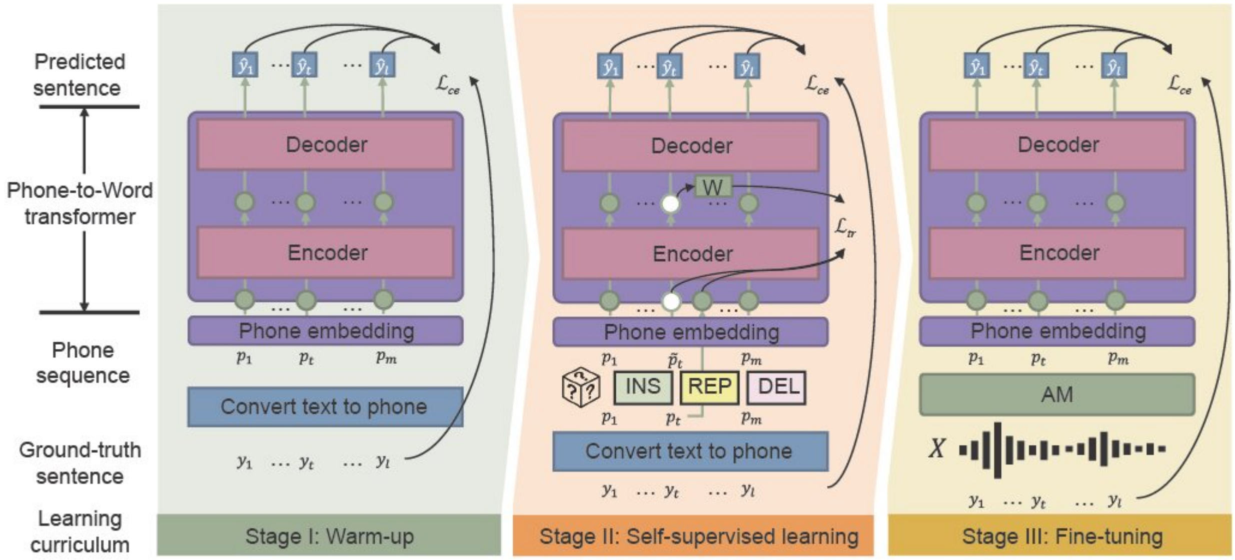
TechXplore

Researchers propose new and more effective model for automatic speech recognition

September 2 2022



The phonetic-semantic pre-training (PSP) framework uses "noise-aware curriculum" learning to effectively improve the performance of ASR in noisy environments. integrating warm-up, self-supervised learning, and fine-tuning. Credit: *CAAI Artificial Intelligence Research*, Tsinghua University Press

Popular voice assistants like Siri and Amazon Alexa have introduced automatic speech recognition (ASR) to the wider public. Though decades in the making, ASR models struggle with consistency and reliability, especially in noisy environments. Chinese researchers developed a framework that effectively improves the performance of ASR for the chaos of everyday acoustic environments.

Researchers from the Hong Kong University of Science and Technology and WeBank proposed a new framework—phonetic-semantic pre-training (PSP) and demonstrated the robustness of their new model against synthetic highly noisy speech datasets.

Their study was published in *CAAI Artificial Intelligence Research* on Aug. 28.

"Robustness is a long-standing challenge for ASR," said Xueyang Wu from the Hong Kong University of Science and Technology Department of Computer Science and Engineering. "We want to increase the robustness of the Chinese ASR system with a low cost."

ASR uses machine-learning and other artificial intelligence techniques to automatically translate speech into text for uses like voice-activated systems and transcription software. But new consumer-focused applications increasingly call for voice recognition to work better—handle more languages and accents, and perform more reliably in real-life situations like video conferencing and live interviews.

Traditionally, training the acoustic and language models that comprise ASR requires large amounts of noise-specific data, which can be time- and cost-prohibitive.

The acoustic model (AM) turns words into a "phones," which are sequences of basic sounds. The language model (LM) decodes phones into natural-language sentences, usually with a two-step process: a fast but relatively weak LM generates a set of sentence candidates, and a powerful but computationally expensive LM selects the best sentence from the candidates.

"Traditional learning models are not robust against noisy acoustic model outputs, especially for Chinese polyphonic words with identical pronunciation," Wu said. "If the first pass of the learning model decoding is incorrect, it is extremely hard for the second pass to make it up."

The newly proposed framework PSP makes it easier to recover misclassified words. By pre-training a model that translates the AM outputs directly to sentence along with the full context information, researchers can help the LM efficiently recover from the noisy outputs of the AM.

The PSP framework allows the model to improve through a pre-training regime called noise-aware curriculum that gradually introduces new skills, starting easy and gradually moving into more complex tasks.

"The most crucial part of our proposed method, Noise-aware Curriculum Learning, simulates the mechanism of how human beings recognize a sentence from noisy speech," Wu said.

Warm-up is the first stage, where researchers pre-train a phone-to-word transducer on a clean phone sequence, which is translated from unlabeled text data only—to cut back on the annotation time. This stage "warms up" the model, initializing the basic parameters to map phone sequences to words.

In the second stage, self-supervised learning, the transducer learns from more complex data generated by self-supervised training techniques and functions. Finally, the resultant phone-to-word transducer is fine-tuned with real-world speech data.

The researchers experimentally demonstrated the effectiveness of their framework on two real- life datasets collected from industrial scenarios and synthetic noise. Results showed that the PSP framework effectively improves the traditional ASR pipeline, reducing the relative character error rates by 28.63% for the first dataset and 26.38% for the second.

In next steps, researchers will investigate more effective PSP pre-training methods with larger unpaired datasets, seeking to maximize the effectiveness of pretraining for noise-robust LM.

More information: Xueyang Wu et al, A Phonetic-Semantic Pre-Training Model for Robust Speech Recognition, *CAAI Artificial Intelligence Research* (2022). DOI: 10.26599/AIR.2022.9150001

Provided by Tsinghua University Press

Citation: Researchers propose new and more effective model for automatic speech recognition (2022, September 2) retrieved 27 April 2024 from https://techxplore.com/news/2022-09-effective-automatic-speech-recognition.html