Almost Unsupervised Text to Speech and Automatic Speech Recognition

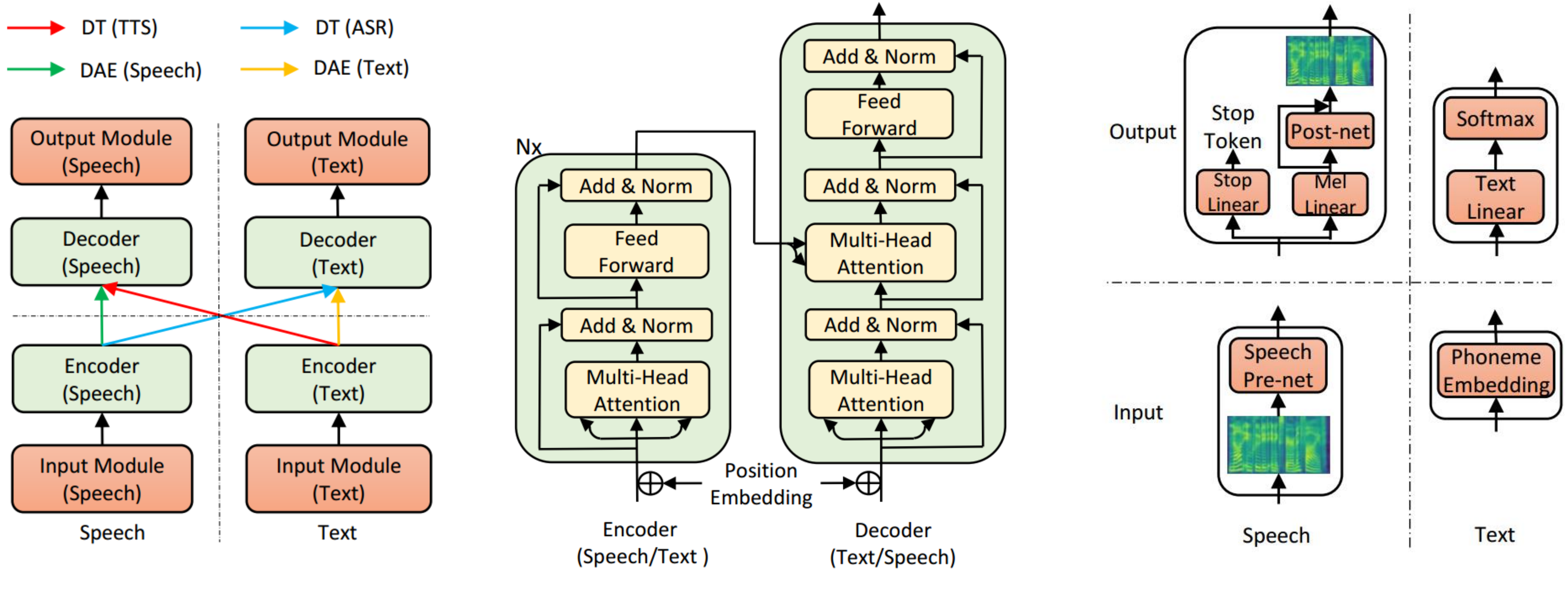
Yi Ren*, Xu Tan*, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu

Microsoft Research
Zhejiang University

Motivation

- ASR and TTS can achieve good performance given large amount of paired data. However, there are many low-resource languages in the world that are lack of supervised data to build TTS and ASR systems.
- We propose a practical way to leverage few paired data and additional unpaired speech and text data to build TTS and ASR systems.

