A BLUEPRINT FOR AI-POWERED SMART SPEECH TECHNOLOGY

Innovative researchers are coming up with cutting-edge solutions to the many TECHNOLOGICAL CHALLENGES OF SPEECH RECOGNITION.

Speech recognition technology is becoming increasingly crucial to our daily lives, and iFLYTEK, based in Hefei, China, has been working on new ways of using this smart technology since the company was founded in 1999.

But, as speech recognition begins to cope with new needs — such as conversion from a complex mess of overlapping speech to text in noisy environments, and human-computer interaction in the era of the Internet of Things — there are technical hurdles which need to be rapidly overcome.

In recent years, the award-winning company has transformed the application of speech recognition from simple scenarios such as voice input and smart call centres, to more complex uses, such as detecting and transcribing speech when there are multiple speakers or multiple languages.

▲ iFLYTEK's Cong Liu introduces children to some of the firm's technology.

COMPLEX SPEECH SCENARIOS

Speech recognition needs to be better at coping in complex scenarios, such as noisy environments, or when multiple speakers speak over one another.

"What we are trying to tackle is the long-standing 'cocktail party' problem. It's very difficult to achieve with existing speech recognition technology," says Cong Liu, executive director of the iFLYTEK Research Institute.

To address the problem of complex speech scenarios, Liu and his team developed a new speech-recognition framework which encompasses two algorithms, Spatial-and-Speaker-Aware Iterative Mask Estimation (SSA-IME) and Spatial-and-Speaker-Aware Acoustic Model (SSA-AM). The SSA-IME combines traditional signal-processing and deep-learning based methods to identify multiple speakers. The SSA-AM then integrates the data extracted by the SSA-IME to recognise each speaker's voice. Furthermore, the estimated text can provide useful information for the SSA-IME to extract more robust spatial- and speaker-aware features. The framework iteratively reduces background noise and interference to speakers' voices, resulting in more accurate speech recognition.

Liu's team won the 6th CHiME speech separation and recognition challenge in 2020 (CHiME-6), with a result of 70% accuracy, but their goal is to improve that rate to 95%.

"The day we accomplish that, we can call the mission completed," he says. "But it will take another three to five years."

MULTILINGUAL SPEECH RECOGNITION

The accuracy of a speech recognition system depends on a large amount of labelled data. Such data sources are rich for widely spoken languages such as Chinese and English, but the problem is harder for other languages. However, with growing globalization, it is urgent to build accurate speech recognition systems for so-called low-resource languages.

In order to achieve this, Liu and his team have developed a framework called Unified Spatial Representation Semi-supervised Automatic Speech Recognition (USRS-ASR), which allows unlabelled data from less commonly spoken languages to be widely used.

Liu's team has used this framework to create an automatic speech recognition (ASR) system baseline with powerful acoustic modelling capabilities, and in 2021, it won the OpenASR Challenge, a major international low-resource ASR contest organized by the U.S. National Institute of Standards and Technology, coming first in 22 categories of 15 languages.

Moving into new areas, Liu and his team have now built a new paradigm of human-computer interaction based on multi-modal information. For example, through a deep understanding of a driver's intent — with the aid of visual perception technology and multi-modal interaction — AI can focus on the target person's voice in noisy settings.

iFLYTEK has adopted its cutting-edge speech technologies in designing many of its products. These include: the smart voice recorder, with an eight-channel microphone array; the smart notebook 'AI Note', specially designed for conference audio recording and audio-text transcription; and the input tools which recognise speech in English, Mandarin Chinese and 25 Chinese dialects without switching between them. ■

科大讯飞 iFLYTEK

www.iflytek.com