f1675de7ad0af8df1aff2e241b812ca18b">https://www.semanticscholar.org/paper/Time-domain-speech-enhancement-using-generative-Pascual-Serr%C3%A0/b0f67f1675de7ad0af8df1aff2e241b812ca18b</a>	2019		<span>Spain</span>	7	VCTK Corpus Demand (noises)	Test Set, Enhancer - WER for ASR-clean / ASR-MTR Clean, None - 11.9 / 14.3 MTR, None - 72.2 / 20.3 MTR, SEGAN - 80.7 / 52.8 MTR, FSEGAN - 33.3 / 25.4	Enhancers: SEGAN, FSEGAN		<span>Speech Enhancement</span>
<a href="#">Whispered-to-voiced Alaryngeal Speech Conversion with Generative Adversarial Networks</a>	<a href="https://arxiv.org/pdf/1808.10687.pdf">https://arxiv.org/pdf/1808.10687.pdf</a>	2018		<span>Spain</span>	6	Custom (CMU Artic corpus)		Adapted SEGAN		<span>Speech Enhancement</span>

Aa Name	🔗 Paper	≡ Year	≡ Affiliation	🌐 Country	# Citations	≡ Datasets	≡ Metrics	≡ Models	≡ Loss Func	🌐 Type
<a href="#">Accent modification for speech recognition of non-native speakers using neural style transfer</a>	<a href="https://asmp-aurasipjournals.springeropen.com/articles/10.1186/s13636-021-00199-3">https://asmp-aurasipjournals.springeropen.com/articles/10.1186/s13636-021-00199-3</a>	2021			2	- English Speech Database Read by Japanese Students (UME-ERJ) - LibriSpeech	WM / Autoencoder / Style Transfer CER: 46.3% / 36.1% / 31.7% WER: 56.8% / 43.2% / 34.9%	CNN for autoencoder CNN-RNN for Style transfer CNN-RNN for ASR		Speech Style Transfer
<a href="#">Improving Unsupervised Style Transfer in end-to-end Speech Synthesis with end-to-end Speech Recognition</a>	<a href="http://speech.ee.ntu.edu.tw/~tlkagk/paper/asr-guided-tacotron.pdf">http://speech.ee.ntu.edu.tw/~tlkagk/paper/asr-guided-tacotron.pdf</a>	2018	National Taiwan University	Taiwan	13	147 hours of American English audiobook data		Tacotron	Reconstruction loss	Speech Style Transfer
<a href="#">Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis</a>	<a href="https://arxiv.org/pdf/1803.09017.pdf">https://arxiv.org/pdf/1803.09017.pdf</a>				317					Speech Style Transfer

BPE		None	3.5	11.5	3.8	12.8
		LSTM	2.9	8.9	3.2	9.9
		Transformer	2.6	8.4	2.8	9.3
		4-gr	4.0	9.6	4.4	10.0
CDp	word		3.4	8.3	3.8	8.8
		+ LSTM	2.2	5.1	2.6	5.5
		Transformer resc.	1.9	4.5	2.3	5.0

Dictionary

- **STT** - Speech To Text
- **SSL** - self supervised learning
- **MTR** - multi-style training
- **GAN** - Generative adversarial networks
- **Loss functions:**
  - **CE Loss** - Cross Entropy loss
  - **CTC Loss** - Connectionist Temporal Classification Loss
- **Metrics:**
  - **WER** - Word error rate
  - **CER** - Character error rate

Datasets

- **Libri-Light**
- **LibriSpeech** - <https://paperswithcode.com/dataset/librispeech>
  - Dataset released in 2015 from audiobooks
- **REVERB** -
- **VoxForge** - <https://paperswithcode.com/dataset/voxforge>
  - ...
- **VoxPopuli**
  - <https://www.semanticscholar.org/paper/VoxPopuli%3A-A-Large-Scale-Multilingual-Speech-Corpus-Wang-Rivi%C3%A8re/6da74c5b1b3cffc236c4b2d75ac46b767f327e62>
- **CHiME** - noisy far-field multi-ch conversational
- **TED-LIUM 3** - <https://arxiv.org/abs/1805.04699>
  - Dataset released in 2018 from ted talks
- <https://www.openslr.org/resources.php>

[Python datasets](#)

Packed sequences

<https://stackoverflow.com/questions/51030782/why-do-we-pack-the-sequences-in-pytorch>