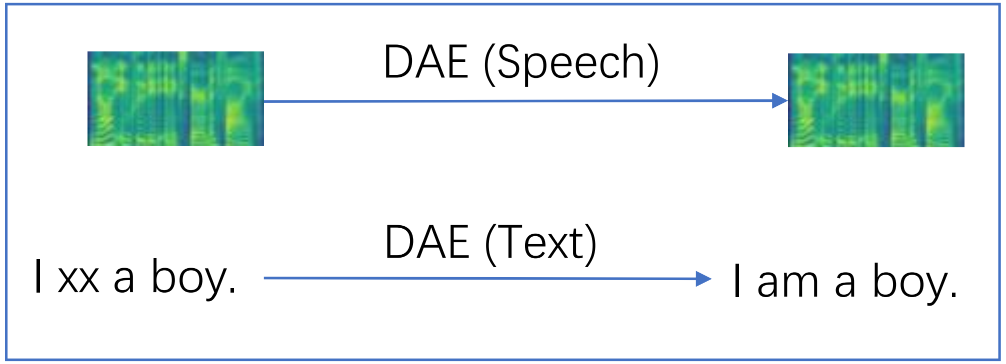
Model Architecture

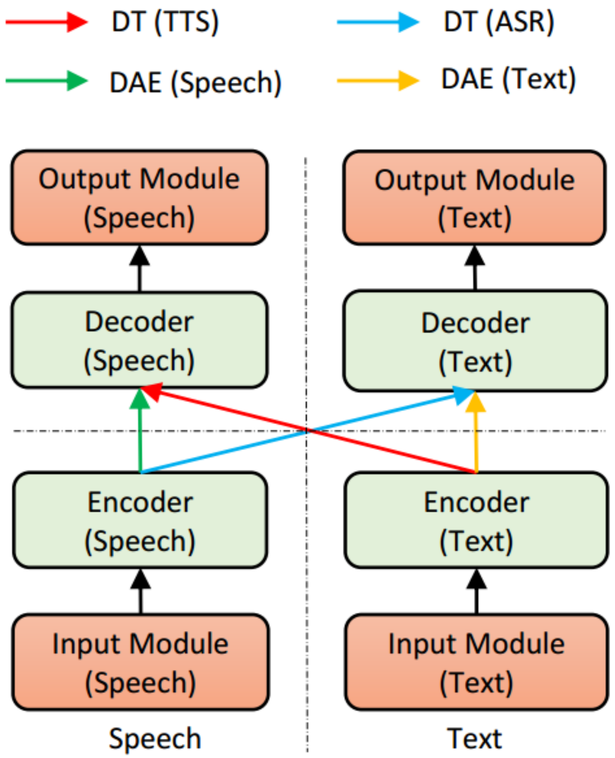


(a) Unified training flow (b) Encoder/Decoder for speech/text (c) Input/Output module for speech/text

Denoising Auto-Encoder

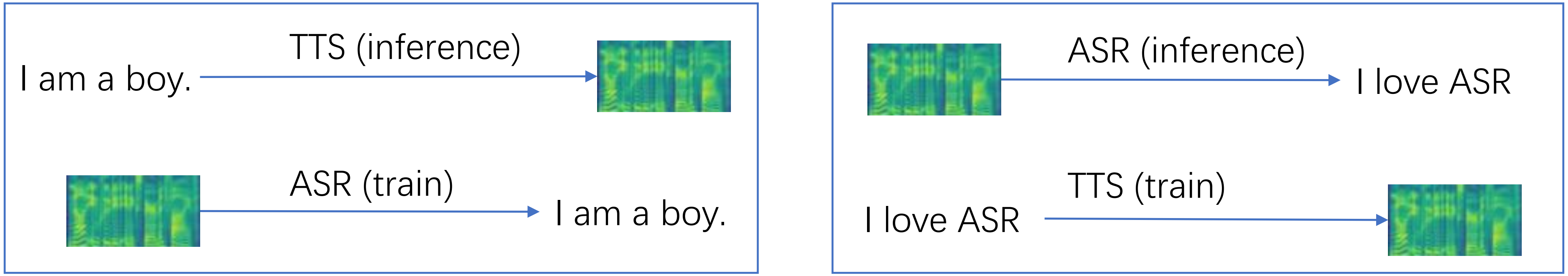
- We adopt denoising auto-encoder to build these capabilities. (Green and yellow lines)
  - Representation extraction: how to understand the speech or text sequence.
  - Language modeling: how to model and generate sequence in speech and text domain.





