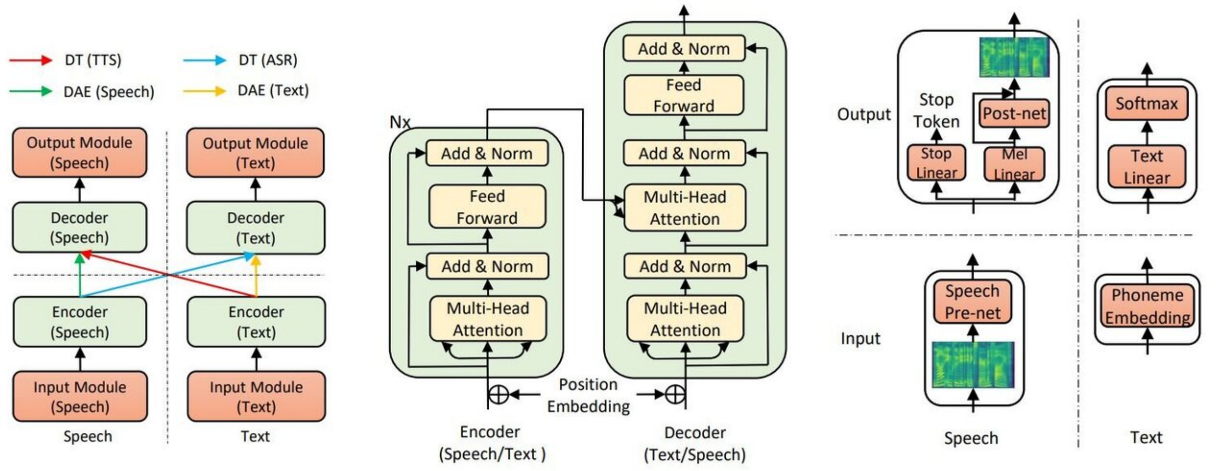
TechXplore

Only few hundred training samples bring human-sounding speech in Microsoft TTS feat

May 29 2019, by Nancy Cohen



(a) Unified training flow

(b) Encoder/Decoder for speech/text

(c) Input/Output module for speech/text

The overall model structure for TTS and ASR. Credit: Yi Ren, Xu Tan et al.

Microsoft Research Asia has been drawing applause for pulling off text to speech requiring little training—and showing "incredibly" realistic results.

Kyle Wiggers in *VentureBeat* said text-to-speech algorithms were not new and others quite capable but, still, the team effort at Microsoft still has an edge.

Abdullah Matloob in *Digital Information World*: "Text-to-speech conversion is getting smart with time, but the drawback is that it will still take an excessive amount of training time and resources to build a natural-sounding product."

Looking for a way to shrug off burdens of training time and resources to create output that was natural-sounding, Microsoft Research and Chinese researchers discovered another way to convert text-to-speech.

Fabienne Lang in *Interesting Engineering*: Their answer turns out to be an AI text-to-speech using 200 voice samples (only 200) to create realistic-sounding speech to match transcriptions. Lang said, "This means approximately 20 minutes' worth."

That the requirement was only 200 audio clips and corresponding transcriptions impressed Wiggers in *VentureBeat*. He also noted that the researchers devised an AI system "that leverages unsupervised learning—a branch of machine learning that gleans knowledge from unlabeled, unclassified, and uncategorized test data."

Their paper is up on arXiv. "Almost Unsupervised Text to Speech and Automatic Speech Recognition" is by Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu. Author affiliations are Zhejiang University, Microsoft Research and Microsoft Search Technology Center (STC) Asia.

In their paper, the team said that the TTS AI utilizes two key components, a Transformer and denoising auto-encoder, to make it all work.

"Through the transformers, Microsoft's text-to-speech AI was able to recognize speech or text as either input or output," said an article in *Edgy* by Rechelle Fuertes.

Tyler Lee in *Ubergizmo* provided a definition of transformer: "Transformers...are deep neural networks designed to emulate the neurons in our brain.."

MathWorks had a definition for autoencoder. "An autoencoder is a type of artificial neural network used to learn efficient data (codings) in an unsupervised manner. The aim of an auto encoder is to learn a representation (encoding) for a set of data, denoising autoencoders is typically a type of autoencoders trained to ignore 'noise' in corrupted input samples."

Did results of their experiment show their idea is worth chasing? "Our method achieves 99.84% in terms of word level intelligible rate and 2.68 MOS for TTS, and 11.7% PER for ASR [automatic speech recognition] on LJSpeech dataset, by leveraging only 200 paired speech and text data (about 20 minutes audio), together with extra unpaired speech and text data."

Why this matters: This approach may make text to speech more accessible, said reports.

"Researchers are continually working to improve the system, and are hopeful that in the future, it will take even less work to generate lifelike discourse," said Lang.

The paper will be presented at the International Conference on Machine Learning, in Long Beach California later this year, and the team plans to release the code in the coming weeks, said Wiggers.

Meanwhile, the researchers are not yet walking away from their work in presenting transformations with few paired data.

"In this work, we have proposed the almost unsupervised method for text to speech and automatic speech recognition, which leverages only few paired speech and text data and extra unpaired data... For future work, we will push toward the limit of unsupervised learning by purely leveraging unpaired speech and text data, with the help of other pre-training methods."

More information: Almost Unsupervised Text to Speech and Automatic Speech Recognition: speechresearch.github.io/unsuper/

© 2019 Science X Network

Citation: Only few hundred training samples bring human-sounding speech in Microsoft TTS feat (2019, May 29) retrieved 2 May 2024 from https://techxplore.com/news/2019-05-samples-human-sounding-speech-microsoft-tts.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.