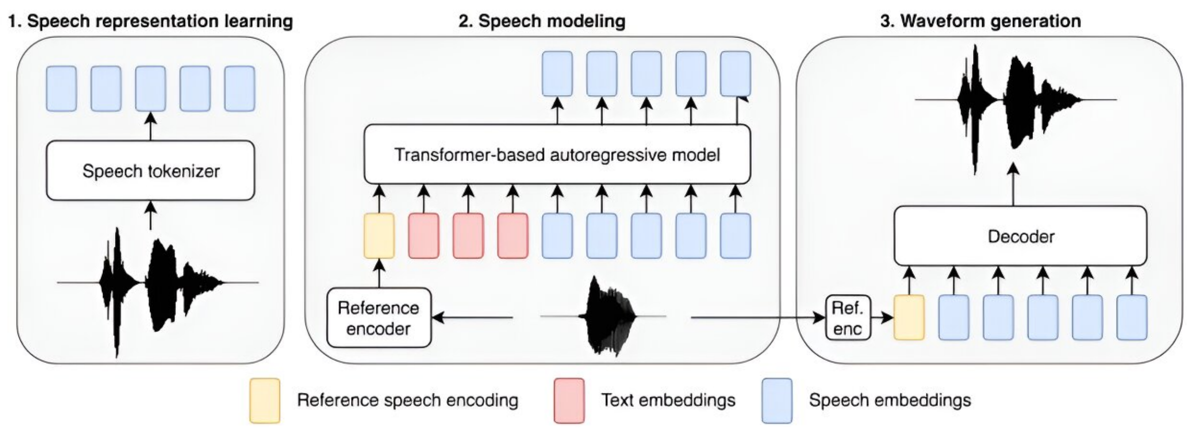
Amazon unveils largest text-to-speech model ever made

February 17 2024, by Bob Yirka



An overview of BASE TTS. The speech tokenizer (1) learns a discrete representation, which is modeled by an autoregressive model (2) conditioned on text and reference speech. The speechcode decoder (3) converts predicted speech representations into a waveform. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2402.08093

A team of artificial intelligence researchers at Amazon AGI announced the development of what they are describing as the largest text-to-speech model ever made. By largest, they mean having the most parameters and using the largest training dataset. They have published a paper on the *arXiv* preprint server describing how the model was developed and trained.

LLMs like ChatGPT have gained attention for their human-like ability to answer questions intelligently and create high-level documents. But AI is still making its way into other mainstream applications, as well. In this new effort, the researchers attempted to improve the ability of a text-to-speech application by increasing its number of parameters and adding to its training base.

The new model, called Big Adaptive Streamable TTS with Emergent abilities, (BASE TTS for short) has 980 million parameters and was trained using 100,000 hours of recorded speech (found on public sites), most of which was in English. The team also gave it examples of spoken words and phrases in other languages to allow the model to correctly pronounce well-known phrases when it encounters them—"au contraire," for example, or "adios, amigo."

The team at Amazon also tested the model on smaller data sets, hoping to learn where it develops what has come to be known in the AI field as an emergent quality, in which an AI application, whether an LLM or text-to-speech application, suddenly seems to break through to a higher level of intelligence. They found that for their application, a medium-sized dataset was where the leap to a higher level occurred, at 150 million parameters.

They also noted that the leap involved a host of language attributes, such as the ability to use compound nouns, to express emotions, to use foreign words, to apply paralinguistics and punctuation and to ask questions with the emphasis placed on the right word in a sentence.

The team says that BASE TTS will not be released to the public—they fear it might be used unethically—instead, they plan to use it as a learning application. They expect to apply what they have learned thus far to improve the human-sounding quality of text-to-speech applications in general.

More information: Mateusz Łajszczak et al, BASE TTS: Lessons from building a billion-parameter Text-to-Speech model on 100K hours of data, *arXiv* (2024). DOI: 10.48550/arxiv.2402.08093

www.amazon.science/publication … n-100k-hours-of-data