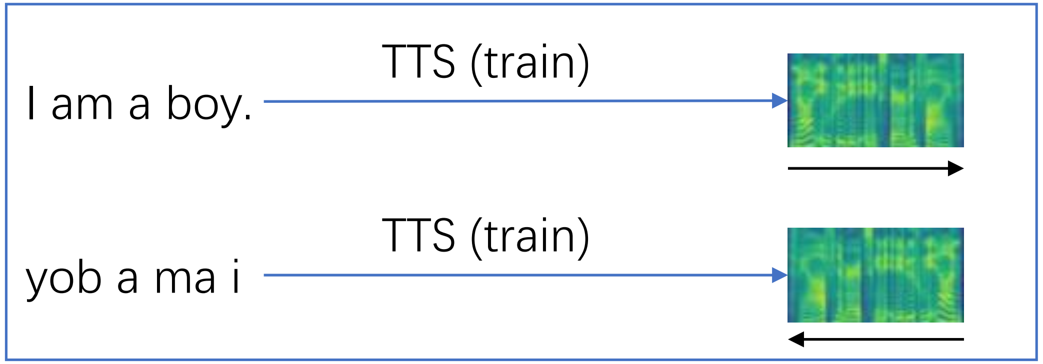
Dual Transformation

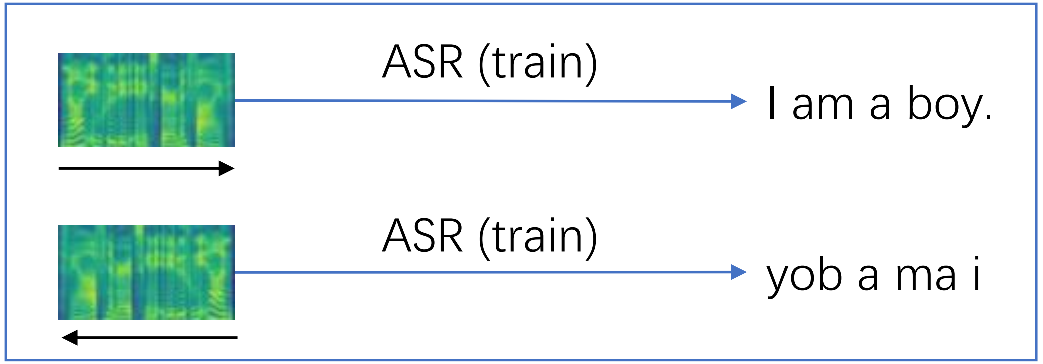
- Dual transformation is the key component to leverage the dual nature of TTS and ASR, and develop the capability of speech-text conversion.



Bidirectional Sequence Modeling

- Sequence generation suffers from **error propagation problem**, especially for the Speech sequence, which is usually longer than text.
- Due to dual transformation, the later part of the sequence is always of low quality.
- We propose the bidirectional sequence modeling (BSM) that generates the sequence in both left-to-right and right-to-left directions.





Audio Samples

Text	Printing then for our purpose may be considered as the art of making books by means of movable types.	A further development of the Roman letter took place at Venice.
Paired-200		
Our method		

