Magic Data

Who are we?

Magic Data Technology is an AI data service provider. We are committed to providing a wild range of customized data services in the fields of automatic speech recognition, text to speech, computer vision recognition and Natural Language Processing. With human-in-the-loop data processing, we significantly improved the efficiency and quality of AI data labeling. We have collected more than 100,000 hours of standard multilingual speech corpora under various scenarios. We help our clients gain easy and timely access to data with accuracy up to 99% and also provide them customized solutions. Magic Data employs a vast team of skilled data specialists and has a wide network of consultants around the world to assist with specific data needs.

Why Us

Security

Our machine-learning powered data platform provides data encryption and supervision services throughout the data lifecycle. We are accredited by ISO9001, ISO27001, CMMI3, we ensure the data quality and security.

Quality

In the process of data collection, real-time review is conducted to ensure the quality. In the data annotation process, human-machine collaboration methodology is adopted to improve the accuracy and consistency.

Variety

Our scenario-based data collection and annotation services are valuable for a vast variety of AI domains such as virtual assistant & chatbot, intelligent customer service, smart home controls, and content censoring.

Scale

We can help to build and improve AI systems with over 300,000 contractors and 50+ languages and dialects.

Scan QR code for more information：

Email：business@magicdatatech.com

website: www.magicdatatech.com

Selected Language Datasets

We provide valuable and reliable training data to empower your AI models. You can find datasets in different languages, styles, and solutions. Our datasets can improve your AI models' performance, thus accelerating the commercialization of AI initiatives.

Dataset Style	Dataset Name
Read Speech Corpus	American English Speech Corpus
	Bahasa Indonesian Speech Corpus
	Chinese-English Code-Mixing Speech Corpus
	French Speech Corpus
	Guangzhou Cantonese In-Vehicle Speech Corpus
	Guangzhou Cantonese Speech Corpus
	Japanese Speech Corpus
	Korean Speech Corpus
	Mandarin Chinese Speech Corpus
	Minnan Dialect Speech Corpus
	Peninsular Spanish Speech Corpus
	Shanghai Dialect Speech Corpus
	Shanxi Dialect Speech Corpus
	Thai Speech Corpus
	Wuhan Dialect Speech Corpus
Conversational Speech Corpus	Bahasa Indonesian Conversational Speech Corpus
	Brazilian Portuguese Conversational Speech Corpus
	English Conversational Telephone Speech Corpus
	French Conversational Speech Corpus
	German Conversational Speech Corpus
	Guangzhou Cantonese Conversational Speech Corpus
	Hangzhou Dialect Conversational Speech Corpus
	Italian Conversational Speech Corpus
	Japanese Conversational Speech Corpus
	Japanese English Conversational Speech Corpus
	Korean Conversational Speech Corpus
	Korean English Conversational Speech Corpus
	Malay Conversational Speech Corpus
	Mandarin Chinese Conversational Speech Corpus
	Mandarin Chinese Conversational Telephone Speech Corpus
	Peninsular Arabic Conversational Speech Corpus

	Peninsular Spanish Conversational Speech Corpus
	Shanghai Dialect Conversational Speech Corpus
	Sichuan Dialect Conversational Speech Corpus
	Turkish Conversational Speech Corpus
	Uyghur Conversational Speech Corpus
Speech Corpus for TTS	American English Speech Corpus for TTS
	Chinese female customer service TTS dataset
	Chinese female voice emotion TTS dataset
	Mandarin Chinese Speech Corpus for TTS

OPENSOURCE DATASETS:

A package of 9.97 hours of high quality, multi-language conversational speech datasets collected from native speakers, ready for your AI or ML product or service.

Language	Duration	Speech Contents	Recording Environments	Audio Parameters	Equipment
Italian	1.26 h	Spontaneous Conversations	Indoor Environments	16 kHz, 16 bits	Mobiles
German	1.05 h	Spontaneous Conversations	Indoor Environments	16 kHz, 16 bits	Mobiles
French	1.01 h	Spontaneous Conversations	Indoor Environments	16 kHz, 16 bits	Mobiles
Spanish	2.21 h	Spontaneous Conversations	Indoor Environments	16 kHz, 16 bits	Mobiles
Yemeni Arabic	2.13 h	Spontaneous Conversations	Indoor Environments	16 kHz, 16 bits	Mobiles
Brazilian Portuguese	2.31 h	Spontaneous Conversations	Indoor Environments	16 kHz, 16 bits	Mobiles

Application Scenarios: Virtual Assitants, Consumer Robot Controls, Smart Home Controls, Car Infotainment, Security and Authentication etc.

For more information, please contact: business@magicdatatech.com

Industrial online forum

Theme: Data sets your model --Which data strategy should be adopted to achieve better performance?

Data: October 28th

Time: 20:15-20:45 GMT+8

Experts: Daniel Povey, Xuelu Zhang, Gaofeng Cheng

Link: https://zoom.com.cn/j/67686637114

Meeting ID: 67686637114