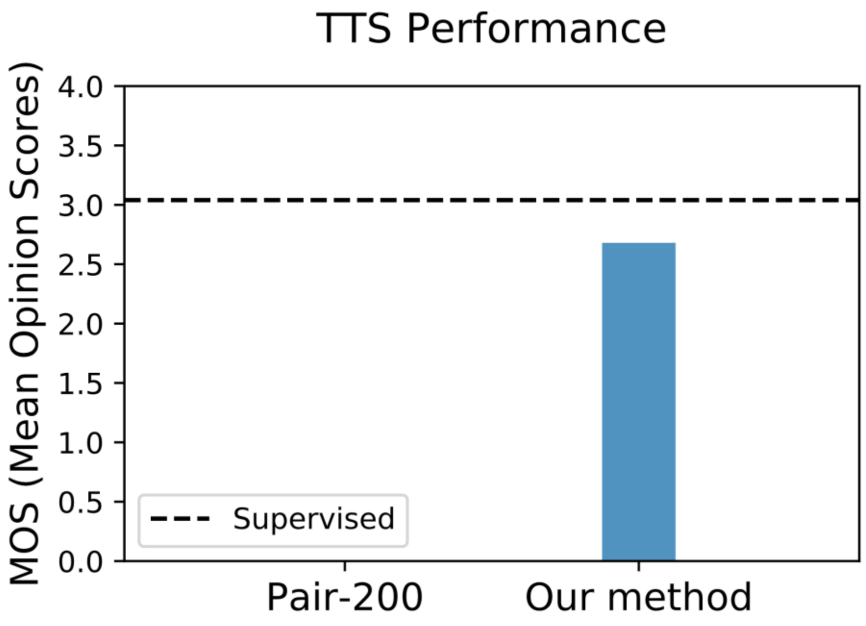
Results

Our Method: leverages 200 paired data + 12300 unpaired data
Pair-200: leverages only 200 paired data
Supervised: leverages all the 12500 paired data
GT: the ground truth audio
GT (Griffin-Lim): the audio generated from ground truth mel-spectrograms using Griffin-Lim algorithm

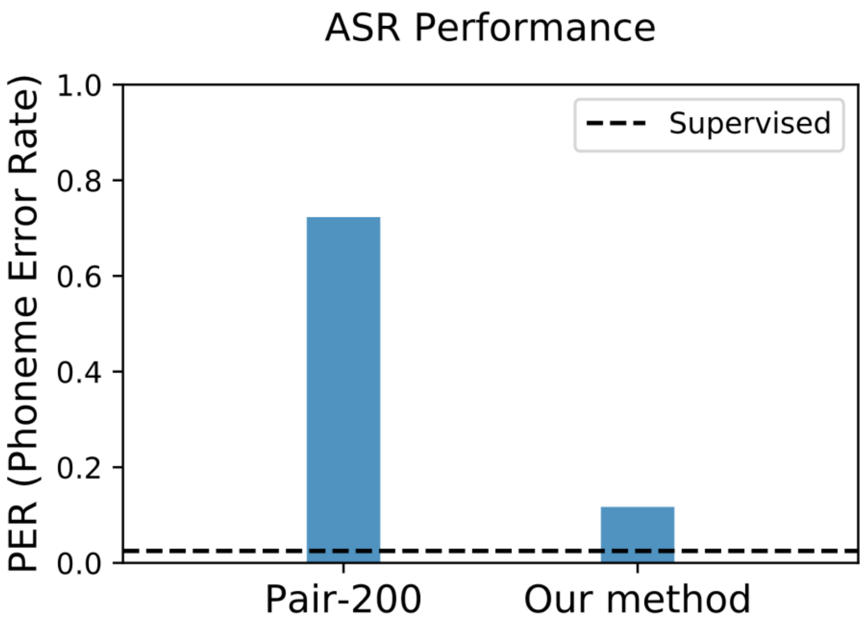
Method	MOS (TTS)	PER (ASR)
GT	4.54	-
GT (Griffin-Lim)	3.21	-
Supervised	3.04	2.5%
Pair-200	Null	72.3%
Our Method	2.68	11.7%

Results

Method	MOS (TTS)	PER (ASR)
GT	4.54	-
GT (Griffin-Lim)	3.21	-
Supervised	3.04	2.5%
Pair-200	Null	72.3%
Our Method	2.68	11.7%



The higher, the better



The smaller, the better

- Our method only leverages 200 paired speech and text data, and additional unpaired data
- Greatly outperforms the method only using 200 paired data
- Close to the performance of supervised method (using 12500 paired data)

Thanks!

Experiments

- Training and evaluation setup
  - Datasets
    - LJSpeech contains 13100 audio clips and transcripts, approximately 24 hours.
  - Evaluation
    - TTS: Intelligibility Rate and MOS (mean opinion score)
    - ASR: PER (phoneme error rate)

Analysis

- Ablation Study on **different components of our method**

Method	MOS (TTS)	PER (ASR)
Pair-200	Null	72.3%
Pair-200+DAE	Null	52.0%
Pair-200+DAE+DT	2.11	15.3%
Pair-200+DAE+DT+BSM	2.51	11.7%

Analysis

Method	MOS (TTS)	PER (ASR)
Pair-200	Null	72.3%
Pair-200+DAE	Null	52.0%
Pair-200+DAE+DT	2.11	15.3%
Pair-200+DAE+DT+BSM	2.51	11.7%